



Proceedings

ITISE 2017

**International work-conference
on Time Series**

Granada

September, 18-20 2017

Volumen 2

Proceedings ITISE 2017.

International work-conference on Time Series

Editors and Chairs

Olga Valenzuela
Fernando Rojas
Héctor Pomares
Ignacio Rojas

University of Granada

ISBN: 978-84-17293-01-7

Legal Deposit: GR 1369-2017

Edit and Print: Godel Impresiones Digitales S.L.

All rights reserved to authors. The total or partial reproduction of this work is strictly prohibited, without the strict authorization of the copyright owners, under the sanctions established in the laws.

Preface

We are proud to present the set of final accepted papers for the fourth edition of the ITISE 2017 conference "International work-conference on Time Series" held in Granada (Spain) during September, 18-20, 2017.

The ITISE 2017 (International work-conference on Time Series) seeks to provide a discussion forum for scientists, engineers, educators and students about the latest ideas and realizations in the foundations, theory, models and applications for interdisciplinary and multidisciplinary research encompassing disciplines of computer science, mathematics, statistics, forecaster, econometric, etc, in the field of time series analysis and forecasting.

The aims of ITISE 2017 is to create a friendly environment that could lead to the establishment or strengthening of scientific collaborations and exchanges among attendees, and therefore, ITISE 2017 solicits high-quality original research papers (including significant work-in-progress) on any aspect time series analysis and forecasting, in order to motivating the generation, and use of knowledge and new computational techniques and methods on forecasting in a wide range of fields.

The list of topics in the successive Call for Papers has also evolved, resulting in the following list for the present edition:

1. Time Series Analysis and Forecasting.

- Nonparametric and functional methods
- Vector processes
- Probabilistic Approach to Modeling Macroeconomic Uncertainties
- Uncertainties in forecasting processes
- Nonstationarity
- Forecasting with Many Models. Model integration
- Forecasting theory and adjustment
- Ensemble forecasting
- Forecasting performance evaluation
- Interval forecasting
- Econometric models
- Econometric Forecasting
- Data preprocessing methods: Data decomposition, Seasonal adjustment, Singular spectrum analysis, Detrending methods, etc.

2. Advanced method and on-Line Learning in time series.

- Adaptivity for stochastic models
- On-line machine learning for forecasting
- Aggregation of predictors
- Hierarchical forecasting
- Forecasting with Computational Intelligence
- Time series analysis with computational intelligence

- Integration of system dynamics and forecasting models

3. High Dimension and Complex/Big Data.

- Local Vs Global forecast
- Techniques for dimension reduction
- Multiscaling
- Forecasting Complex/Big data

4. Forecasting in real problem.

- Health forecasting
- Telecommunication forecasting
- Modelling and forecasting in power markets
- Energy forecasting
- Financial forecasting and risk analysis
- Forecasting electricity load and prices
- Forecasting and planning systems
- Real time macroeconomic monitoring and forecasting
- Applications in: energy, finance, transportation, networks, meteorology, health, research and environment, etc.

After a careful peer review and evaluation process (each submission was reviewed by at least 2, and on the average 2.9, program committee members or additional reviewer), 121 contributions are presenting in this proceedings (accepted for oral, poster or virtual presentation, according to the recommendations of reviewers and the authors' preferences.

In this edition of ITISE, we are honored to have the following invited speaker:

1. Prof. Dr. Fredj Jawadi , Associate Professor of Economics (MCF-HDR) at the University of Evry, France.
2. Prof. Dr. Joerg Breitung, Professor in the Center of Econometrics and Statistics, University of Cologne, Germany
3. Dr. Travis J. Berge, Senior Economist. Board of Governors of the Federal Reserve System, USA.
4. Dr. Anna Korzeniewska, Faculty, Department of Neurology at Johns Hopkins University School of Medicine, Baltimore MD, USA
5. Dr. Joan Paredes, Senior Scientist, Dr. Joan Paredes, European Central Bank, Frankfurt am Main, Germany.
6. Dr. Pekka Koponen, Senior Scientist, D.Sc.Tech, VTT Technical Research Centre of Finland, Energy Systems, P.O. Box 1000, FI-02044 VTT, Finland

During ITISE 2017 several Special Sessions will be carried out. Special Sessions will be a very useful tool in order to complement the regular program with new and emerging topics of particular interest for the participating community. Special Sessions that emphasize on multi-disciplinary and transversal aspects, as well as cutting-edge topics are especially encouraged and welcome.

This fourth edition of ITISE was organized at the Universidad de Granada, with the help of the Spanish Chapter of the IEEE Computational Intelligence Society and Spanish Network Time Series (RESET). We wish to thank to our main sponsor the institutions Faculty of Science, Dept. Computer Architecture & Computer Technology and CITIC-UGR from the University of Granada for their support. We wish also to thank to the Dr. Veronika Rosteck and Dr. Eva Hiripi, Springer, Associate Editor, for their interest in the future editing a book series of Springer from the best papers of ITISE 2017.

We would also like to express our gratitude to the members of the different committees and to the reviewer for their support, collaboration and good work.

September, 2017
Granada

ITISE Editors and Chairs
Olga Valenzuela
Fernando Rojas
Hector Pomares
Ignacio Rojas

Steering and Local Committee

Amaury Lendasse	University of Iowa
Hector Pomares	University of Granada
Fernando Rojas	University of Granada
Ignacio Rojas	University of Granada
Olga Valenzuela	University of Granada

Program Committee

Adnan Sozen	Juan María Palomo
Ahlame Douzal	Julien Chevallier
Alan Wee-Chung Liew	Junsoo Lee
Alberto Guillén	K. Muraleedharan
Alexandra Spitz-Oener	Kalle Saastamoinen
Alexey Koronovskii	Katerina Tsakiri
Alicia Troncoso	Kit Yan Chan
Aman Ullah	Konstantinos Spiliopoulos
Amaury Lendasse	Krzystof Siwek
Anke Meyer-Baese	Lazaros Iliadis
Ansgar Steland	Leonid Sheremetov
Antonio J. Rivera Rivas	Leopold Soegner
Antonio Montañés	Leszek Borzemski
Asesh Roychowdhury	Loli Perez
Athanasios Sfetsos	Luca Faes
Axel Werwatz	Luis Javier Herrera
Bormin Huang	Manfred Deistler
C.Germán Castellanos Dominguez	María Dolores Gadea
Calvin Wong	María José Del Jesús Díaz
Carlos Henrique Ribeiro Lima	Marc Hallin
Caroline Uhler	Marcel Ausloos
Cecilio Tamarit	Marco Lippi
Chang-Yong Lee	Martin Wagner
Charilaos Kourogiorgas	Matteo Barigozzi
Charles Efferson	Mehdi Vafakhah
Chor Foon Tang	Micael Castanheira
Christian Brownlees	Miquel Montero Torralbo
Christian Gourieroux	Mohammed Rezaul Karim
Christoph Winter	Naoufel Cheikhrouhou
Christopher Burke	Narayanan Kumarappan
Chunshien Li	Nicolás Marín Ruiz
Claudia Villalonga	Olga Valenzuela
Dalia Kriksciuniene	Oresti Banos
Daniel Castillo	P.C. Nayak
Daniel Peña Sanchez	Panos Pardalos
David Giles	Paulo Cortez

Dimitris Varoutas	Paulo Rodrigues
Dorel Aiordachioaie	Pei-Chann Chang
Elmar Lang	Peter Gloesekoetter
Erol Egrioglu	Peter M. Robinson
Eros Pasero	Philipp Sibbertsen
Ferda Halicioglu	Philippe Weil
Fernando Pérez de Gracia	Pierpaolo D'Urso
Fernando Rojas	Plamen Ivanov
Fionn Murtagh	Popescu Theodor Dan
Florian Zimmermann	Ragulskis Minvydas
Francisco Estrada	Rajendra Udyavara Acharya
Francisco Martínez Álvarez	Rebecca Killick
François Schmitt	Ricardo de Andrade Araújo
Fredj Jawadi	Rosangela Ballini
Fuxia Cheng	Ruqiang Yan
G.S. YIN	Ryszard Tadeusiewicz
Gabriel Pérez Quirós	Sajjad Ahmad
Gani Aldashev	Salah Bourennane
Georg Görg	Samrad Jafarian-Namin
Gerhard Rünstler	Shilu Tong
Germán Gutiérrez Sánchez	Siem Jan Koopman
Gilney Zebende Zebende	Silke Hüttel
Guy Mélard	Slawek Zadrozny
Héctor Pomares	Suresh Sethi
Hakan Aladag	Tatiana Afanaseva
Hashem Pesaran	Thomas Epper
Heather Ruskin	Tin-Chih Toly Chen
Hisham El-Shishiny	Tobias Preis
Hooi Hooi Lean	Tomas Cipra
Hui Liu	Toshihisa Funabashi
Ignacio Rojas	Tsangyao Chang
Ildar Batyrshin	Tzung-Pei Hong
Ilhan Ozturk	Ulrich Foelsche
Irina Perfilieva	V. Jothiprakash
Isaias Lima	Vadlamani Ravi
Janusz Kacprzyk	Vijay P. Singh
Janusz Miskiewicz	Vladas Pipiras
Javier Hualde	Wei-Chiang Hong
Jeff Yao	Wei-Xing Zhou
Jesús Gonzalo	Wieslaw M. Macek
Jianbo Gao	Willem K. M. Brauers
Jing Shi	William A. Barnett
João Catalão	Witold Pedrycz
José L. Aznarte	Wolfgang K Härdle
José María Amigó García	Yixiao Sun
Josep Lluís Carrion-i-Silvestre	Yucheng Dong
Josu Arteche	Yukun Bao
Juan Manuel Galvez	

PROCEEDINGS ITISE 2017

Advanced in Time Series Forecast

Alternative Solution for the Adjustment of Defect Liability Period in Construction	1
<i>Kichang Jeong, Woo-Ram Kim and Jaeseob Lee</i>	
Time Series Anomaly Detection with Discrete Wavelet Transforms and Maximum Likelihood Estimation	11
<i>Markus Thill, Wolfgang Konen and Thomas Baeck</i>	
Robust Multivariate Time Series Analysis in Nonlinear Models with Autoregressive and t-Distributed Errors	23
<i>Hamza Alkhatib, Boris Kargoll and Jens-André Paffenholtz</i>	
Kurtosis Computations and Black-Scholes Model with GARCH Volatility	37
<i>Muhammad Sheraz</i>	
Robust estimation of covariance and correlation functions of a stationary multivariate process	47
<i>Higor Cotta, Valdério Reisen, Pascal Bondon and Wolfgang Stummer</i>	

Applications in Time Series

Forecasting German Crash Numbers: The Effect of Meteorological Variables	59
<i>Kevin Diependaele, Heike Martensen, Markus Lerner, Andreas Schepers, Frits Bijleveld and Jacques J.F. Commandeur</i>	
Safety stock calculation based on kernel bandwidth estimates that minimize inventory costs	74
<i>Carlos Ruiz-Cañadas and Juan R. Trapero</i>	
Forecasting Diffusion Investments in FinTech Using Diffusion Models	76
<i>Miriam Scaglione and Simone Dimitriou</i>	
Multiple seasonal Holt-Winters improvement for the special events forecast using Discrete-Interval Multiple Seasonalities	91
<i>Juan Carlos García-Díaz and Oscar Trull</i>	
Chaos Neural Network for Ultra-Long Period Pseudo-Random Number Generator	102
<i>Hitoaki Yoshida, Yukito Kon and Takeshi Murakami</i>	
Joint Multifractal Description of the Influence of Climatic Variables on Reference Evapotranspiration Time Series	114
<i>Ana Belén Ariza Villaverde, Pablo Pavón Domínguez, Juan María Gómez López, Eduardo Gutiérrez de Ravé Agüera and Francisco José Jiménez Hornero</i>	
Generalized nonparametric method for analyzing economic data inconsistent with the model of single rational representative consumer	117
<i>Nikolay Klemashev and Alexander Shananin</i>	
Geomagnetic Storms, Earthquakes and their Influence on GNSS Coordinate Time Series ..	122
<i>Inese Varna, Janis Balodis and Diana Haritonova</i>	

Forecasting Power Output of Photovoltaic Systems Using Linear, Non-Linear and Enhanced Models	129
<i>Georgia Xanthopoulou, Athanasios Salamanis, Dionysios Kehagias, Ioannis Antoniou, Charalampos Bratsas and Dimitrios Tzovaras</i>	
Extreme value analysis of geomagnetic activity based on the data from Canadian geomagnetic observatories	141
<i>Lidia Nikitina, Larisa Trichtchenko, David Boteler and Callum Heggart</i>	
Estimation of the crustal velocity field in Baza and Galera faults from GPS position time series in 2009-2012	146
<i>Antonio J. Gil, et.al.</i>	
Advanced Symbolic Time Series Analysis in Cyber Physical Systems	155
<i>Roland Ritt, Paul O'Leary, Christopher Josef Rothschedl and Matthew Harker</i>	
A Non-stationary Index-flood Model With Local Likelihood Smoothing for Drought Assessment	162
<i>Filip Strnad, Martin Hanel, Vojtěch Moravec and Adam Vizina</i>	
Forecasting of Demand on Raw for Dairy Products	173
<i>Marina Arkhipova, Viacheslav Sirotin and Kirill Arkhipov</i>	
Spark and Solr: a powerful and ergonomic combination for online search in the Big Data environment (case of the UAE)	181
<i>Karim Aoulad Abdelouarit, Boubker Sbihi and Noura Aknin</i>	
Dynamical evolution of the community structure of complex network inherent in seismic time series	192
<i>Norikazu Suzuki</i>	
Quantitative characterization of intracellular calcium signals	195
<i>Iker Malaina, Carlos Bringas, Alberto Pérez-Samartín, Luis Martinez and Ildefonso Martínez de La Fuente</i>	
Real-Time Radioactive Precursor of the April 16, 2016 Mw 7.8 Earthquake in Ecuador ...	207
<i>Theofilos Toulkeridis, Fernando Mato, Katerina Toulkeridis-Estrella, Juan Carlos Perez Salinas, Santiago Tapia and Walter Fuertes</i>	
Local selection of learning data for neural networks in prediction of PM10 pollution	220
<i>Krzysztof Siwek and Stanislaw Osowski</i>	
Intelligent approach to vehicle routes planning base on artificial neural networks prediction model	232
<i>Daniel Kubek and Paweł Wiecek</i>	
Electricity price forecasting using a hybrid time series model	246
<i>Büşra Taş and Ceylan Yozgatligil</i>	
Forecasting Intraday Risk Measures using Multiplicative Component GARCH Model and Multimodal Distributions	249
<i>Aymeric Thibault and Pascal Bondon</i>	
Astronomical Time Delay Estimations	254
<i>Mariko Kimura, Hyungsuk Tak and Taichi Kato</i>	

Period Analysis in Astronomy by using Lasso	266
<i>Keisuke Isogai</i>	
Analyzing Spatial Dissimilarities via Effective-Time Series	270
<i>Madalina Olteanu and Julien Randon-Furling</i>	
Sequential Motor Unit Number estimation	282
<i>Peter Ridall</i>	
The electricity consumption in selected sectors of the Polish economy	295
<i>Marek Kott</i>	

Bio-medical Time Series Analysis

A GIS-based Model for Cholera Forecast	305
<i>Dau Xuan Hoang and Thi Ngc Anh Le</i>	
Correlation Dimension Estimation from EEG Time Series for Alzheimer Disease Diagnostics	316
<i>Martin Dlask and Jaromir Kukal</i>	
An application of the GAM-PCA-VAR model to respiratory disease and air pollution data	319
<i>Márton Ispány, Juliana Bottoni de Souza, Valderio A. Reisen, Glauro C. Franco, Pascal Bondon and Jane Meri Santos</i>	

Chaos and Random in Time Series

Cryptanalysis of a Random Number Generator Based on a Chaotic Ring Oscillator	321
<i>Salih Ergun</i>	
Factors Affecting Randomness in Pseudo-Random Number Series Extracted from Chaotic Time Series of Logistic Map and Chaos Neural Network	331
<i>Hitoaki Yoshida, Masatomo Sasaki, Takeshi Murakami, Shogo Shimono and Satoshi Kawamura</i>	

Computational Intelligence methods for Time Series

Exploring a century of Savoy history using hidden-Markov models with Beta-inflated distributions	343
<i>Julien Alerini and Madalina Olteanu</i>	
Comparing Three Time Series Segmentation Methods via Novel Evaluation Criteria	355
<i>Huynh Thi Thu Thuy, Vo Thi Ngoc Chau and Duong Tuan Anh</i>	
Eigenvalues distribution limit of covariance matrices with AR processes entries	367
<i>Zahira Khettab and Tahar Mourid</i>	
An Incremental von Mises Mixture Framework for Modelling Human Activity Streaming Data	379
<i>Eris Chinellato, Kanti Mardia, David Hogg and Anthony G. Cohn</i>	
Simulation of Defect Prediction over Time in Building Façade	390
<i>Woo-Ram Kim, Kichang Jeong, Yongdeok Jeon, Jinhong Park, Heeyoung Jeong and Jae-Seob Lee</i>	

Signal Classification using Covariance Matrices: A Riemannian Geometry Framework.....	400
<i>Shaelyn G. Divins, Joshua S. Beard, Nenad Mijatovic, Anthony O. Smith, Adrian M. Peter, Dean A. Clauter and Rana Haber</i>	
Combining Support Vector Regression with Scaling Methods for Highway Tollgates Travel Time and Volume Predictions	411
<i>Amanda Yan Lin, Mengcheng Zhang and Selpi Selpi</i>	

Data preprocessing methods: Data decomposition, seasonal adjustment, singular spectrum analysis, detrending methods

Comparative analysis of criteria for filtering time series of word usage frequencies.....	422
<i>Inna Belashova and Vladimir Bochkarev</i>	
Educational Data Mining: A Case Study of Data Pre-Processing and Investigation of Students' Academic Achievement for Artificial Intelligence Classifier	432
<i>Usamah Mat and Norlida Bunyamin</i>	
Telescope: A Hybrid Forecast Method for Univariate Time Series.....	444
<i>Marwin Züfle, André Bauer, Nikolas Herbst, Valentin Curtef and Samuel Kounev</i>	
The analysis of variability of short data sets based on Mahalanobis distance calculation and surrogate time series testing	452
<i>Teimuraz Matcharashvili, Natalia Zhukova, Tamaz Chelidze, Evgeni Baratashvili, Tamar Matcharashvili and Manana Janiashvili</i>	
Rainfall Measurements from Commercial Cellular Networks	463
<i>Reason L. Machete, Leonard A. Smith and Nnyaladzi Batisani</i>	
Understanding Instantaneous frequency detection: A discussion of Hilbert-Huang Transform versus Wavelet Transform	474
<i>Maximiliano Bueno Lopez, Marta Molinas and Geir Kulia</i>	

Deep Learning and Time Series Analysis

Deep Learning for Detection of BGP Anomalies	487
<i>Marijana Cosovic, Slobodan Obradovic and Emina Junuz</i>	
Abnormal State Prediction based on Deep Learning using Multiple Time Series Production Process Data.....	499
<i>Shigeru Fujimura and Wen Song</i>	
Human Gait Recognition by Deep Convolutional Activation Feature of Recurrence Plot for Accelerometer Time Series.....	503
<i>Yusuke Manabe</i>	

Dimensionality reduction and Similarity measures for Time series data analysis and its applications

Design Aircraft Engine Bivariate Data Phases using Change-Point Detection Method and Self-Organizing Maps.....	512
<i>Cynthia Faure, Jean-Marc Bardet, Jérôme Lacaille and Madalina Olteanu</i>	

Linear Trend Filtering via Adaptive Lasso.....	524
<i>Matus Maciak</i>	
A novel genetic algorithm based similarity measure for time series classification.....	536
<i>Basabi Chakraborty and Sho Yoshida</i>	
A time series clustering technique to analyze the stock market movement after the budget announcement.....	548
<i>Arup Mitra, Saptarsi Goswami, Basabi Chakraborty, Arun Jalan and Amlan Chakrabarti</i>	
An Efficient Anomaly Detection in Quasi Periodic Time-series Data - A Case Study with ECG.....	563
<i>Goutam Chakraborty, Takuya Kamiyama, Hideyuki Takahashi and Tetsuo Kinoshita</i>	
New Hybrid Feature Selection Algorithm based on Consistency Measures and Simulated Annealing Search.....	575
<i>Adrian Pino Angulo, Kilho Shin and Takako Hashimoto</i>	
On methods to assess the significance of community structure in networks of financial time series.....	585
<i>Argimiro Arratia and Martí Renedo</i>	
Minimizing the Number of Probes and Maximizing Classification Performance for P300 BCI speller.....	597
<i>Weilun Wang, Horie Shigeki and Goutam Chakraborty</i>	

Econometric Forecasting

Untangling the inefficiency of hotel industry: the Portuguese Teixeira Duarte Hotel chain analysis.....	609
<i>Nuno Ferreira and Manuela de Oliveira</i>	
Determining macroeconomic indicators to implement a short-term forecasting model for VAT revenue.....	616
<i>Maria Del Camino Gonzalez Vasco and Cesar Pérez Lopez</i>	
Combining forecasts to capture realized volatility dynamics.....	639
<i>Danilo Carità, Giovanni De Luca and Giampiero M. Gallo</i>	
Time series and artificial intelligence with genetic algorithms hybrid approach for rare earths price prediction.....	649
<i>Fernando Sanchez Lasheras, Sergio Luis Suárez Gómez, Maria Victoria Riesgo García, Alicia Krzemień and Ana Suárez Sánchez</i>	
Predicting the financial status of companies using data balancing and classification methods.....	661
<i>Huthaifa Aljawazneh, Antonio Mora García and Pedro Castillo Valdivieso</i>	

Econometric models

Change Point Detection in Autoregression Without Variability Estimation.....	674
<i>Barbora Pestova and Michal Pesta</i>	

Distance Between VAR Models and its Application to Spatial Differences Analysis in the Relationship GDP - Unemployment Growth Rate in Europe	686
<i>Francesca di Iorio and Umberto Triacca</i>	
A least-squares approach to estimate the impulse-response function of a general linear model	696
<i>Miguel Jerez and Alfredo Garcia-Hiernaux</i>	
Recovering the background noise of a Levy-driven CARMA process using an SDDE approach	707
<i>Mikkel Slot Nielsen and Victor Rohde</i>	

Energy Forecasting

Fuel Consumption Estimation for Climbing Phase	719
<i>Jingjie Chen and Yongping Zhang</i>	
Energy Prediction of Access Points in Wi-Fi Networks Using Time Series Modeling	730
<i>David Rodríguez Lozano, Juan A. Gomez-Pulido and Arturo Durán Domínguez</i>	
A Combination of Variational Mode Decomposition with Neural Networks on Household Electricity Consumption Forecast	740
<i>Vanessa Haykal, Hubert Cardot and Nicolas Ragot</i>	
Nonparametric panel stationarity testing. An application to crude oil production	752
<i>Manuel Landajo, María José Presno and Paula Fernández González</i>	
Detection of temperature break point for gas storage	764
<i>Andrzej Szczurek, Andrzej Kielbik and Monika Maciejewska</i>	
An econometric analysis of the merit order effect in electricity spot price: the Germany case	774
<i>Francois Benhmad and Jacques Percebois</i>	
Pattern sequence similarity based techniques for wind speed forecasting	786
<i>Neeraj Bokde, Alicia Troncoso, Gualberto Asencio-Cortés, Kishore Kulat and Francisco Martínez-Álvarez</i>	
Improving the performance of machine learning models by integrating partly physical control response models in short-term forecasting of aggregated power system loads	795
<i>Pekka Koponen, Harri Niska and Reino Huusko</i>	

Ensemble forecasting

A new approach to nowcast economic time series using ensembles of hidden Markov and Arima models	807
<i>Álvaro Gómez-Losada and Panayotis Christidis</i>	
Ensemble Learning Framework for Predicting Project Cost Overrun Levels in Construction Procurement Auctions	809
<i>Hyosoo Moon, Trefor P. Williams and Moonseo Park</i>	

Time Series Forecasting applying Data Transformation and Neural Networks Ensembles ..	820
<i>German Gutierrez, M. Paz Sesmero Lorente and Araceli Sanchis</i>	

Forecasting Complex/Big data

Dynamics of Memory in Investor Attention to Energy Market	829
<i>Ravi Prakash Ranjan and Malay Bhattacharyya</i>	
Sparse Granger-Causal Network Learning via the Depth Wise Group LASSO – An Application of ADMM for Large Vector Autoregressions	841
<i>Ryan J. Kinnear and Ravi R. Mazumdar</i>	
Development of a Routing Procedure to Assist an Earth Systems Model with Long Term Coastal Discharge Predictions	853
<i>Josefine Wilms and Marcus Thatcher</i>	
Short-term Stream Flow Forecasting at Australian River Sites using Data-driven Regression Techniques	865
<i>Melise Steyn, Josefine Wilms, Willie Brink and Francois Smit</i>	
An Implementation of HMM Classier in High Dimensions Based on MapReduce	877
<i>Badreddine Benyacoub</i>	
Performance Analysis of Time Series Forecasting of Ebola Casualties Using Machine Learning Algorithms	885
<i>Manish Kumar Pandey and Karthikeyan Subbiah</i>	
Hidden Markov Models for monitoring Circadian Rhythmicity in Telemetric Activity Data	899
<i>Barbel Finkenstadt</i>	
Forecasting via Fokker-Planck using conditional probablilites	913
<i>Chris Montagnon</i>	
Forecasting of CO2 emissions based on Preprocessing Techniques	922
<i>Lida Barba, Guillermo Machado, Lorena Molina, Ana Congacha, Jorge Delgado and Lady Espinoza</i>	
Analysis of Buildings Energy Losses Using Smart Monitoring	939
<i>Nivine Attoue and Isam Shahrour</i>	
Forecasting UK House Prices During Turbulent Periods	946
<i>Alisa Yusupova and Efthymios Pavlidis</i>	
Impact of weather forecasting accuracy over the electric demand predictions quality	960
<i>Eduardo Caro, Jesús Juan and Paula Cernuda</i>	
A New Approach for Time Series Decomposition and Prediction	964
<i>Yading Yue, Guangan Zhuang, Rong Zhang, Jianchun Zhao and Lichun Liu</i>	
Short-term time series forecasting based on internal smoothing of Pade interpolants	974
<i>Minvydas Ragulskis, Kristina Lukoseviciute, Tadas Telksnys and Zenonas Navickas</i>	

Future of Mathematical and Logical Structures behind Time Series Analysis and History

The Dependence Structures of Non-Stationary Bivariate INAR(1) Processes: The Case of the Bivariate Poisson Innovations	985
<i>Naushad Mamode Khan, Yuvraj Sunecher and Vandna Jowaheer</i>	
Similarity Analysis of Time Interval Data Sets Regarding Time Shifts and Rescaling	995
<i>Marc Haßler, Sabina Jeschke and Tobias Meisen</i>	
Financial variables and the real economy: Evidence using a data based procedure of Simultaneous Structural Model Design	1007
<i>Roger Hammersland</i>	
Logical Comparison Measures in Classification of Data	1035
<i>Kalle Saastamoinen</i>	

Macroeconomic analysis

Macroeconomic Forecasting using Approximate Factor Models with Outliers	1047
<i>Ray Yeutien Chou, Tso-Jung Yen and Yu-Min Yen</i>	
Testing Granger-causality on macroeconomic time series: a bootstrap approach	1050
<i>Matteo Farnè and Angela Montanari</i>	
An implied rating software system	1054
<i>Ventsislav Nikolov</i>	
Fiscal Regime Shifts, and Household Expectations on Policy Dynamics	1064
<i>Diederik Kumps and Peter Claeys</i>	

Nonparametric and functional methods

Robust autocovariance estimation from the frequency domain	1073
<i>Higor Cotta, Valdério Reisen and Pascal Bondon</i>	
Event Related Causality analysis of electrocorticographic (ECoG) time series as diagnostic tool for epileptic surgery	1075
<i>Anna Korzeniewska, Piotr Franaszczuk and Nathan Crone</i>	
Sieves Estimators and Predictors for Functional Autoregressive Processes	1083
<i>Tahar Mourid and Nesrine Kara-Terki</i>	
Modeling of p-order persistent time series by the modified Langevin equation	1089
<i>Zbigniew Czechowski</i>	
Bootstrap confidence intervals for conditional density function in Markov processes	1094
<i>Inés Barbeito Cal, Ricardo Cao and Dimitris Politis</i>	
Forecasting with functional Time Series	1098
<i>Fatiha Messaci and Sara Leulmi</i>	
Time Series predictor based on deterministic and stochastic assumptions	1108
<i>Pedro Cadahia, José Manuel Bravo Caro, Manuel Emilio Gegundez-Arias and Antonio Golpe</i>	

Functional Data Classification by Discriminative Interpolation with Features	1120
<i>Rana Haber, Anand Rangarajan, Nenad Mijatovic, Anthony O. Smith and Adrian M. Peter</i>	

Nonstationarity Analysis in Time Series

A Modified EM Algorithm for Parameter Estimation in Linear Models with Time-Dependent Autoregressive and t-Distributed Errors	1132
<i>Boris Kargoll, Mohammad Omidalizarandi, Hamza Alkhatib and Wolf-Dieter Schuh</i>	
Copulas for Modeling the Relationship between the Inflation and the Exchange Rates	1146
<i>Laila Ait Hassou, Fadoua Badaoui, Cyrille Okou Guei, Amine Amar, Abdelhak Zoglat and Elhadj Ezzahid</i>	
Fractal analysis applied to light curves of pulsating stars	1157
<i>Sebastiano de Franciscis, Javier Pascual Granado, Juan Carlos Suárez and Rafael Garrido Haba</i>	

Recent Developments on Time-Series Modelling

Method for modeling and analysis of natural time series	1163
<i>Oksana Mandrikova, Nadezhda Fetisova and Yury Polozov</i>	
A New Estimation Technique for AR(1) Model with Long-tailed Symmetric Innovations ..	1175
<i>Aysen Dener Akkaya and Özlem Turker Bayrak</i>	
Modeling and analysis of the cosmic rays variations during periods of heliospheric disturbances on the basis of wavelet transform and neural networks	1185
<i>Oksana Mandrikova and Timur Zalyaev</i>	
Multidimensional Time-Frequency Analysis of the CAPM	1187
<i>Roman Mestre and Michel Terraza</i>	
Prediction of High-Dimensional Time-Series with Exogenous Variables Using Extended Koopman Operator Framework in Reproducing Kernel Hilbert Space	1206
<i>Jia-Chen Hua, Farzad Noorian, Philip H.W. Leong, Gemunu Gunaratne and Jorge Gonçalves</i>	

Structural Time Series Models

Nonlinear Dynamical Analysis of Twitter Time Series	1219
<i>Andrey V. Dmitriev, Vitaly Silchev, Victor Dmitriev and Svetlana Maltseva</i>	
Interpolation of ARMA processes with infinitely divisible white noise	1231
<i>Argimiro Arratia, Alejandra Cabaña and Enrique Cabaña</i>	
Analysis of time series of earthquake occurrence in Caucasus	1240
<i>T. Matcharashvili, N. Zhukova, E. Mepharidze, A. Sborshikov</i>	

Untangling the inefficiency of hotel industry: the Portuguese Teixeira Duarte Hotel chain analysis

Abstract. In this study the technical efficiency was analyzed for four hotels of the Teixeira Duarte Group - a Portuguese hotel chain. An efficiency ranking was established for these four Portuguese hotels units using Stochastic Frontier Analysis. This methodology allowed discriminating between measurement error and systematic inefficiencies, enabling the identification of the main inefficiency causes. The results showed that distance to the airport and the higher price of accommodations promote efficiency. Additionally, hotels with many standard rooms and sea views are likely to achieve higher levels of efficiency.

Keywords: Hotel industry; Efficiency; Stochastic Frontier Analysis (SFA)

1 Introduction

The industry of tourism has great strategic significance for the Portuguese economy due to its capability to generate wealth and to create employment opportunities (World Tourism Organization, 2011). In fact, this is an economic sector where Portugal has clear competitive advantages, due to the existing high-quality infrastructures, highly qualified human resources, and to natural diversity and pristine environments. Portugal has exceptional resources in terms of geographic location, temperate climate, security, historical and cultural heritage, high-quality beaches, natural diversity (of species and environments), and a competitive high quality coastal touristic development.

Concerning the Portuguese hotel sector studies and according to the Atlas Hospitality (2005), the touristic Portuguese market is highly segmented, with hotel groups owning 63.8% of integrated housing units, while the remaining 36.2% belong to independent entrepreneurs.

The hotel sector is an important component of the tourism industry, challenged by a competitive atmosphere managed by different pressure factors and driven by supply and demand (International Labour Office - GDFHTS/2010, 2010).

Teixeira Duarte (TD), a renowned Portuguese hotel chain, was founded in 1921 as a family company, and today is one of the Portuguese largest economic groups. Teixeira Duarte has a successful trajectory established through the sustainable growth in the civil construction sector. The expansion of the hotel industry in Portuguese-speaking countries were decisive factors to consolidate its privileged economic situation. In fact, outside Portugal, there are TD Hotels in the main cities of Angola and Mozambique, whereas in Portugal its main hotels are located in the southern region of the Algarve region, although several units can also be found on the coast of Alentejo (Southwest coast) and in the centre area of the country.

(see Teixeira Duarte in the World in the homepage of Teixeira Duarte, 2016).

Despite the competitiveness and excellence displayed by the Group, it is essential to guarantee that levels of performance are improved or, at least, maintained. In fact, for management purposes, maintaining efficiency implicates the use of scarcer inputs, and the production of additional outputs. It also means performing the assigned roles and preventing possible inaccuracies that can impede the progress of an industry.

In this context, the aim of this study is to analyze the efficiency of the four TD Group hotel units based in Portugal, namely: The Lagoas Park Hotel, the Sinerama Aparthotel, the Eva Hotel and the Oriental Hotel, in order to identify factors affecting efficiency, and analyze what must be altered to promote better performances.

This paper is organized as follows: section 2 presents the Teixeira Duarte Group; section 3 briefly reviews the literature; section 4 outlines the main steps in the chosen methodology; section 5 contains the experimental results; section 6 reviews the main conclusions and the limitations of the paper and suggests further work.

Teixeira Duarte (TD) Group.

The Teixeira Duarte Group currently employs more than 13,000 workers, and operates in 16 countries in seven different sectors such as construction, concessions and services, Real Estate, hotel services, distribution, energy, and automobile.

Business Indicators (Teixeira Duarte Group)	Year				
	2010	2011	2012	2013	2014
Average number of workers	13036	11182	10853	12011	13261
Turnover	1380	1200	1383	1581	1680
Operating income	1445	1263	1440	1630	1716
Net debt	1067	927	990	1176	1293
Total equity	562	333	326	361	485
Total net assets	2721	2753	2767	2779	2954

Table 1. - The main indicators of Teixeira Duarte Group's business (the book values are expressed in million euros. Total Equity includes non-controlled interests).

In non-consolidated terms, and in order to provide an overall view of the total activity of the TD Group during 2014, we disclose that its operating income in the construction sector reached the total value of 1,027,221€, reflecting an overall slight decrease of 0.7% regarding 2013 (source: TD Annual Reports (2012, 2013, 2014)).

Disregarding new contracts that may arise, the Group has already assured business levels in the construction sector for the foreign markets which, in spite of the current adverse circumstances of the domestic market, achieve 904,808€ for 2015. In this context the Hotel Service Business represents 4.7% of the entire Group's income (TD homepage, 2016).

After the first experience in 1974 in the Algarve, the Teixeira Duarte Group resumed its activity in the Hotel Services sector in the 1990's in Sines, and currently operates 10 hotels, four of which are located in Portugal, three in Angola and three in Mozambique, covering a total of 2,908 beds and 1,465 rooms. According to TD Group, their services are based on Tradition, Quality, Comfort and Kindness (see hotel services from TD homepage, 2016).

Hotels in Africa		Hotels in Portugal
Angola	Mozambique	
Hotel Alvalade, Luanda	Hotel Avenida, Maputo	Hotel Eva, Faro
Hotel Baía, Luanda	Hotel Tivoli Maputo, Maputo	Hotel Oriental, Portimão
Hotel Trópico, Luanda	Tivoli Hotel Beira, Beira	Lagoas Park Hotel, Oeiras
		Sinerana, Sines

Table 2. – Location (city and country) of the hotels of the Teixeira Duarte (TD) Group.

Eva Hotel (4-star hotel) is acknowledged as a benchmark of quality in Faro, both for leisure or business stays. The hotel was recently renovated in order to be architectonically integrated into the historical and commercial downtown area of Faro. The Oriental Hotel (4-star hotel), with a characteristic oriental style, is situated in one of the most popular sun and sea Portuguese touristic destination. The Lagoas Park Hotel (4-star hotel) is located in one of the largest business centres of the country, providing all conditions needed for business meetings and for leisure, given its congress centre and its privileged location, fairly close to the beaches of Cascais, to Sintra, as well as to several other interesting touristic sites. Sinerama Hotel (3-star hotel) is located in Bay of Sines, in the vicinities of the Castle of Sines, and of the Vasco da Gama Museum. The hotel provides a family and quiet environment (www.tdhotels.com/pt).

2 Literature review

The efficiency analysis within the touristic hotel sector has been widely studied over the years. Among available literature on this subject, the Stochastic Frontier Analysis (SFA) methodology must be emphasised as a frequent approach. In fact, authors such as Anderson *et al.* (1999a) analysed the estimation of the managerial efficiency of 48 hotels in the USA during 1994. In a subsequent paper, Anderson *et al.* (1999b) applied both Data Envelopment Analysis (DEA) and SFA to estimate the efficiency of 31 corporate

travel management departments. Also, Wang *et al.* (2007) used a one-stage SFA approach to analyze technical efficiency of 66 international hotels in Taiwan from 1992 to 2002, and also incorporated the Malmquist productivity index in the results. Likewise, Chen (2007) examined the cost efficiency of 55 international hotels in Taiwan using an SFA model.

In fact, the use of the SFA approach can be found in many other studies, and often combined with other methodologies (e.g. Wang *et al.*, 2007; Pérez-Rodríguez & Acosta-González, 2007; Assaf *et al.*, 2010 and Hu *et al.*, 2010).

The approaches to hotel industry efficiency analysis can be driven by different perspectives, for example, Narayan & Sharma (2013) analyzed different tourist markets and the relationship between hotel industry and its macroeconomic contribution (e.g. Kreishan, 2010; Assaf & Josiassen, 2012; Assaf & Barros, 2011; Hathroubi *et al.*, 2014 and Jarboui *et al.*, 2015). Concerning the Portuguese hotel industry, the efficiency analysis has been addressing by different approaches mostly by Barros (e.g. Barros, 2004; Barros & Alves, 2004; Barros & Santos, 2004; Barros, 2005a,b; Barros & Mascarenhas, 2004,2006; Barros & Santos, 2006; Barros *et al.*, 2010; Barros & Machado, 2010; Barros *et al.*, 2011; Oliveira & Pedro, 2013, 2014; and Oliveira *et al.*, 2013).

The issues approached vary between the economic efficiency analysis and the determinants that contribute to the economic efficiency. For instance, Barros *et al.* (2011) used a DEA model to estimate the efficiency determinants of Portuguese hotel groups from 1998 to 2005. Barros *et al.* (2010) analysed the length of stay off golf tourists in the Algarve whereas Barros and Machado (2010) contributed to the relevant literature by analysing the determinants of the length of stay, in this instance, of foreign tourists in Madeira Island.

Concerning the Oliveira & Pedro (2013, 2014) work, the authors studied a sample of 28 prestige hotels in the Algarve (Portugal) to compare both DEA and SFA approaches in order to measure cost, allocative and technical efficiencies. Oliveira *et al.* (2013) use DEA to investigate and compare the efficiency of Portuguese hotels in Algarve in terms of the influence of star ratings, golf courses and location on hotel efficiency. Therefore, both methods are widely applied to this economic sector, supporting the methodology choice for the present study.

3 Methods and materials

Dataset

For the stochastic frontier analysis, the data collected from Teixeira Duarte Group database comprises data from 01/01/2011 to 30/06/2015 (Table 1), and relates only to the Portuguese Hotels, to incorporate hotels facing similar seasonality patterns and having common operational periods and homogenous quality of services. A total of 216 observations was gathered, corresponding to the 54 months (since January 2011 to June 2015) per hotel. The chosen output variable was the Operating profits. Table 3 defines the remaining inputs and exogenous variables.

TD Hotels SFA model	
Output	
	Operating profits (euros)
Inputs	
	Operating costs (euros)
	Employees (number)
Exogenous variables	
	Lodging price-range (euros)
	Standard rooms (number)
	Sea view (0=no; 1=yes)
	Airport distance (Kms)

Table 3. - *Output, Inputs and Exogenous variables used in the stochastic frontier model.*

Data analysis

Using a stochastic frontier model, where each Decision Maker Unit (DMU) is denoted by i , the individual operating profit is obtained using the following production function (Battese & Coelli, 1995):

$$\ln(y_i) = x_i\beta + (v_i - u_i) \quad (1)$$

where $i = 1, 2, \dots, N$; y_i measures the operating profits of the i th hotel; x_i is a $1 \times K$ vector corresponding to the inputs (operating costs and employees); and β is a $1 \times K$ vector of unknown scalar parameters to be estimated. For this model, the traditional error term ε is composed of two distinct terms ($v_i - u_i$) for each DMU where the error term v_i , similarly to traditional regression models, is assumed to be independent and identically distributed as $N(0, \sigma_v^2)$. Random variation in output caused by factors beyond DMUs control, such as measurement errors in dependent variables or explanatory variables eventually omitted, is captured by the v_i error term. The error term u_i is a non-negative random variable, accounting for the existence of technical inefficiency in production following a half-normal $u_i \sim |N(0, \sigma_u^2)|$ distribution.

According to Battese & Coelli (1995), the inefficiency distribution parameter can also be specified as the inefficiency model:

$$u_i = \delta_0 + z_i\delta + \omega_i \quad (2)$$

where δ represents a vector of parameters to be estimated, z_i is a vector of DMU specific effects (lodging price range, standard room, the existence of sea view and airport distance), that determine technical inefficiency, and ω_i is distributed following $N(0, \sigma_w^2)$. All observations either lie on, or are beneath, the stochastic production frontier, and this is assured by $u_i \geq 0$ in Equation (2). The variance terms are parameterized by replacing σ_v^2 and σ_u^2 with $\sigma^2 = \sigma_v^2 + \sigma_u^2$ and $\gamma = \frac{\sigma_u^2}{(\sigma_v^2 + \sigma_u^2)}$ according to Battese & Coelli (1995). The value of γ ranges between 0 and 1, where 1 indicates that all of the deviation from the frontier is entirely due to technical inefficiency (Coelli et al., 1998). The technical efficiency (TE) of each DMU is expressed as follows:

$$TE_i = \frac{E(Y_i|u_i, X_i)}{E(Y_i|u_i=0, X_i)} = e^{-u_i} \quad (3)$$

where E is the expectation operator; thus, the measure of technical efficiency is based on a conditional expectation given by Equation (3), considering that the value of $v_i - u_i$ evaluated at the maximum value of Y_i is conditional on $u_i = 0$ (Battese & Coelli, 1995).

The parameters of the stochastic frontier model (1) and the technical inefficiency model (2) were estimated using the FRONTIER version 4.1 software (Coelli, 1996).

4 Results

The SFA model results confirms that the inclusion of the inefficiency effects is highly significant (at the 1% significance level) in the analysis of Operating Profits (the estimate for the variance is close to one – $\gamma = 0.999$ in Table 5), indicating that 99.9% of the random variation in Operating Profit is due to inefficiency.

The mean efficiency of the four hotel units is presented in Table 4, and indicates that the Lagoas Park Hotel is the most efficient hotel unit, contrasting with the Oriental Hotel (the least efficient one).

Analyzing the yearly evolution, Eva and Lagoas Park recorded an increase in efficiency. Nevertheless, for the last one the efficiency level has decreased slightly during the analyzed 6 months of 2015.

Hotel	2011	2012	2013	2014	2015	mean efficiency per hotel
Eva	0.542	0.545	0.558	0.602	0.625	0.569
Lagoas	0.597	0.633	0.643	0.669	0.650	0.637
Oriental	0.414	0.411	0.465	0.502	0.485	0.453
Sinerama	0.662	0.632	0.524	0.570	0.418	0.577
TD hotels' mean efficiency per year	0.556	0.559	0.547	0.586	0.545	0.560

Table 4. – Mean efficiency scores per hotel unit and per year (from 2011 to 2015) of the Portuguese hotels of Teixeira Duarte (TD) Group.

The SFA and the inefficiency models results are presented in Table 5.

Variable	Coef.	Std. Error
Stochastic frontier model		
constant	2.790 **	0.116
ln(operating costs)	0.841 **	0.278
ln(employees)	-0.343 **	0.126
Inefficiency model		
constant	-0.125 **	0.014
Lodging price-range	-0.042 **	0.010
Standard rooms	-0.024 **	0.008
Sea view	-0.034 **	0.008
Airport distance	10.953 **	1.603
Variance parameter		
γ	0.999 **	0.000

** significant at 1%.

Table 5. - The results of the SFA and of the inefficiency models from 2011 to 2015 for the Portuguese hotels of Teixeira Duarte (TD) Group.

Moreover, it must be emphasized that the Sinerama Hotel has been losing efficiency since 2011, whereas the Oriental Hotel did not indicate any pattern regarding the variation in the efficiency levels from 2011 to 2015. In both models all variables are statistically significant at the 1% significance level. The SFA results indicate that the hotels with higher “operating costs” and less “employees” are the ones that achieved higher operating profits. This could seem contradictory but in fact, a hotel unit with high operating costs could mean that have plenty of services (or just services with high quality) to their customers rather than just a high inefficient operation. In the case of the studied hotels, “Oriental” is by far the best hotel unit of the group, praised frequently by guests who refer the five-star’s service quality. Additionally, the number of employees if variable along the year. Some of them are fixed through the twelve months to ensure the quality services with their experience whereas some others are recruited only in the high seasons (Carnival, Easter, Summer and Christmas festivities). In this scenario wins the hotel unit with the smallest but better-trained team that will coach the seasonal teams and overall, be able to ensure the best service along the entire year. Concerning the Inefficiency model, the “airport distance” is the most important factor that contributes to inefficiency (highlighted by the positive coefficient). This is not a surprising result since the proximity of the airport affects the transportation’s cost to the hotel unit. The shorter the distance the smallest the cost (with plenty of transport facilities alternatives). In this context, “Sinerana” is definitely in disadvantage however, the site in which this hotel is inserted is very particular (Natural Park of South-West Alentejo and the Costa Vicentina), meaning that their customers are determined to go there despite the distance. Notwithstanding, in comparison with the remaining, this unit is underprivileged by airport’s distance. The authors of Pedro and Marques (2013) in their study of Portuguese hotels in Algarve analyzed the influence of star ratings, golf courses and location on hotel efficiency. Their conclusions pointed that star rating is not a significant determinant of efficiency but the location and the existence of golf courses may have some relevance. In this case, those hotels that do not possess golf courses are the more efficient, confirming that frequently guests chose hotel units for specific purposes (such location) and not only by quality services or transportation facilities. In a previous study of Barros and Santos (2006) the economic efficiency of the biggest Portuguese hotel chains (15 hotels observed from 1998 to 2002) was analyzed concluding that scale is the main factor in explaining hotel efficiency and did not found any specific regional or property characteristics affecting the results. These contradictory studies reveal that this sector is very sensitive to changes being supported by the guests’ opinion and preferences rather than facilities, price ranges or location.

Concerning the present study and the price range, the “lodging price range” revealed a negative coefficient (a positive impact), meaning that more high prices contribute positively to efficiency.

Similarly, a hotel with many “standard rooms” and “sea view” also achieves higher levels of efficiency.

In one of the more recent works, Barros et al. (2010) addressed to Algarve’s golf hotel units concluded that elderly golfers might stay for an extended period and the type of hotel in which the golf tourists are accommodated can also increase the length of stay. In this case the proximity and quality of beaches were not of relevance to this type of tourist.

Therefore, a plausible general conclusion is that revenue efficiency determinants can change from one hotel unit to another. What seems to be important and relevant to some guests can be irrelevant to others. This legitimises our junction of different hotel unit type, considering the main purpose of assessing the TD Group hotel chain efficiency.

5 Conclusions and Final Remarks

The aim of the present research was to evaluate the efficiency of the Teixeira Duarte hotel chain in Portuguese mainland. The use of SFA allowed to assess the level of efficiency of each DMU (hotel unit), and simultaneously to highlight the factors that significantly affect the performance of the hotel units.

The achieved results showed that an efficient hotel should be placed in a location in the vicinity of an airport, and be equipped with standard rooms and – preferentially – sea view. High lodging price range revealed not to be a problem to efficiency levels, since high prices favour efficiency improvement.

However, as it was also presented, different authors concluded different determinants for efficiency. Additionally, the authors of the present study are aware of its limitations. Firstly, and more important, the small time-window (yet the available one) secondly, the limited number of hotel units (but again, there were included all the TD Group hotel units in Portugal mainland). Besides this limitation we observed also that the number of hotels included in this study has different dimensions, production characteristics, and locations turning possible biased comparisons. Finally, the reduced number of covered factors or determinants. Indeed, it would be an asset to this analysis to add some additional factors regarding tourist experience valuation, such as satisfaction and length of stay (Assaf & Josiassen, 2012); (Barros & Machado, 2010).

Despite these limitations we think that the present study could highlight some interesting results concerning this particular Portuguese hotel chain. Notwithstanding, results should be carefully considered in the management strategies adopted by the TD Hotel Group.

6 References

1. Anderson, R.I., Fish, M., Xia, Y. & Michello, F. (1999b). Measuring efficiency in the hotel industry: a stochastic approach. *International Journal of Hospitality Management*, 18, 45-57.
2. Anderson, R.I., Lewis, D. & Parker, M.E. (1999a). Another look at the efficiency of corporate travel management departments. *Journal of Travel Research*, 37, 267-272.
3. Assaf, A. & Josiassen, A. (2012). Time-varying production efficiency in the health care foodservice industry: a Bayesian method? *Journal of Business Research*, 65(5), 617-625.
4. Assaf, A., Barros, C.P. & Josiassen, A. (2010). Hotel efficiency: A bootstrapped metafrontier approach. *International Journal of Hospitality Management*, 29(3), 468-475.
5. Assaf, A.G. & Barros, C. (2011). Bayesian cost efficiency of Luanda, Angola hotels. *The Service Industries Journal*, 31(9), 1549-1559.
6. Atlas Hospitality (2015). Available at http://atlasdahotelaria.com/2015/downloads/deloitte_atlas_da_hotelaria_2015_web_en.pdf. Consulted on February 19, 2016.
7. Barros, C.P. & Alves, P. (2004). Productivity in the tourism industry. *International Advances in Economic Research*, 10, 215-225.
8. Barros, C.P. & Santos, C.A., (2006). The measurement of efficiency in Portuguese hotels using data envelopment analysis. *Journal of Hospitality & Tourism Research*, 30(3), 378-400.
9. Barros, C.P., Butler, R. & Correia, A. (2010). The length of stay of golf tourism: A survival analysis. *Tourism Management*, 31, 13-21.
10. Barros, C.P., Botti, L., Peypoch, N. & Solonandrasana, B. (2011). Managerial efficiency and hospitality industry: the Portuguese case. *Applied Economics*, 43, 2895-2905.
11. Barros, C.P. & Machado, L.P. (2010). The Length of Stay in Tourism, *Annals of Tourism Research*, 37(3), 692-706.
12. Barros, C.P. & Mascarenhas, M.J. (2004). Technical and allocative efficiency in a chain of small hotels. *International Journal of Hospitality Management*, 24(3), 415-436.
13. Barros, C.P. & Machado, L. P. (2011). The Length of Stay in Tourism, *Annals of Tourism Research*, 37(3), 692-706.
14. Barros, C.P. & Mascarenhas, M.J. (2006). The measurement of efficiency in Portuguese hotels with DEA. *Journal of Hospitality & Tourism Research*, 30(3), 378-400.
15. Barros, C.P. (2004). A stochastic cost frontier in the Portuguese hotel industry. *Tourism Economics*, 10, 177-192.
16. Barros, C.P. (2005a). Measuring efficiency in the hotels: An illustrative example. *Annals of Tourism Research*, 32(2), 456-477.

17. Barros, C.P. (2005b). Evaluating the efficiency of small hotel chain with a Malmquist productivity index. *International Journal of Tourism Research*, 7(3), 173-184.
18. Battese, G.E. & Coelli, T.J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20, 325-332.
19. Chen, C.F. (2007). Applying the stochastic frontier approach to measure hotel managerial efficiency in Taiwan. *Tourism Management*, 28, 696-702.
20. Coelli, T., Prasada, R. & Battese, G. (1998). An introduction to efficiency and productivity analysis. Boston, Massachusetts, USA: Kluwer Academic Press.
21. Coelli, T.J. (1996). A guide to FRONTIER version 4.1: a computer program for stochastic frontier production and cost function estimation.
22. Hathroubi, S., Peypoch, N. & Robinot, E. (2014). Technical efficiency and environmental management: The Tunisian case. *Journal of Hospitality and Tourism Management*, 21, 27-33.
23. Hu, J.L., Chiu, C.N., Shieh, H.S. & Huang, C.H. (2010). A stochastic cost efficiency analysis of international tourist hotels in Taiwan. *International Journal of Hospitality Management*, 29, 99-107.
24. International Labour Office - Developments and challenges in the hospitality and tourism sector - Issues paper for discussion at the Global Dialogue Forum for the Hotels, Catering, Tourism Sector, Genova (2010). Available at http://www.ilo.org/wcmsp5/groups/public/---ed_dialogue/---sector/documents/meetingdocument/wcms_162202.pdf
25. Jarboui, S., Guetat, H. & Boujelbène, Y. (2015). Evaluation of hotels performance and corporate governance mechanisms: Empirical evidence from Tunisian context. *Journal of Hospitality and Tourism Management*, 25, 30-37.
26. Kreishan, F. M. (2011). Time-series evidence for tourism-led growth hypothesis: A case study of Jordan. *International Management Review*, 7(1), 89-93.
27. Narayan, P. & Sharma, S. (2013). Does tourism predict macroeconomic performance in Pacific Island countries? *Economic Modelling*. Elsevier.
28. Oliveira, R. and Pedro, M.I. (2014). Cost efficiency of Portuguese hotels in the Algarve: a comparative analysis using mathematical and econometric approaches, *Tourism Economics*, 20(4), 797-812.
29. Oliveira, R., Isabel Pedro, M. & Cunha Marques, R. (2013). Efficiency and its determinants in Portuguese hotels in the Algarve. *Tourism Management*, 36, 641-649.
30. Pérez-Rodríguez, J.V. & Acosta-González, E. (2007). Cost efficiency of the lodging industry in the tourist destination of Gran Canaria (Spain). *Tourism Management*, 28, 993-1005.
31. TD Annual Reports (2012, 2013, 2014). Available at <http://www.teixeiraduarte.pt/investors/financial-information/annual-reports.html>
32. Teixeira Duarte hotel-services (2016). Available at <http://www.teixeiraduarte.pt/business-sectors/hotel-services.html>
33. Teixeira Duarte in the World in Teixeira Duarte's homepage (2016). Available at <http://www.teixeiraduarte.pt/group/teixeira-duarte-in-the-world.html>
34. Wang, Y.H., Lee, W.F. & Wong, C.C., (2007), Productivity and efficiency analysis of international tourist hotels in Taiwan: an application of the stochastic frontier approach. *Taiwan Economic Review*, 35, 87-114.
35. World Tourism Organization (2011). Investing in energy and resource efficiency. World Tourism Organization. Available at http://www.unep.org/resourceefficiency/Portals/24147/scp/business/tourism/greenecomony_tourism.pdf

Determining macroeconomic indicators to implement a short-term forecasting model for VAT revenue.

César Perez López and Camino González Vasco

Cesar.perez@ief.minhap.es
Camino.gonzalez@ief.minhap.es

Institute for Fiscal Studies
Spain

February, 2017

ABSTRACT:

Macroeconomic indicators are a good source of information for short-term forecasting due to several reasons: they cover different areas of the economy and provide faster modes of dissemination. In this study we use a set of indicators to obtain a valid forecast for VAT revenue using a blend of statistical methods such as transfer functions and principal component analysis. The objective is to enforce parsimony and avoid multicollinearity problems with minimum information loss.

We apply the proposed method to quarterly data beginning in 1995 and ending in 2016, providing out of sample estimations for the four quarters of 2017.

Keywords: Principal Components Regression, VAT forecasting, forecast combination, generated regressors.

JEL classification numbers: E62, C51, H68, C22.

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Spanish Institute of Fiscal Studies.

1. Introduction

Forecasting of revenues is an issue of crucial relevance to governments in ensuring stability in tax and expenditure policies. Just as demand analysis and forecasting in the private sector is of critical importance because sales sustain the financial health of business, adequate and predictable tax and non-tax revenues underpin the financial sustainability and stability of government.

The importance of revenue forecasting in public budgeting has increased with governments shifting from annual cash-based budgets to medium-term budgeting as fiscal policy design and implementation have paid more attention to medium term constraints and the importance of budgeting for multiyear financial commitments. To address these goals, many countries have built a Medium-Term Budgetary Framework (MTBF)¹, a system of projections tailored to obtain values of revenues and expenditures in future periods.

These MTBF also serve as meaningful tools that provide a starting point in addressing compliance problems and supporting evasion deterrents. The main drawback of these prediction systems is that their construction involves the use of techniques related to time series theory, transfer function models and multivariate analysis.

In particular, regarding to Value-Added Tax models, there are several approaches in the literature that either calculate or forecast the VAT revenue.

Most of the models that compute tax revenue of VAT are focused on estimating the VAT base. Prominent examples of such strategy are the models based on the National Accounts approach, used by the U.S. Department of the Treasury and by the International Monetary Fund² for numerous countries, the models based on the Sectoral Approach, and the Input-Output Models approach that is also used for simulation³.

If we focus on VAT revenue forecasting, we often find in the literature methodologies grounded on the GDP based tax forecasting models. As a first step, the models require the construction of data series for tax revenues and their bases for each tax. All these tax bases are assumed to be predetermined and are obtained from macroeconomic variables derived from national accounts and balance of payments aggregates. These historical data series of tax revenues have embedded in them the effects of increases in national income or expenditures, as well as discretionary changes made in the tax system over time. For the VAT revenue model we present in this paper, these changes brought about by discretionary changes are introduced by dummy variables.

¹ See the IMF working paper "Medium-Term Budgetary Frameworks - Lessons for Austria from International Experience" by Erik J. Lundback

² See e.g. H.H. Zee and J.P. Boding "Aspects of introducing a Value-Added Tax in Sri Lanka" paper prepared for the International Monetary Fund, Fiscal Affairs Department, (August 1992)

³ See "Tax Analysis and Revenue Forecasting-Issues and Techniques" – by Glenn P. Jenkins, Chu-Yan Kuo and Gangadhar P. Shukla, Harvard Institute for International Development, Harvard University, for a detailed description of these models.

The next step for setting up the GDP based forecasting models is to establish an exact relationship between the tax revenue and the economic variables (ie proxy base). In order to do this, it is necessary to determine the correct base for each tax using the *National Accounts*. Subsequently, it is necessary to find out which component of the *National Account* corresponds most closely to the base for a particular tax.

In the case of Value-Added Tax, tax revenues are linked with *Total Consumption Expenditure on Goods and Services*⁴. This could be written as a transfer function and a regression analysis is carried out to forecast future revenue collections.

Obviously the predictive ability of this type of models is limited and error margins are large. This is partly because tax revenues are highly sensitive to a wide variety of economic variables and specifically to the economic cycle, and our ability to forecast the path of the economy using only one explanatory variable in a transfer function is restricted.

In order to address this problem, we could explore the possibility of introducing additional variables covering different areas of the economy (*Domestic Demand, Labour Market* and *Activity Indicators*, among others), but the high degree of linear dependency among this indicators would cause multicollinearity in the model.

Therefore, we propose principal component analysis applied to the entire set of numerical independent variables, to provide orthogonal regressors for the transfer function, ensuring the lack of multicollinearity with little information loss and increasing the forecasting accuracy.

This approach has the advantage of considering the behavioral responses of certain economic sectors (such as *Tourism Industry* or *Construction*) to the introduction of changes in the existing tax laws, and reciprocally, it is able to capture the influence of a decrease in one specific sector on the VAT revenue.

This paper is organized as follows. Section II outlines the derivation of the model employed and describes the estimation technique and the empirical framework. Section III presents the data set. Section IV shows the estimation results. The last section provides the main conclusions of this study.

2- Estimation Strategy.

Starting from a set of indicators relative to different areas of the economy (*Construction and Services Activity Indicators, Private Consumption variables, Labour Market Indicators* and *External Trade Indicators*) we propose a principal component analysis as a dimension reduction technique for the set of independent variables. The next step is to use the first and second principal components as inputs to the transfer function to estimate the VAT revenue.

⁴ The Institute for Fiscal Studies's public finance forecasting model, which was used to produce forecasts in Britain in each Green Budget up to 2013, was based on the assumption that VAT revenues grew in line with nominal consumer spending. For further information see "Forecasting the PSBR Outside Government: The IFS Perspective" by Christopher Giles and John Hall, Fiscal Studies Volume 19, Issue 1 (February, 1998)

The ultimate goal in principal components analysis is to find the minimum number of dimensions that are able to explain the largest variance contained in the initial set of indicators. We intend to simplify the information which gives us the correlation matrix to make it easier to interpret.

Principal component analysis was originated by Pearson (1901) and later developed by Hotelling (1933). The application of principal components is discussed by Rao (1964), Cooley and Lohnes (1971), and Gnanadesikan (1977). Exceptional statistical treatments of principal components are found in Kshirsagar (1972), Morrison (1976), and Mardia, Kent, and Bibby (1979).

Given a data set with p numeric variables, we can compute up to p principal components. Each principal component is a linear combination of the original variables, with coefficients equal to the eigenvectors of the correlation or covariance matrix. The eigenvectors are customarily taken with unit length. The principal components are sorted by descending order of the eigenvalues, which are equal to the variances of the components.

The principal components meet the following properties (Rao 1964; Kshirsagar 1972):

- The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables.
- The principal component scores are jointly uncorrelated. This property ensures the lack of multicollinearity when we use them as input variables in a regression model.
- The first principal component has the largest variance of any unit-length linear combination of the observed variables. The j th principal component has the largest variance of any unit-length linear combination orthogonal to the first $j-1$ principal components. The last principal component has the smallest variance of any linear combination of the original variables.
- The scores on the first j principal components have the highest possible generalized Variance of any set of unit-length linear combinations of the original variables.
- The first j principal components provide a least squares solution to the model:

$$Y = XB + E$$

Where:

Y is an $n \times p$ matrix of the centered observed variables;
 X is the $n \times j$ matrix of scores on the first j principal components;
 B is the $j \times p$ matrix of eigenvectors;
 E is an $n \times p$ matrix of residuals;

Our goal is to minimize the trace of $E'E$. That means that the first j principal components are the best linear predictors of the original variables among all possible sets of j variables, although any nonsingular linear transformation of the first j principal components would provide an equally good prediction.

In geometric terms, the j -dimensional linear subspace spanned by the first j principal components provides the best possible fit to the data points as measured by the sum of squared perpendicular distances from each data point to the subspace. This is in contrast to the geometric interpretation of least squares regression, which minimizes the sum of squared vertical distances.

3- The data set

Our starting point was an extensive collection of time series data comprised of quarterly indicators on a wide range of economic areas valued at current prices (raw data) covering the period from 1995 onwards.

We next selected a subset of indicators, taking into account various attributes: high correlation to the VAT revenue at current prices and quarterly variation rate, speed of publication, operability (easy access), coverage, cyclical sensitivity and frequency.

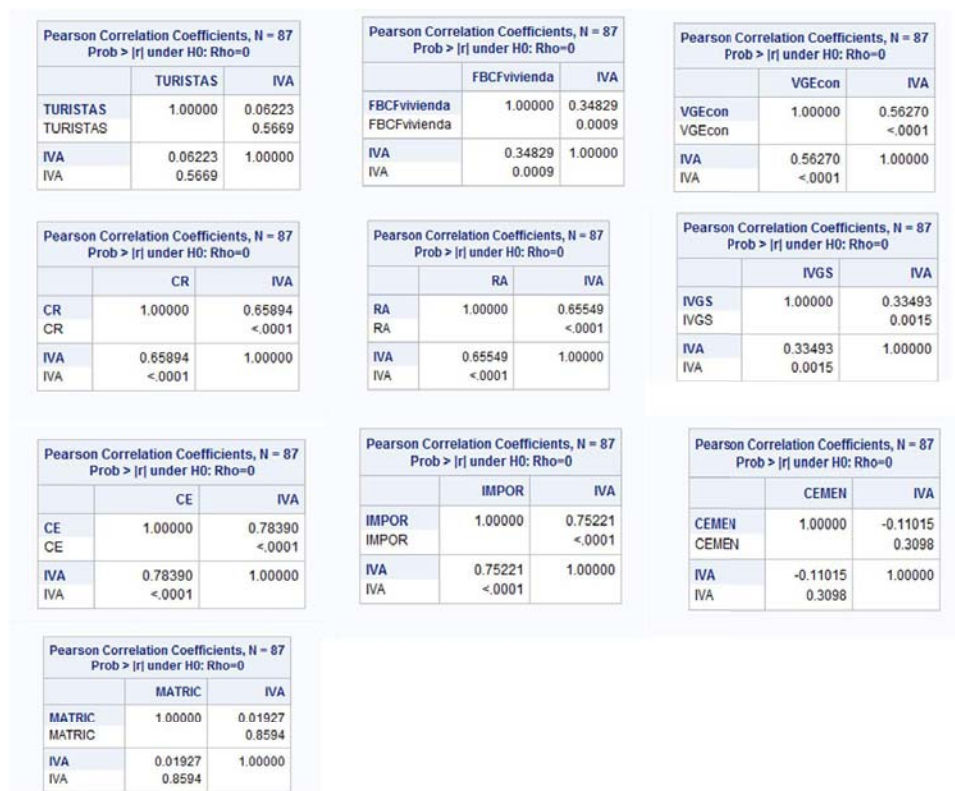


FIGURE 1: Pearson Correlation Coefficients and associated p-values for VAT tax revenue and the resulting ten partial indicators⁵.

The next step in the process was to identify the underlying cyclical pattern of the indicators. This goal required the removal of two factors: long term trends and high frequency noise. We decided to remove these factors in a single step using a Fixed length Symmetric Band-Pass Filter (Baxter-King).

⁵ The correlation coefficients of *Cement Apparent Consumption* and *Passenger Car Registrations* are low compared to the rest of the indicators. We considered this two reference series as useful indicators because of their cycle pattern (Figures 2-4), their relatively short publication lags, and because they belong to economic areas which are sensitive to policy changes.

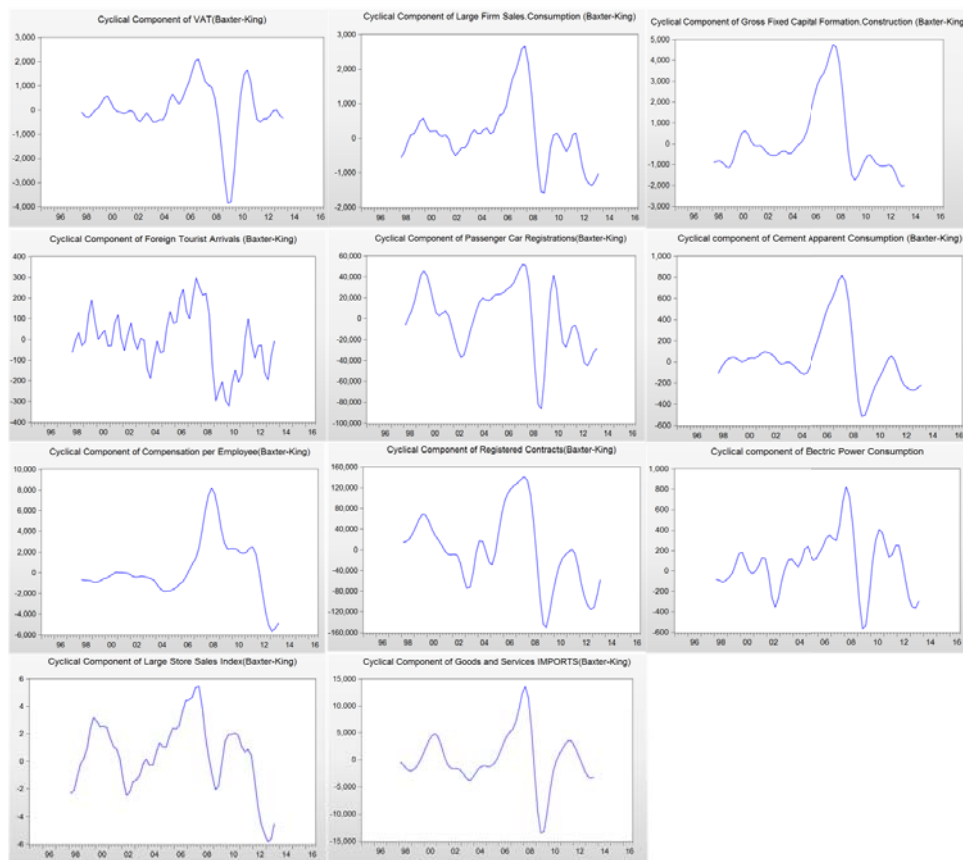


FIGURE 2: Cyclical patterns of VAT revenue and selected partial indicators obtained by a Baxter-King filter.

If the cyclical profiles are highly correlated⁶, the indicator would provide a signal, not only to approaching turning points, but also to developments over the whole cycle. The cross correlation function between the cyclical component of the partial indicators and the cyclical component of the VAT revenue, provides invaluable information on cyclical conformity. The location of the peak of the cross-correlation function is a good indicator of average lead time.

⁶ The methodology guideline “OECD System of Composite Leading Indicators” prepared by Gyorgy Gyomai and Emmanuelle Guidetti in April 2012, specifies this approach to select the reference series based on cyclical profiles.

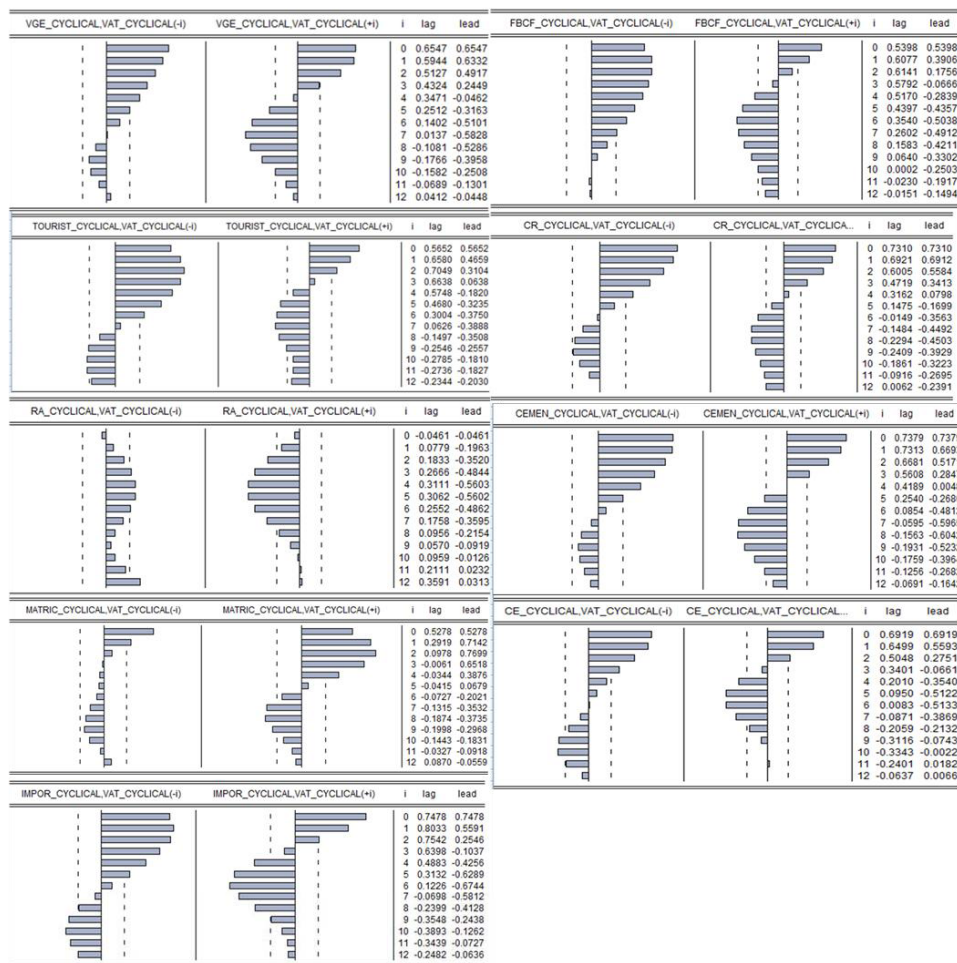


FIGURE 3: Cross Correlograms of the cyclical component of VAT revenue and the cyclical components of the selected partial indicators obtained by a Baxter-King filter.

The cross-correlation analysis of the two cyclical components shows that the cyclical component of *Cement Apparent Consumption* is highly correlated to the cyclical component of VAT revenue at lag=0 (Rho=0.7379). A leading relationship (lag=1, Rho=0.7313) could be rationalised on the basis that construction is probably the sector which reacts most quickly to changes in financial conditions.

Figures 3 feature similar results for the cyclical component of *Passenger Car Registrations*; the cross-correlation analysis of the two cyclical components shows that the cyclical component of *Passenger Car Registrations* is highly correlated to the cyclical component of VAT revenue at lag=0 (Rho=0.5278) but the maximum occurs at lag=2 (Rho=0.7699). That result reveals a lagging relationship between these two variables.

Similar cyclical pattern analysis of the rest of the candidate reference series are shown in Figure 3. Specifically, for selected partial indicators, *Fixed Capital Formation in Construction*, *Goods and Services Imports*, *Compensation per Employee* and *Foreign Tourists Arrivals* display movements that precede those of the VAT revenue (average lead times are two, one, four and two quarters, respectively). *Large Store Sales Index*, *Large Firm Sales (Consumption)*, *Cement Apparent Consumption*, *Electric Power Consumption* and *Registered Contracts* are more significant in providing contemporaneous information. *Passenger Car Registrations* and performs as lagging indicator.

Note that whereas the correlation value of the peak provides a measure of how well the cyclical profiles of the indicators match, the size of the correlations cannot be the only indicators used for component selection.

Higher correlations in quarterly variation rate maintain a similar structure and correspond to general indicators (*Electric Power Consumption*), consumption indicators (*Large Firm Sales in Consumption Goods and Services*), construction indicators (*Gross Fixed Capital Formation in Construction*) and services indicators (*Foreign Tourists Arrivals*).

We fitted the model using a training data set from $t=1$ (first quarter of 1995) to $t=T$ (last quarter of 2014) and then we tested its performance computing one-step ahead forecasts on a test data set (first, second and third quarter of 2015). Once we have checked the predictive ability of the model, and since the latest update of the VAT revenue released by the Spanish Tax Agency corresponds to the third quarter of 2016, we provide forecasts for the last quarter of 2016 and the four quarters of 2017. The latest predictions are obtained by extending the partial indicators using seasonal ARIMA models.

The following sections provide a more detailed description of the various steps highlighted above.

4-Estimation results

4.1. Finding the two orthogonal regressors.

As indicated before, the purpose of the principal component analysis is to compute two variables that best summarize all ten partial indicators.

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 10 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.75530185	3.26313005	0.5755	0.5755
2	2.49217180	1.60767047	0.2492	0.8247
3	0.88450133	0.42084794	0.0885	0.9132
4	0.46365339	0.23280994	0.0464	0.9596
5	0.23084345	0.15759913	0.0231	0.9826
6	0.07324432	0.02041474	0.0073	0.9900
7	0.05282959	0.02794817	0.0053	0.9953
8	0.02488142	0.01218430	0.0025	0.9977
9	0.01269712	0.00282137	0.0013	0.9990
10	0.00987574		0.0010	1.0000

TABLE 1: Eigenvalues of the Correlation Matrix

Results of the principal component analysis are displayed on Table 1. We compute principal components from the correlation matrix. The set of partial indicators show a high correlation between the variables, validating the relevance of prior principal component analysis to avoid problems of multicollinearity.

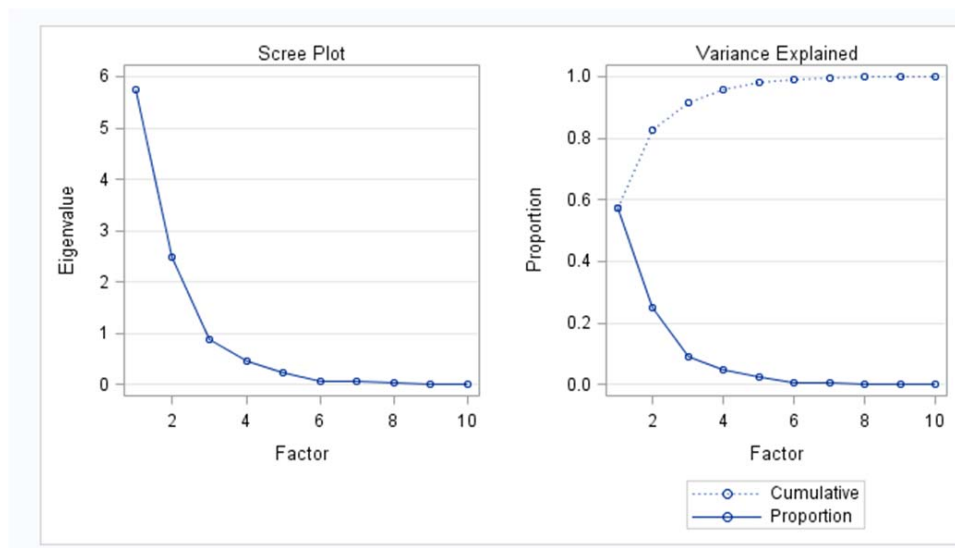


FIGURE 4: Scree Plot and Variance Explained Plot.

The Scree Plot on the left in Figure 4 shows that the eigenvalue of the first component is approximately 5,8 and the eigenvalue of the second component is largely decreased to 2,5. The Variance Explained Plot on the right in Figure 4 shows that the first two principal components account for nearly 82% of the total standardized variance, which indicates that two components provide a good summary of the data.

Factor Pattern			
		ivafactor1	ivafactor2
IVGS	IVGS	0.80375	0.42794
VGEcon	VGEcon	0.97638	-0.00563
MATRIC	MATRIC	0.29556	0.79694
IMPOR	IMPOR	0.90220	-0.36185
RA	RA	0.87476	-0.41368
FBCFvivienda	FBCFvivienda	0.79814	0.52059
CEMEN	CEMEN	0.23950	0.93610
CE	CE	0.89822	-0.27121
TURISTAS	TURISTAS	0.44321	-0.36425
CR	CR	0.88984	-0.13522

TABLE 3: Factor Pattern of the two principal components.

The factor pattern (Table 3) shows that the first component (labeled " Ivafactor1") has large positive loadings for all ten variables. The second component is basically a contrast of *Large Store Sales Index* (0,428), *Passenger Car Registrations* (0,797), and the two construction indicators (*Gross Fixed Capital Formation in Construction* and *Cement Apparent Consumption*) against the rest.

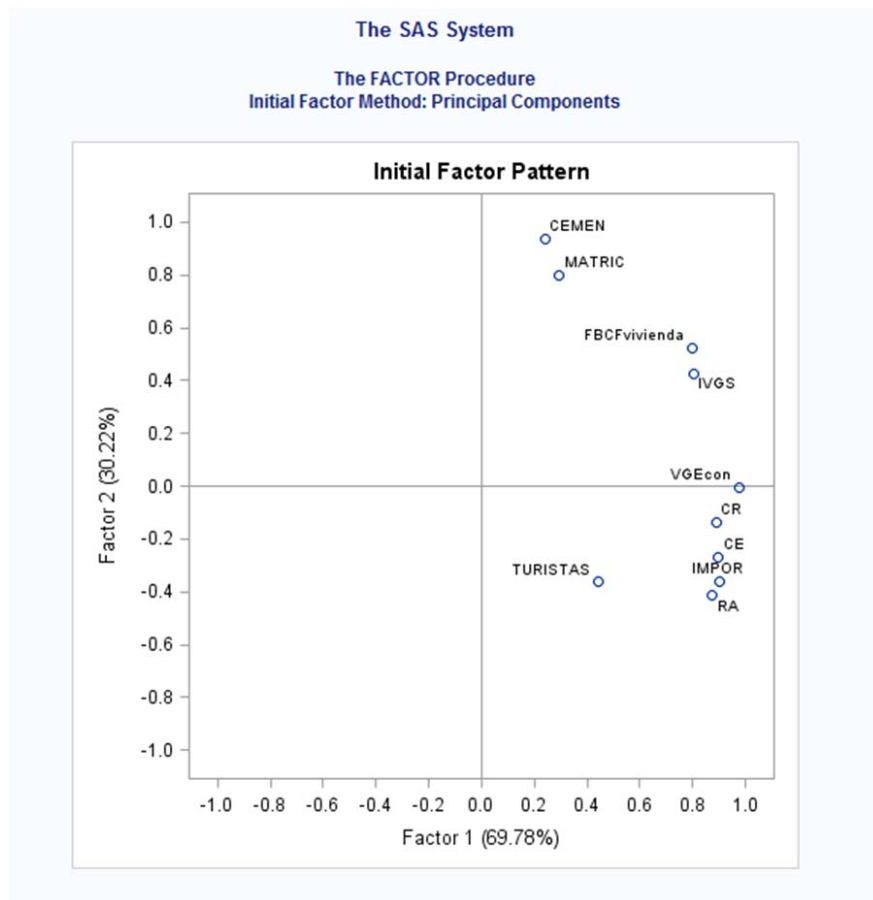


FIGURE 5: Unrotated Factor Pattern Plot of the two Principal Components.

The unrotated factor pattern (Figure 5) reveals three clusters of variables, with the variables *Cement Apparent Consumption* and *Passenger Car Registrations* at the positive end of Factor2 axis, and *Compensation per Employee*, *Goods and Services Imports* and *Electric Power Consumption* at the negative side. The rest of the variables remain between these two clusters.

The results of the Varimax rotation are shown in Table 4 and Figure 6.

Standardized Scoring Coefficients			
		ivafactor1	ivafactor2
IVGS	IVGS	0.05547	0.21427
VGEcon	VGEcon	0.15518	0.06859
MATRIC	MATRIC	-0.08647	0.31212
IMPOR	IMPOR	0.20298	-0.06673
RA	RA	0.20731	-0.08762
FBCFvivienda	FBCFvivienda	0.03910	0.24767
CEMEN	CEMEN	-0.11858	0.35883
CE	CE	0.18721	-0.03395
TURISTAS	TURISTAS	0.13088	-0.10081
CR	CR	0.16316	0.01505

TABLE 4: Standardized Factor Scoring Coefficients

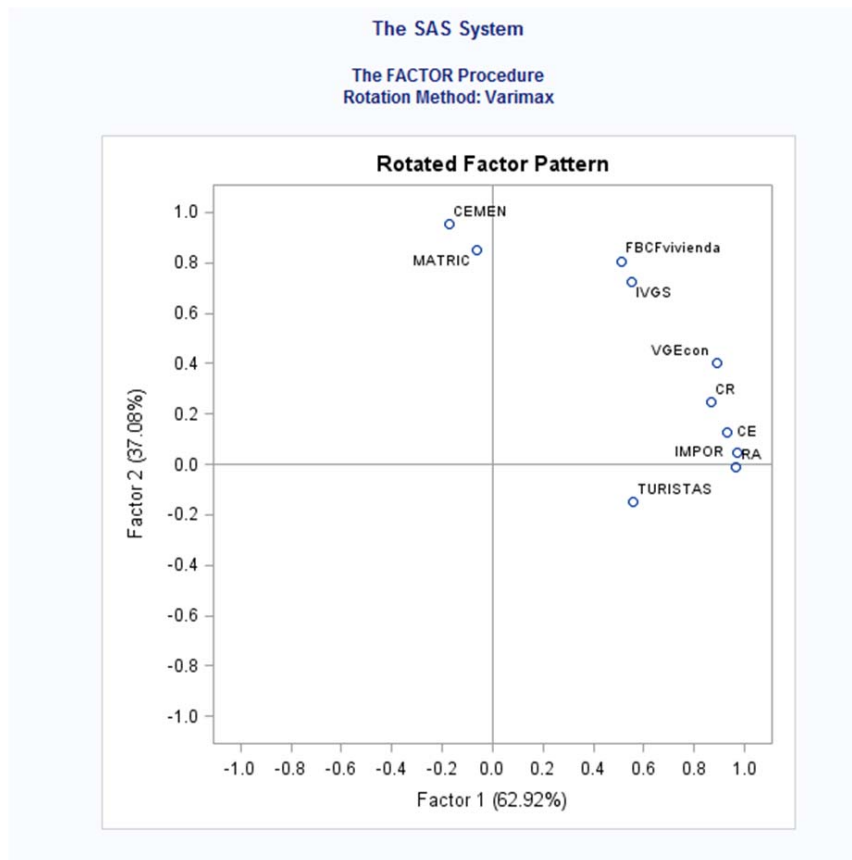


FIGURE 6: Graphical Plot of the VARIMAX Rotated Factor Loadings.

The graphical plot of the Varimax-rotated factor loadings clearly features that *Cement Apparent Consumption* and *Passenger Car Registrations* cluster together on Factor 2 axis, while *Compensation Per Employee*, *Goods and Services Imports*, *Electric Power Consumption*, *Foreign Tourists Arrivals*, *Registered Contracts*, *Large Firm Sales (Consumption)* cluster together on the Factor 1 axis. The standardized scoring coefficients of *Gross Fixed Capital Formation (Construction)* and *Large Store Sales Index* are larger in factor 2 than in factor 1.

4.2-Determining the transfer function.

The second part of the methodology makes use of these factors as input variables for a transfer function:

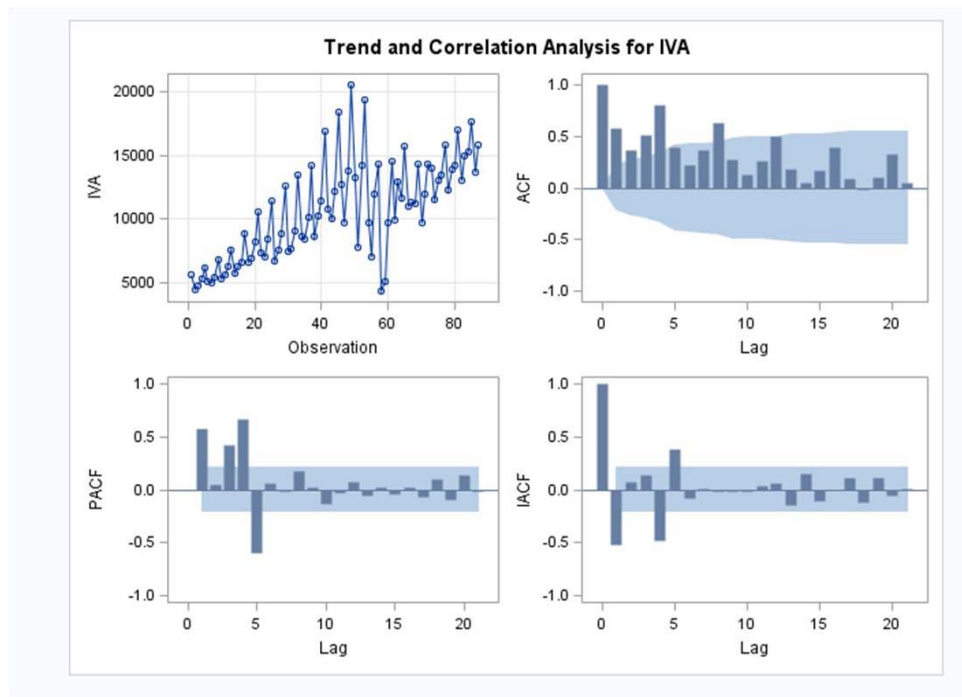


FIGURE 7: Correlation analysis panel for *VAT revenue*. Sample Autocorrelation Function plot (ACF), Partial Autocorrelation Function plot (PACF) and Sample Inverse Autocorrelation Function plot (IACF) of *VAT revenue*.

We introduce in the model two level shifts corresponding to the second quarter of 2010 and the third quarter of 2012 (VAT reform). The parameter estimates table and goodness-of-fit statistics for this model are shown in the conditional Least Squares Estimation table (Table 5).

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MU	9317.3	598.35751	15.57	<.0001	0	IVA	0
MA1,1	0.31931	0.10693	2.99	0.0037	1	IVA	0
AR1,1	0.92926	0.04840	19.20	<.0001	4	IVA	0
NUM1	1294.0	348.39201	3.71	0.0004	0	ivafactor1	0
NUM2	2336.5	246.59014	9.48	<.0001	0	ivafactor2	0
NUM3	4994.5	439.69931	11.36	<.0001	0	Is2010q2	0
NUM4	2351.5	402.43262	5.84	<.0001	0	Is2012q3	0

TABLE 5: Table of parameter estimates. Method : Conditional Least Squares.

As shown in table 5, all parameters are statistically significant, although the moving average parameter MA1,1 is close to the 5% significance level.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.20	4	0.3790	0.055	-0.178	0.014	0.015	-0.061	0.083
12	12.89	10	0.2301	0.236	0.036	-0.011	-0.093	0.062	-0.136
18	17.31	16	0.3660	-0.053	0.075	-0.005	0.028	-0.047	-0.169
24	23.63	22	0.3669	-0.108	0.032	0.059	0.140	-0.122	-0.053

TABLE 6: Check for White Noise Residuals.

The autocorrelations checks on the residuals (Table 6) features there is no autocorrelation of residuals at any lag. Test statistics fail to reject the no-autocorrelation hypothesis at a high level of significance ($p = 0.3790$ for the first six lags). This result seems fairly robust to changes in the number of lags.

The probability of white noise is clearly high (Figure 10).

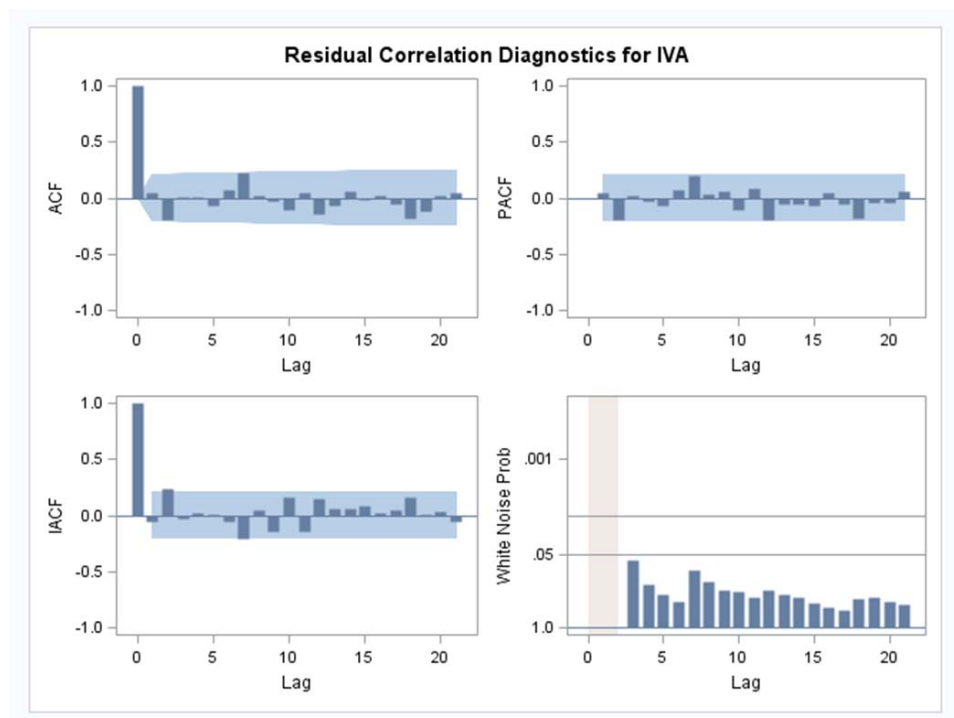


FIGURE 8: Correlation analysis panel for residuals. ACF, PACF and IACF plots of the residuals. White Noise Probability plot.

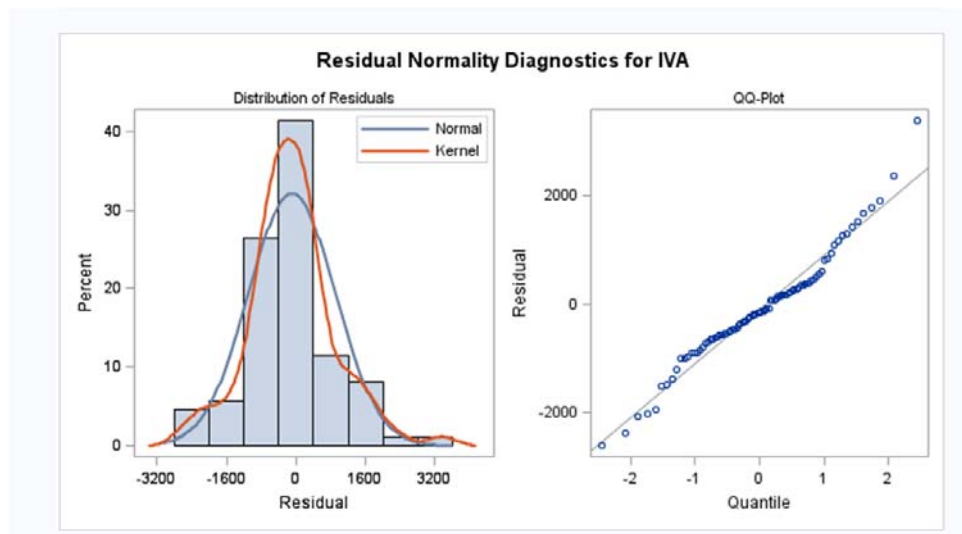


FIGURE 9: Residual Normality Diagnostics.

As showed in Figure 9 residuals of the model follow a *Normal* distribution.

4.3-Out of sample forecasts.

1. For the first quarter of 2015, the observed *VAT revenue* at current prices in Millions Euros was: 16997,655. Table 7 shows the predicted values for *VAT revenue* by the model.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
81	16804.4492	1114.6663	14619.7434	18989.1550

TABLE 7: Forecast and Confidence Limits of *VAT revenue*. Out of Sample estimations. First quarter of 2015. Million Euros.

2. For the second quarter of 2015, the observed *VAT revenue* at current prices in Millions Euros was: 13032,929. Table 8 shows the predicted values for *VAT revenue* by the model.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
82	13349.7536	1106.5157	11181.0226	15518.4846

TABLE 8: Forecast and Confidence Limits of *VAT revenue*. Out of Sample estimations. Second quarter of 2015. Million Euros.

3. For the third quarter of 2015, the observed VAT revenue at current prices in Millions Euros was: 14976,823. Table 9 shows the predicted values for *VAT revenue* by the model.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
83	15000.9714	1098.8438	12847.2771	17154.6657

TABLE 9: Forecast and Confidence Limits of *VAT revenue*. Out of Sample estimations. Third quarter of 2015. Million Euros.

Once we have checked the predictive ability of the model, and since the latest update of the VAT revenue released by the Spanish Tax Agency corresponds to the third quarter of 2016, we extended the partial indicators using seasonal ARIMA models to provide forecast for the last quarter of 2016 and the four quarters of 2017.

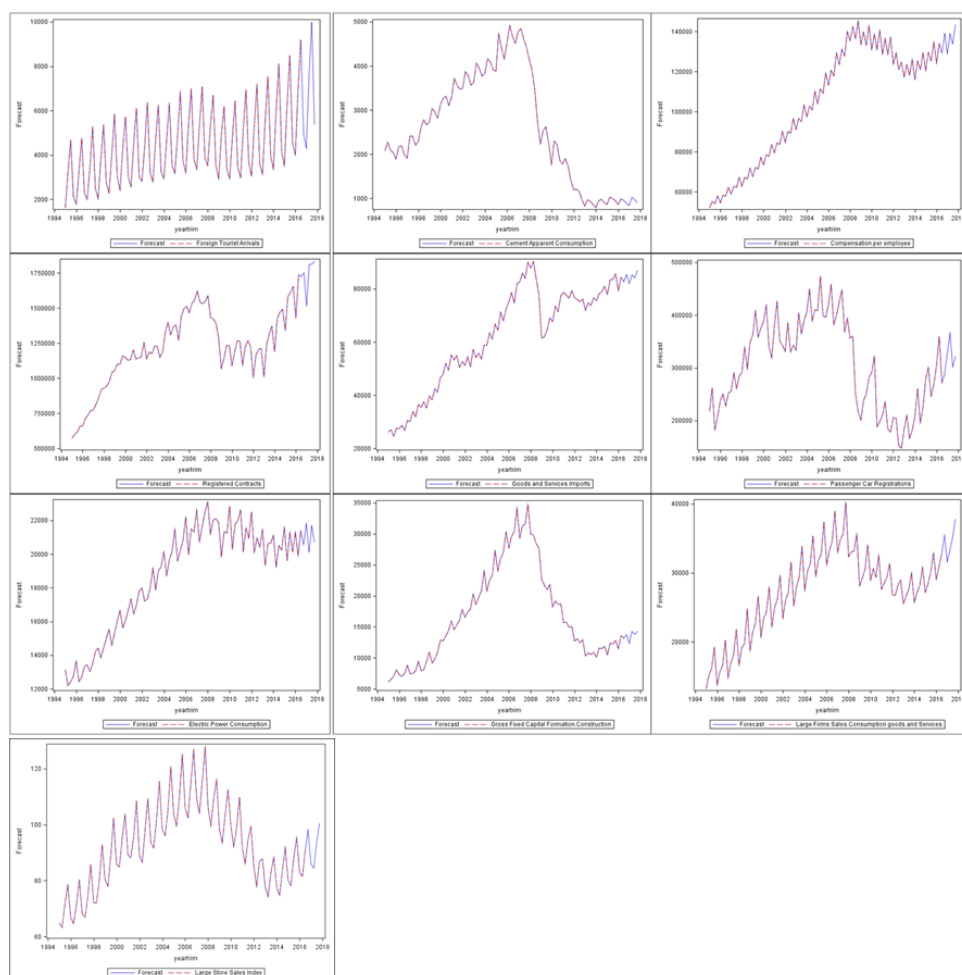


FIGURE 10: Extended partial indicators. Four quarters of 2017.

4.5-Forecasts from extended partial indicators.

We also provide forecasts for the last quarter of 2016 and the four quarters of 2017 obtained from the extended partial indicators.

Forecast last quarter 2016. M.E.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
88	15846.4051	1035.2207	13817.4097	17875.4005

Forecast first quarter 2017. M.E.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
89	18082.1236	1028.8106	16065.6918	20098.5554

Forecast second quarter 2017. M.E.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
90	14332.6736	1022.5182	12328.5748	16336.7723

Forecast third quarter 2017. M.E.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
91	16480.5134	1016.3398	14488.5241	18472.5028

Forecast fourth quarter 2017. M.E.

Forecasts for variable IVA				
Obs	Forecast	Std Error	95% Confidence Limits	
92	16525.9196	1010.2720	14545.8229	18506.0164

Table 10: Forecasts and Confidence Limits of VAT revenue. Last quarter of 2016. Four quarters of 2017. Million Euros.

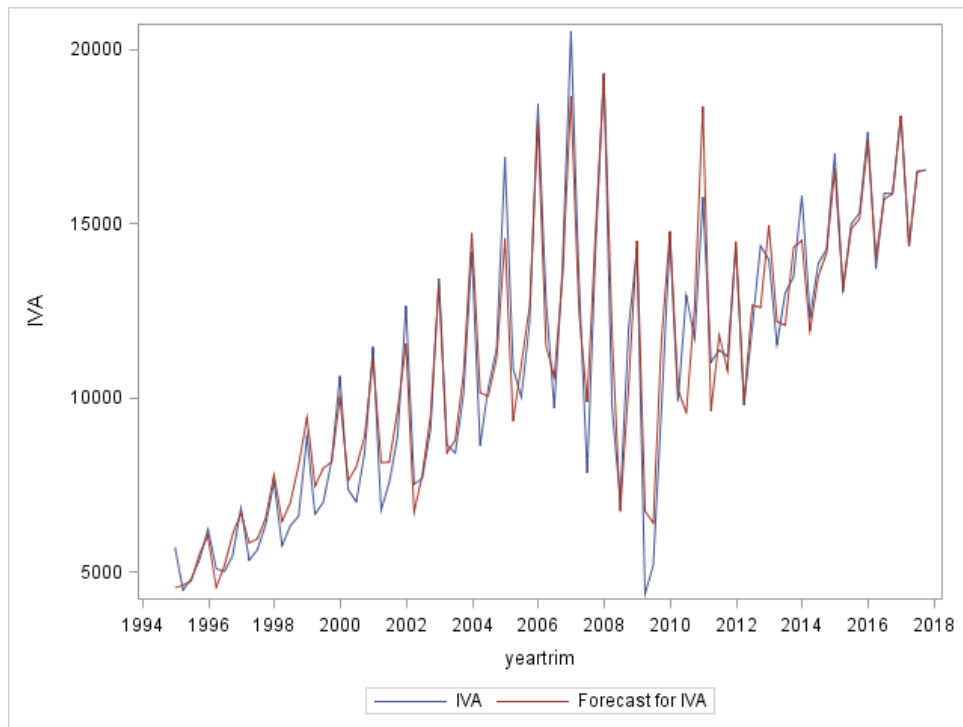


FIGURE 11: VAT revenue and forecasts. Four quarters of 2017.

5-Conclusions:

As mentioned in the introduction, the final aim of this paper is to propose a methodology that successfully combines principal components analysis and transfer function theory to forecast VAT revenue. This approach offers advantages to Value-Added Tax forecasting models based on the National Accounts approach, and specifically, to those using *Total Consumption Expenditure on Goods and Services* as the only input explanatory variable. The pre-selection of the reference series and the dimension reduction technique enables to incorporate in advance changes in specific fields of the economy that may affect tax revenues.

The analysis of which partial indicators contain useful leading or lagging information about the dependent variable and the filtering process aimed to identify underlying cyclical pattern of the candidate component series was not simple. The approach taken does not in any sense attempt to construct an optimal set of partial indicators, but has the more limited aim of assessing which indicators contain information that is useful for VAT revenue short-term forecasting. Among other criteria, we selected those variables that exhibited a cyclical profile highly correlated to the cyclical pattern of VAT revenue. *Fixed Capital Formation in Construction, Goods and Services Imports*, and *Foreign Tourists Arrivals* were found to add significant leading information to the model. The usefulness of the rest of indicators arises from the contemporaneous relationships between the variables, and their inclusion in the model found some support in less sophisticated methods such as correlation analysis. Consumption indicators selected have also the advantage of having shorter publications lags than the National Accounts.

The output factors obtained from the dimension reduction technique were highly significant as explanatory variables in the transfer function. Thus, their influence has been crucial for achieving such high predictive power of the model.

6-References:

- Arnold, J. Brys, B. Heady, C., Johansson, A. Schwellnus, C. and Vartia, L. (2011).** *Tax policy for economic recovery and growth*, The Economic Journal, 121, 59-80.
- Artola, C., and Galán, E. (2012).** *Tracking the future on the Web: construction of leading indicators using internet searches*. Banco de España. Documentos Ocasionales N.º 1203.
- Bandholz, Harm (2005)** *New Composite Leading Indicators for Hungary and Poland*. Ifo Working Paper, No. 3.
Available at <http://www.econstor.eu/handle/10419/73832>.
- Central Bank of Spain (2015)** *Quarterly Report on the Spanish Economy*. Economic Bulletin, December 2015. Available at
<http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEconomico/descargar/15/Dic/Files/be1512-coye.pdf>
- Danninger, S. (2005).** Revenue forecasts as performance targets.
- Duncan, G., Gorr, W., & Szczypula, J. (1993).** Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting. *Management Science*, 39(3), 275-293.
- European Commission. (2016)** *Country Report Spain 2016*. Including an In-Depth Review on the prevention and correction of macroeconomic imbalances: (February 26, 2016). Commission staff working document.
Available at http://ec.europa.eu/europe2020/pdf/csr2016/cr2016_spain_en.pdf
- European Commission (2015),** *VAT Rates Applied in the Member States of the European Union*. Situation at 1st September 2015– Taxud.c.1(2015) - EN
Available at
http://ec.europa.eu/taxation_customs/resources/documents/taxation/vat/how_vat_works/rates/vat_rates_en.pdf.
- Glenn P. J., Chu-Yan K. and Gangadhar P. S. (2000)** *Tax Analysis and Revenue Forecasting-Issues and Techniques*. Harvard Institute for International Development, Harvard University.
- Giles, C. and Hall, J. (1998)** “Forecasting the PSBR Outside Government: The IFS Perspective” *Fiscal Studies* Volume 19, Issue 1.
- Golosov, M. (2002).** Tax revenue forecasts in IMF-supported programs.
- Gyomai, G. and E. Guidetti (2012)** *OECD system of composite leading indicators*. Available at <http://www.oecd.org/std/leading-indicators/41629509.pdf>

Jenkins, G. P., Kuo, C. Y., & Shukla, G. (2000). Tax analysis and revenue forecasting. *Cambridge, Massachusetts: Harvard Institute for International Development, Harvard University.*

Labeaga, J. M. and A. López (1994), *Estimation of the welfare effects of indirect tax changes on Spanish households: an analysis of the 1992 VAT reform*", *Investigaciones Económicas*, Vol. XVIII(2), May, pp.289-311.

Leal, T., Pérez, J. J., Tujula, M., & Vidal, J. P. (2008). Fiscal forecasting: lessons from the literature and challenges. *Fiscal Studies*, 29(3), 347-386.

Le Minh, T. (2007). Estimating the VAT base: method and application. *Tax Notes International*, 46, 42.

Legeida, N., & Sologoub, D. (2003). *Modeling value added tax (VAT) revenues in a transition economy: Case of Ukraine*. Institute for economic research and policy consulting working paper, (22), 1-21.

Lundback, E.J. (2008). *Medium-Term Budgetary Frameworks - Lessons for Austria from International Experience*. IMF working paper WP/08/163
Available at <https://www.imf.org/external/pubs/ft/wp/2008/wp08163.pdf>.

Michael Keen (2013) *The Anatomy of the VAT*. IMF Working Paper. Fiscal Affairs Department. WP/13/111. Available at
<https://www.imf.org/external/pubs/ft/wp/2013/wp13111.pdf>.

Pavlik, M. (2008). *The Usage Of the Dummy Variable for VAT Forecasting of the Tax Administration in the Slovak Republic*. *Prace Naukowe Uniwersitetu Ekonomicznego we Wrocławiu, Ekonometria*, 21, 40-54.

Pérez, C. (2006) *Econometría de las series temporales*. Prentice Hall.

Pérez, C. (2007) *Econometría básica*. Prentice Hall.

Pérez, C. (2008) *Econometría avanzada. Técnicas y herramientas..* Prentice Hall.

Pérez, C. (2010) *El Sistema Estadístico SAS*. Garceta Grupo Editorial

Pérez, C. (2013) *Análisis multivariante de datos*. Garceta Grupo Editorial

Pike, T., & Savage, D. (1998). *Forecasting the public finances in the Treasury*. *Fiscal Studies*, 19(1), 49-62.

Sancak, C., Velloso, R., & Xing, J. (2010). *Tax revenue response to the business cycle*.

Slobodnitsky, T., & Drucker, L. (2008). VAT Revenue Forecasting in Israel. *Ministry of Finance, State Revenue Administration*, The Maurice Falk Institute for Economic Research in Israel Ltd.

Sung, M. J. (1999). *Estimation of Tax Evasion in Global Income Tax and VAT for Enhancing the Accuracy of Revenue Forecasting.* , Korea Institute of Public Finance, Séoul.

Țitan, E., Boboc, C., Ghita, S., Todose, D (2011) *Econometric Analysis of the Correlations between the Social Security Budget and the Main Macroeconomic Aggregates in Romania.* Theoretical and Applied Economics Volume XVIII (2011), No. 2(555), pp. 117-126.

Wawire, N. H. W. (2011). Determinants of value added tax revenue in Kenya.

Zee, H.H. and Boding, J.P (1992) Aspects of introducing a Value-Added Tax in Sri Lanka. Paper prepared for the International Monetary Fund, Fiscal Affairs Department.

Combining forecasts to capture realized volatility dynamics

Giovanni De Luca¹, Giampiero M. Gallo², and Danilo Carità¹

¹ Università degli Studi di Napoli "Parthenope"

² Università degli Studi di Firenze

Abstract. In this work we provide the findings of a forecast combination analysis carried out on the realized volatility series of three market indexes (DAX, CAC, AEX). Two volatility types (5 minutes, kernel) have been considered. The results suggest that forecasts computed through combining models are generally more accurate than those provided by single models. However, the choice of the latter can affect significantly the goodness of the results.

Keywords: Realized volatility, forecast combinations, loss functions.

1 Introduction

Volatility is a central parameter for many financial decisions including the pricing and hedging of derivative products as well as the development of efficient risk management methods. Most of the volatility models presented in the literature are based on the empirical detection that volatility is time-varying and that periods of high volatility tend to cluster (Ané 2006). The forecasting process of such an important measure represents a major issue.

In literature there exists a wide variety of models that are able to estimate volatility forecasts, but they are, almost by definition, simple and incomplete (Raviv 2016). An improvement in the forecasts accuracy can be achieved combining forecasts originated from different types of models. Forecast combinations have been used successfully in empirical work in diverse areas such as forecasting Gross National Product, currency market volatility, inflation, money supply, stock prices, meteorological data, city populations, outcomes of football games, wilderness area use, check volume and political risks (Timmermann 2006).

The aim of this paper is to forecast the daily realized volatility one-step-ahead for a one-year period with both single and combining models. Thereafter we will compare the predicted values with the actual data by means of a number of loss functions. To carry out our analysis we have used data on realized volatility from 01/01/2008 to 31/12/2016 of three market indexes (DAX, CAC, AEX)

The remainder of the paper is organized as follows. Section 2 describes the data, the models adopted and the loss functions used for evaluating the different forecasts. Section 3 presents the results of the analysis while Section 4 concludes.

2 Data and Methodology

This study focuses on the realized volatility of three European market indexes:

- DAX 30 (*Deutsche Aktienindex 30*) is a blue chip stock market index consisting of the 30 major German companies trading on the Frankfurt Stock Exchange;
- CAC 40 (*Cotation Assistée en Continu*) represents a capitalization-weighted measure of the 40 most significant values among the 100 highest market caps on the Euronext Paris;
- AEX (*Amsterdam Exchange Index*) is a stock market index composed of Dutch companies that trade on Euronext Amsterdam, composed of a maximum of 25 of the most frequently traded securities on the exchange.

The time series of the indexes are provided by the Oxford-Man Institute of Quantitative Finance by means of its own website (*Oxford-Man Institute of Quantitative Finance Realized Library* 2017). For each asset, the dataset contains the realized volatility collected every 5 minutes, the realized kernel volatility and the daily returns, covering the period from 01/01/2008 to 31/12/2016.

Three different models have been chosen to create the single forecasts:

1. *Asymmetric Multiplicative Error Model* (AMEM) (Engle 2002; Engle and Gallo 2006), which for a basic (1,1) order has the following structure:

$$\begin{aligned} rv_t &= \mu_t \xi_t \\ \mu_t &= \omega + \alpha_1 rv_{t-1} + \beta_1 \mu_{t-1} + \gamma D_{t-1} rv_{t-1} \end{aligned} \quad (1)$$

with $\omega > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, $\alpha_1 + \beta_1 < 1$. D_t is a dummy variable that takes the value of 1 if the return at time t is negative and 0 otherwise;

2. *Asymmetric Power Multiplicative Error Model* (APMEM), which for the usual (1,1) order is given by:

$$\begin{aligned} rv_t &= \mu_t \xi_t \\ \mu_t^\delta &= \omega + \alpha_1 rv_{t-1}^\delta + \beta_1 \mu_{t-1}^\delta + \gamma D_{t-1} rv_{t-1}^\delta \end{aligned} \quad (2)$$

with $\omega > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, $\alpha_1 + \beta_1 < 1$, $\delta > 0$. This model is a generalization of the basic MEM and follows the APGARCH philosophy (Ding, Granger, and Engle 1993);

3. *Asymmetric Heterogeneous AutoRegressive Model* (AHAR), that is the HAR model (Corsi 2009) with a leverage effect term:

$$rv_t = c + \beta^{(d)} rv_{t-1} + \beta^{(w)} rv_{t-1}^{(w)} + \beta^{(m)} rv_{t-1}^{(m)} + \epsilon_t^{(d)} \quad (3)$$

where:

(d) stands for the time horizons of one day;

$rv_{t-1}^{(w)}$ is the weekly realized volatility which at time t is given by the average

$$rv_t^{(w)} = \frac{1}{5} \left(rv_t^{(d)} + rv_{t-1d}^{(d)} + \dots + rv_{t-4d}^{(d)} \right) \quad (4)$$

$rv_{t-1}^{(m)}$ is the monthly realized volatility which at time t is given by the average

$$rv_t^{(m)} = \frac{1}{22} \left(rv_t^{(d)} + rv_{t-1d}^{(d)} + \dots + rv_{t-21d}^{(d)} \right) \quad (5)$$

As a preliminary analysis, in Fig. (1) we compare the forecasts obtained using the three models above-mentioned for the year 2016 (coloured lines) with the actual values of the volatility (dashed black line) for the DAX 5-minutes series. The chart shows that all models react satisfactorily to positive peaks of volatility, whereas they are not able to achieve a suitable degree of accuracy when volatility reaches a local minimum. This issue, which is common also to the other observed time series, can be overcome by combining the forecasts of two models, as we will see thereafter.

The combining methods are based on the following two combination models:

- *comb1* model, based on a simple unconstrained Ordinary Least Squares estimates of the weights. The one-step-ahead forecast is given by

$$rv_T(1) = \alpha + \beta_1 f_T^{(1)}(1) + \beta_2 f_T^{(2)}(1) \quad (6)$$

with $f_T^{(1)}(1)$ and $f_T^{(2)}(1)$ denote, respectively, the first and second model forecasts;

- *comb2* model, with the combination given by

$$rv_T(1) = \alpha + (\beta_1 + \delta_1 D_{t-1}) f_T^{(1)}(1) + (\beta_2 + \delta_2 D_{t-1}) f_T^{(2)}(1) \quad (7)$$

which includes a dummy variable D_t that takes the value 1 if rv_t is lower than rv_{t-1} and 0 otherwise. The ratio of this choice is given by the consideration that, as we have mentioned before, the forecast of volatility is often far from the actual realized volatility while this is decreasing.

To compare the results of the combination schemes with those that can be reached by exclusively relying on a single model, we compute four loss functions:

1. *Mean Square Error* (MSE);
2. *Mean Absolute Error* (MAE);
3. *Quasi-Likelihood* (QLIKE), defined as

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{rv_{T+i}}{rv_{T+i-1}(1)} - \ln \left(\frac{rv_{T+i}}{rv_{T+i-1}(1)} \right) - 1 \right] \quad (8)$$

with rv_{T+i} being the observed value of the realized volatility and $rv_{T+i-1}(1)$ is the one-step-ahead forecast for time $T+i$, $i = 1, \dots, n$;

4. a new measure, given by

$$\frac{1}{n} \sum_{i=1}^n \left(1 + \left(\frac{|\epsilon_{T+i}|}{rv_{T+i}} \right)^m I(\epsilon_{T+i} > 0) \right) |\epsilon_{T+i}| \quad (9)$$

where $\epsilon_{T+i} = rv_{T+i} - rv_{T+i-1}(1)$. This measure is an extension of the MAE (we will call it Asymmetric MAE, or AMAE): it reduces to $|\epsilon_{T+i}|$ when the indicator function is 0 (overestimation of the volatility) and is given by $\left(1 + \left(\frac{|\epsilon_{T+i}|}{rv_{T+i}}\right)^m\right) |\epsilon_{T+i}|$ when the indicator function is 1 (underestimation of the volatility).

For the computation of the forecast combinations, we start by splitting the data into an estimation and training set and a test set. The former is again split into two parts, the first to estimate the parameters of the model, the second (the training period) to estimate the weights to be attributed to the single forecasts. The latter, the test set, will be used for the evaluation of the different models. We have chosen to take into account two different training periods in our analysis: a four-years training period and a three-years training period. For instance, with a four-years training period, we estimate the parameters of the models using observations from 02/01/2008 to 31/12/2011, then compute one-step-ahead forecasts from 02/01/2012 to 31/12/2015; these forecasts are used to estimate the weights of the combinations, finally the one-step-ahead forecast for 02/01/2016 is produced. Then, we estimate the parameters of the models using observations from 03/01/2008 to 02/01/2012, compute one-step-ahead forecasts from 03/01/2012 to 02/01/2016 to estimate the weights of the combinations, and the one-step-ahead forecast for 03/01/2016 is produced. And so on.

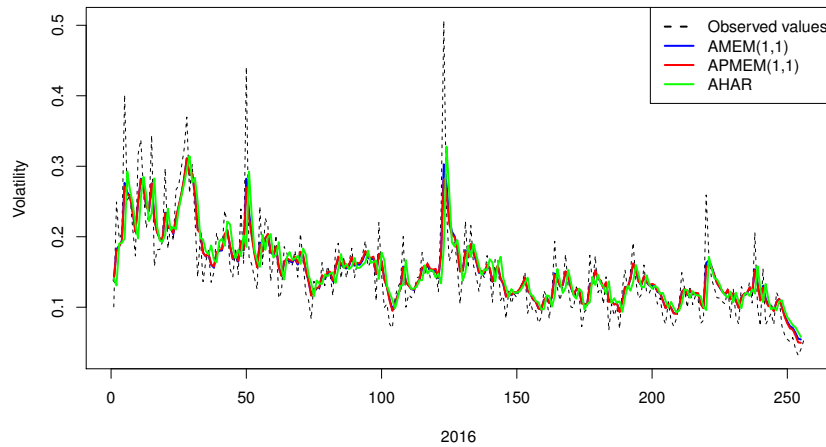


Fig. 1. Comparison among observed realized volatility (5 minutes) for year 2016 and AMEM(1,1), APMEM(1,1) and AHAR forecasts - DAX dataset

3 Comparisons among forecasting models

In this section we will show the findings of our analysis. For each model we display the value of the four loss functions mentioned above between the forecasts and the observed values. Since two of the three single models we have used (AMEM and APMEM) are very similar each other, we will present first a comparison between AMEM and AHAR and then between APMEM and AHAR, along with the combination schemes we have described in Sect. 2.

3.1 AMEM vs AHAR

The order of the two single models is defined using the Ljung-Box test on the residuals of the in-sample analysis of the two models. We have selected an AMEM(1,1) for DAX dataset, an AMEM(1,2) for CAC and AEX, and an AHAR with a second lag term (rv_{t-2}) for all datasets.

Table 1. Comparison among AMEM(1,1), AHAR and combination schemes (in bold the smallest values) - DAX dataset

Series	Training period	MSE				MAE			
		AMEM (1,1)	AHAR	comb1	comb2	AMEM (1,1)	AHAR	comb1	comb2
rv 5 min.	4 years	0.254	0.293	0.254	0.248	3.459	3.671	3.458	3.361
	3 years	0.254	0.293	0.254	0.250	3.459	3.671	3.461	3.418
rv kernel	4 years	0.206	0.251	0.206	0.200	3.124	3.386	3.125	3.021
	3 years	0.206	0.251	0.206	0.203	3.124	3.386	3.126	3.082
Series	Training period	QLIKE				AMAE ($m = 2$)			
		AMEM (1,1)	AHAR	comb1	comb2	AMEM (1,1)	AHAR	comb1	comb2
rv 5 min.	4 years	4.231	4.979	4.235	4.246	3.679	3.925	3.679	3.596
	3 years	4.231	4.979	4.244	4.338	3.679	3.925	3.682	3.644
rv kernel	4 years	3.505	4.312	3.505	3.508	3.296	3.599	3.297	3.205
	3 years	3.505	4.312	3.511	3.594	3.296	3.599	3.299	3.259

Table 1 shows the results for the first comparison, i.e. AMEM(1,1) and AHAR models along with combined forecasts on DAX data. We can see that *comb2* model performs very well for almost all indicators, only QLIKE prefers AMEM(1,1) model (three times out of four) and *comb1*.

Findings provided by Table 2 for CAC dataset are very similar, the only difference is that QLIKE prefers *comb2* for *rv* kernel with a training period of four years instead of *comb1*.

Table 2. Comparison among AMEM(1,2), AHAR and combination schemes (in bold the smallest values) - CAC dataset

Series	Training period	MSE				MAE			
		AMEM (1,2)	AHAR	comb1	comb2	AMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	0.243	0.292	0.243	0.239	3.164	3.502	3.169	3.098
	3 years	0.243	0.292	0.243	0.242	3.164	3.502	3.169	3.126
rv kernel	4 years	0.240	0.295	0.240	0.233	3.150	3.546	3.161	3.045
	3 years	0.240	0.295	0.241	0.236	3.150	3.546	3.160	3.083

Series	Training period	QLIKE				AMAE ($m = 2$)			
		AMEM (1,2)	AHAR	comb1	comb2	AMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	3.809	4.660	3.831	3.836	3.375	3.752	3.381	3.323
	3 years	3.809	4.660	3.838	3.891	3.375	3.752	3.383	3.349
rv kernel	4 years	3.855	4.810	3.877	3.834	3.364	3.800	3.376	3.270
	3 years	3.855	4.810	3.878	3.895	3.364	3.800	3.375	3.306

Table 3. Comparison among AMEM(1,2), AHAR and combination schemes (in bold the smallest values) - AEX dataset

Series	Training period	MSE				MAE			
		AMEM (1,2)	AHAR	comb1	comb2	AMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	0.251	0.293	0.251	0.247	3.003	3.170	2.993	2.960
	3 years	0.251	0.293	0.251	0.251	3.003	3.170	2.995	3.006
rv kernel	4 years	0.219	0.257	0.219	0.213	2.947	3.180	2.948	2.884
	3 years	0.219	0.257	0.219	0.216	2.947	3.180	2.950	2.922

Series	Training period	QLIKE				AMAE ($m = 2$)			
		AMEM (1,2)	AHAR	comb1	comb2	AMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	3.925	4.706	3.941	3.960	3.233	3.434	3.225	3.198
	3 years	3.925	4.706	3.949	4.023	3.233	3.434	3.228	3.240
rv kernel	4 years	3.854	4.677	3.872	3.834	3.153	3.419	3.155	3.095
	3 years	3.854	4.677	3.874	3.886	3.153	3.419	3.157	3.129

The results for AEX dataset (Table 3) are very similar. There is a predominance of *comb2* model, but the AMEM(1,1) performs well too particularly according to QLIKE (in three cases out of four), whereas *comb1* is chosen as best model once in both MAE and AMAE.

So far we have evaluated the available forecasts by means of the numerical values provided to the loss functions. Before doing so, however, we need to assess if the forecast series are different from a statistical point of view too. To this end, we have used the Conditional Predictive Ability (CPA) test (Giacomini and White 2006) to make pair-wise comparisons among all forecasting models ($\alpha = 0.05$).

The test shows different results according to the different datasets.³ As regards DAX, we accept the null hypothesis of equal accuracy for *comb1* and AMEM(1,1) (for both training periods) and for *comb2* and AMEM(1,1) (for the 3 years training period only). The same holds for CAC dataset, except for *rv* 5 minutes with a 4 years training period where AMEM(1,2) and both combining models have the same accuracy. Findings are rather dissimilar for AEX dataset: looking at *rv* 5 minutes, we can observe that no model (neither combination or single model) is better than others, whereas in the *rv* kernel case we reject the alternative hypothesis for the comparisons between *comb1* and AMEM(1,2), AMEM(1,2) and AHAR (for both 3 and 4 years training periods) and *comb2* and AMEM(1,2) (for the 3 years training period only).

3.2 APMEM vs AHAR

In this subsection we want to inspect if a generalization of the AMEM basic model is able to improve the accuracy of the forecasts of combinations. According to the Ljung-Box test, we use an APMEM(1,1) for DAX and an APMEM(1,2) for CAC and AEX.

As shown in Table 4, we have observed an actual improvement in the combined forecasts. Compared to the findings shown in Table 1, this time all the loss functions appoint the smallest value to a combination. The best combination scheme is *comb2* again, with *comb1* preferred only by QLIKE three times out of four.

Unfortunately, the improvement occurred for DAX dataset moving from AMEM to APMEM does not hold for CAC. Indeed the results shown in Table 5 are the same that we can draw from the Table 2 in terms of loss functions values.

Observing the Table 6 we can see that, compared to the Table 3, the transition from AMEM to APMEM has caused an agreement of the loss functions towards the choice of *comb2*, with the single model APMEM(1,2) that has been chosen by QLIKE only (overall, they are the same findings that can be draw from Table 5)

As before, we have measured the statistical meaning of the different forecasts by using the CPA test. Concerning the first two datasets, the results are perfectly

³ For lack of space, tables providing the findings are not shown herein.

Table 4. Comparison among APMEM(1,1), AHAR and combination schemes (in bold the smallest values) - DAX dataset

Series	Training period	MSE				MAE			
		APMEM (1,1)	AHAR	comb1	comb2	APMEM (1,1)	AHAR	comb1	comb2
rv 5 min.	4 years	0.249	0.293	0.249	0.245	3.421	3.671	3.418	3.332
	3 years	0.249	0.293	0.249	0.247	3.421	3.671	3.421	3.383
rv kernel	4 years	0.203	0.251	0.203	0.198	3.101	3.386	3.099	3.006
	3 years	0.203	0.251	0.203	0.200	3.101	3.386	3.101	3.059
Series	Training period	QLIKE				AMAE ($m = 2$)			
		APMEM (1,1)	AHAR	comb1	comb2	APMEM (1,1)	AHAR	comb1	comb2
rv 5 min.	4 years	4.165	4.979	4.152	4.156	3.640	3.925	3.637	3.564
	3 years	4.165	4.979	4.162	4.232	3.640	3.925	3.640	3.607
rv kernel	4 years	3.464	4.312	3.448	3.429	3.274	3.599	3.271	3.187
	3 years	3.464	4.312	3.456	3.506	3.274	3.599	3.273	3.236

Table 5. Comparison among APMEM(1,2), AHAR and combination schemes (in bold the smallest values) - CAC dataset

Series	Training period	MSE				MAE			
		APMEM (1,2)	AHAR	comb1	comb2	APMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	0.246	0.292	0.245	0.240	3.177	3.502	3.191	3.091
	3 years	0.246	0.292	0.246	0.242	3.177	3.502	3.193	3.113
rv kernel	4 years	0.242	0.295	0.242	0.234	3.158	3.546	3.172	3.045
	3 years	0.242	0.295	0.242	0.236	3.158	3.546	3.173	3.077
Series	Training period	QLIKE				AMAE ($m = 2$)			
		APMEM (1,2)	AHAR	comb1	comb2	APMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	3.833	4.660	3.868	3.855	3.389	3.752	3.405	3.318
	3 years	3.833	4.660	3.877	3.913	3.389	3.752	3.407	3,340
rv kernel	4 years	3.865	4.810	3.893	3.840	3.372	3.800	3.387	3.271
	3 years	3.865	4.810	3.899	3.903	3.372	3.800	3.389	3.301

Table 6. Comparison among APMEM(1,2), AHAR and combination schemes (in bold the smallest values) - AEX dataset

Series	Training period	MSE				MAE			
		APMEM (1,2)	AHAR	comb1	comb2	APMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	0.256	0.293	0.254	0.247	3.022	3.170	3.003	2.945
	3 years	0.256	0.293	0.254	0.250	3.022	3.170	3.005	2.989
rv kernel	4 years	0.219	0.257	0.219	0.213	2.950	3.180	2.949	2.878
	3 years	0.219	0.257	0.220	0.215	2.950	3.180	2.950	2.916
Series	Training period	QLIKE				AMAE ($m = 2$)			
		APMEM (1,2)	AHAR	comb1	comb2	APMEM (1,2)	AHAR	comb1	comb2
rv 5 min.	4 years	3.946	4.706	3.971	3.975	3.253	3.434	3.236	3.186
	3 years	3.946	4.706	3.980	4.039	3.253	3.434	3.239	3.226
rv kernel	4 years	3.855	4.677	3.875	3.831	3.156	3.419	3.156	3.091
	3 years	3.855	4.677	3.879	3.886	3.156	3.419	3.158	3.125

equal to those depicted in the previous subsection. As regards AEX, the only difference with findings mentioned above lies in the *rv* 5 minutes with a 4 years training period forecasts. Here, we accept the null hypothesis that all competing models are equally accurate on average, except the comparison between *comb2* and AHAR.

4 Conclusions

A comparison among forecasts provided by single and combination models is investigated. We have found that combining the AHAR model with APMEM instead of AMEM causes an improvement in the accuracy of the forecasts computed using combination schemes, especially the *comb2* model. This finding holds for DAX and AEX datasets and for all training periods, whereas for the CAC index there was not any change in loss function choices when moving from AMEM to APMEM.

References

- Ané, T. (2006). “An analysis of the flexibility of asymmetric power GARCH models”. In: *Computational Statistics & Data Analysis* 51.2, pp. 1293–1311.
- Corsi, F. (2009). “A simple approximate long-memory model of realized volatility”. In: *Journal of Financial Econometrics* 7.2, pp. 174–196.
- Ding, Z., C.W.J Granger, and R.F. Engle (1993). “A long memory property of stock market returns and a new model”. In: *Journal of empirical finance* 1.1, pp. 83–106.
- Engle, R.F. (2002). “Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models”. In: *Journal of Business & Economic Statistics* 20.3, pp. 339–350.
- Engle, R.F. and G.M. Gallo (2006). “A multiple indicators model for volatility using intra-daily data”. In: *Journal of Econometrics* 131.1, pp. 3–27.
- Giacomini, R. and H. White (2006). “Tests of conditional predictive ability”. In: *Econometrica* 74.6, pp. 1545–1578.
- Oxford-Man Institute of Quantitative Finance Realized Library (2017). URL: <http://realized.oxford-man.ox.ac.uk/data>.
- Raviv, E. (2016). “Forecast combinations in R using the ForecastCombinations package A Manual”.
- Timmermann, A. (2006). “Forecast combinations”. In: *Handbook of economic forecasting* 1, pp. 135–196.

Time series and artificial intelligence with a genetic algorithm hybrid approach for rare earth price prediction.

Fernando Sánchez Lasheras¹, Sergio Luis Suárez Gómez², María Victoria Riesgo García³, Alicja Krzemień⁴, Ana Suárez Sánchez⁵,

¹ Department of Construction and Manufacturing Engineering, University of Oviedo, Gijón, Spain

² Prospecting and Exploitation of Mines Department. University of Oviedo, Oviedo. Spain

³ Doctorate Program on Industrial Technologies and Civil Engineering, Polytechnic School, University of Burgos, Burgos, Spain.

⁴ Department of Risk Assessment in Industry, Central Mining Institute (GIG), Katowice, Poland.

⁵ Department of Business Administration, University of Oviedo, Spain

Correspondence: sanchezfernando@uniovi.es

Abstract. Profitable exploitation of rare earths is unusual. These elements are essential for new technologies as well as for the development of the current world economy. The need for rare earths and their demand on markets between 2006 and 2014 was roughly constant, without decrease despite the global crisis, and demand is expected to increase. However, fluctuations in the need for the various rare earths will make it difficult to keep the market equilibrium. With the aim of forecasting the behavior of market prices, multivariate time series processes are used to model and forecast the prices of one of the rare earths, terbium. Although the results achieved are satisfactory, an alternative model is proposed, based on univariate time series for the predictions, together with a support machine model for the multivariate relations with their parameters set with a genetic algorithm, thus improving the forecasts.

Keywords: rare earth; time series; support vector machines; genetic algorithms.

1 Introduction

There are fifteen chemical elements called rare earths or lanthanides. They can be classified as light and heavy rare earths, according to their atomic weight. Light rare earths include the following chemical elements: lanthanum, cerium, praseodymium, neodymium, promethium and samarium; the heavy rare earths are terbium, dysprosium, holmium, erbium, thulium, ytterbium and lutetium. Finally, scandium and yttrium, although not true rare earths, are often considered as such because of their physical and chemical properties. Rare earths follow the Oddo-Harkins law, whereby the percentage composition in which they are found in the different minerals containing them decreases as their atomic number increases. It means that light rare earths are more abundant than heavy rare earths [1].

Rare earths are not common in quantities sufficient to allow profitable exploitation. In addition, these elements are essential for certain civil and military technologies as well as for the development of the current world economy, including wind turbines, catalysts, glassmaking, metallurgy, aerospace, health care and advanced battery systems. According to the Worldwide Threat Assessment [2], 90% of world rare earth mining is now concentrated in China. The European Union has classified some of the rare earths into the list of "critical raw materials" [3]. The importance of light and heavy rare earths is very similar, but heavy rare earths present a much higher risk of supply, due to China's influence in terms of global supply, since it provides 99% of them, compared to 87% of the lightweight ones.

The British Geological Survey [4] also classifies rare earths within the category of "critical materials" with a relative supply risk index of 9.5, on a scale ranging from 1 to 10, due mainly to the concentration of production, the distribution of reserves, the rate of recycling and the difficulty of substitution by other minerals. The United States Department of Energy, in a 2011 report [5], also gave them a high importance. In the study entitled "Critical Material Strategy", neodymium, europium, terbium, dysprosium and yttrium were identified as the rare earths critical for the short and medium term. Dysprosium was the rare earth selected as of major importance for clean energies and with the greatest risk of supply in the medium term, followed by terbium, and neodymium.

According to ERECON (2014) [6], the estimated total consumption of rare earths in Europe in 2012 was 113,250 tons, while consumption in 2010 was 8,050 tons. The historical consumption of rare earths between 2006 and 2014 was roughly constant, without any decrease despite the global crisis, and demand is expected to increase [7] in the coming years.

Although fluctuations in the demand for the various rare earths will make it difficult to achieve market equilibrium [8], in 2014 there were five mining projects in development that would have meant an increase in the supply of rare earths by some 41,100 tons per year, so that these projects alone could have covered about one third of the

total consumption of rare earths in the world. In 2016 and after a fall in prices, some of the projects mentioned were paralyzed [9].

Finally, we would like to point out that there is currently no substitutability for about 45% of rare earth applications, while for another 45% this is only possible at a high cost or with loss of performance.

In this paper, we present an alternative methodology to the classical time series techniques, with the aim of improving the forecasts over the series of prices of the rare earths introduced above. This approach includes the use of artificial intelligence methods to support the modelling of the series, whose parameters are set with the support of a genetic algorithm.

2 Materials and Methods

2.1 Univariate and Multivariate Time Series models.

Univariate time series processes ARMA (autoregressive moving average) define a parametric family of stationary processes that make it possible to model the time structure and behavior of large sets of data measured in homogeneous time lapses. Therefore, these processes also enable predictions for the modelled data [10,11].

The process $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ is called an ARMA(p, q) process if $\{X_t\}$ is stationary and, for all t ,

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q} \quad (1)$$

Where $\{Z_t\}$ is a white noise process of uncorrelated observations with mean zero. Also, Φ and θ are polynomials of degrees p and q respectively, defined as follows,

$$\Phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p \quad (2)$$

$$\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q \quad (3)$$

The equation 1 is usually reformulated as

$$\Phi(B)X_t = \theta(B)Z_t, \quad t = 0, \pm 1, \pm 2, \dots \quad (4)$$

With B the backward shift operator defined as

$$B^j X_t = X_{t-j}, \quad j = 0, \pm 1, \pm 2, \dots \quad (5)$$

As a generalization of ARMA models, ARIMA models include the possibility of modeling non-stationary series due to the existence of tendency, with a differentiation order d . Therefore, the process $\{X_t\}$ will be an ARIMA(p, d, q) if,

$$\Phi(B)(1 - B)^d X_t = \theta(B)Z_t \quad (6)$$

The extension to multivariate modeling, in the case of time series with relations between variables, can be modeled with multivariate processes for vectors of time series, VARMA (vector autoregressive moving average) [12,13].

A k -dimensional process $\{X_t\}$, with $X_t = (X_{t1}, \dots, X_{tk})$, is generated with a stationary and invertible model, VARMA(p, q), if it satisfies the equation,

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t - \theta_1 Z_{t-1} - \dots - \theta_q Z_{t-q} \quad (7)$$

With coefficient matrixes ϕ_i and θ_j , for $i = 1, \dots, p$, $j = 1, \dots, q$ and $\{Z_t\} = (Z_{t1}, \dots, Z_{tk})$ vector of white noise processes.

2.2 Support Vector Machines.

The support vector machines (SVM) are machine learning techniques. This category of techniques is known for its aptitude for approximating multivariate functions [14,15]. Among other mathematical models for similar problems [11,16], SVM belong to the family of machine learning methods and are used to model different types of physical systems through the adaptation of their parameters [17,18] with a training process. These methods are commonly used in classification [19] and regression problems [20,21]. The performance of SVM relies on the training with data previously obtained from the system to be modelled, as training data set.

For classification problems, the vectors from the training data are used to map a feature space of a higher dimension, which depends on the kernel function selected. The classification is made through the separation of classes defined by hyperplanes. These hyperplanes are constructed as an optimized linear solution.

Then, the output of a trained SVM can be formulated as [15,22]:

$$\hat{y}_i = a^T \Phi(x_i) + b \quad (8)$$

With x_i the input vectors from the training set, mapped into a hyperplane via $\Phi(x_i)$, a function that linearizes the relations between inputs and outputs. The parameters are a and b , which are a vector of the same dimension as the image of Φ , and a coefficient, respectively.

The determination of the parameters is made by finding an optimized solution to the problem [23]

$$\min_{a, \varepsilon, \eta_i, \eta'_i} \frac{1}{2} a^T a + C \left(\frac{1}{N} \sum_{i=1}^N (\eta_i + \eta'_i) + v \varepsilon \right) \quad (9)$$

Where the restrictions of the optimization problem are:

$$a^T \Phi(x_i) + b - y_i \leq \varepsilon + \eta_i \quad (10)$$

$$y_i - a^T \Phi(x_i) - b \leq \varepsilon + \eta'_i \quad (11)$$

Where C is a regularization parameter, ε is the tolerance error for each input x_i . Both η and η' are the slack variables, that take positive values. Finally, v is a parameter for the adjustment of the tolerance. The estimation of SVM can be expressed as

$$\hat{y}_i = \sum_{i=1}^N (\beta'_i - \beta) K(x_i, x) + b \quad (12)$$

In this expression, β and β' are the Lagrange multipliers corresponding to the restrictions above. The kernel function K can be defined as $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$.

Some of the most common kernel functions are the following [24]:

- Linear kernel; $K(x_i, x_j) = x_i^T \cdot x_j$.
- Polynomial kernel; with parameters γ , α_0 and, as the polynomial degree, α . This is defined as $K(x_i, x_j) = (\gamma \cdot x_i^T \cdot x_j + \alpha_0)^\alpha$
- Radial basis kernel; with γ as parameter, and expression $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$.
- Sigmoid kernel; with parameters γ and α_0 . It is defined as $K(x_i, x_j) = \tanh(\gamma \cdot x_i^T \cdot x_j + \alpha_0)$.

2.3 Genetic algorithms.

Genetic Algorithms were developed to simulate the evolution of a population in terms of optimizing the survival of the next generation. These algorithms were first developed for chromosomic studies [25], but now, their use has been generalized for optimization problems. Consequently, genetic algorithms work to improve the fitness of the iterations, that is, of the formation of new generations, until a solution is reached. Four basic genetic operators are used in each step as criteria [26–28]:

- **Crossover:** to modify the programming of the chromosomes from one generation to the next. Its behavior is similar to that of biological crossover.
- **Mutation:** used to provide genetic diversity from one generation to the next, randomly altering values of genes before the crossover operation. Usually, uniform random mutation is used.
- **Reproduction:** in each generation, two solutions are selected and mixed with crossover or mutation techniques, in order to create a child solution for the next generation.
- **Elitism:** This is used as accelerator criterion for improving the fitness function, allowing the genetic algorithm to clone the best genomes from one generation to the next.

The optimization problem and the iterations of the genetic algorithm are widely known and formulated in literature [29].

In our case the aim of the algorithm is to find adequate parameters for the SVM. The iterations of the genetic algorithm are focused in minimizing the mean absolute error over the predictions. The steps of the algorithm are as shown in figure 1:

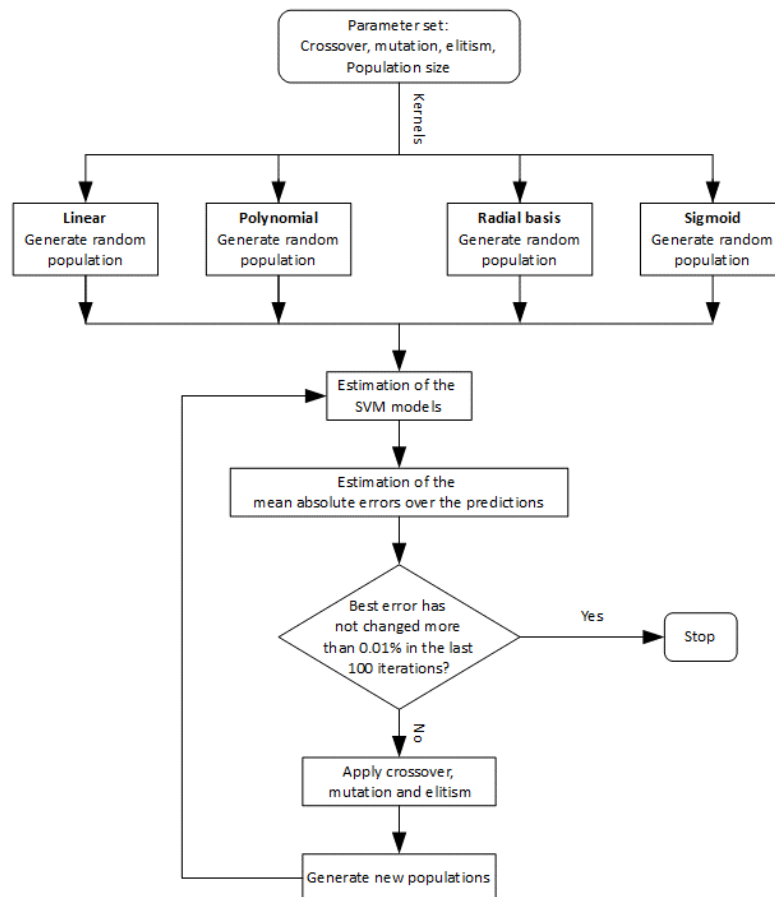


Fig. 1. Genetic algorithm for setting optimal SVM parameters.

The algorithm begins with the setting of parameters for the genetic algorithm, such as crossover, mutation, elitism and population size. After this, an initial population should be created. The set of the initial population has to be a vector with size of the possible variation of parameters for the SVM.

Since we will consider, among the possible variable parameters, the type of kernel used for the SVM, we will have different types of initial populations, based on the type

of kernel. To avoid this problem, we consider a branching of the algorithm. Considering four ways in parallel, in each branch, the genetic algorithm is performed with each possible type of kernel (linear, polynomial, radial basis or sigmoid).

The stop criterion will be satisfied if the error does not change more than 0.01% in the last 1000 iterations of the algorithm. When the stop criterion is not satisfied, a new population has to be created. This is performed by implementing the crossover, mutation and elitism. The process is repeated until the stop criterion is satisfied.

2.4 Data available for the study.

Data is in the form of time series, taken monthly, of the prices of 5 rare earths. The length of the series is not equal; the first and last data can be found on table 1. Once the measurement began, it was taken every month until the last one; therefore there are no missing values.

Table 1. Characteristics of the price series

Series	Rare earth	First date	Last date	Length
1	Dysprosium	March-2003	February-2017	167
2	Europium	March-2003	February-2017	167
3	Neodymium	March-2003	February-2017	167
4	Praseodym	March-2003	February-2017	167
5	Terbium	January-2004	February-2017	127

For the study, data are separated into two sets, a training set and a validation set, for estimating errors in the prediction of the developed models. The validation set will have the last 5 measurements of each series, and the training set will have the rest of the measurements.

3 Results and discussion.

Data considered are time series of prices, which are supposed to be related among themselves. With this hypothesis, a multivariate approach is required in order to model and consider the possible relationships between the time series of prices. Our first approach was to consider VARMA models, where data was available for the whole series (from $t = 25$ to $t = 162$).

However, a search in the usual indicators of fitted VARMA models, such as AIC, BIC or the mean absolute error of predictions over the validation set, led us to discard the models that have a moving average component.

The available data corresponding with the terbium price series is the one to be predicted, with values ranging from a minimum of 196.35 to a maximum of 5,900 US dollars. With VARMA models, the mean absolute error of the predictions represented a high percentage of the range of values of the data of the Terbium price series, 5,703.65 dollars.

Consequently, the modeling was restricted to VAR processes. Mean absolute errors over the prediction of Terbium series data in the validation set are presented in Figure 2.

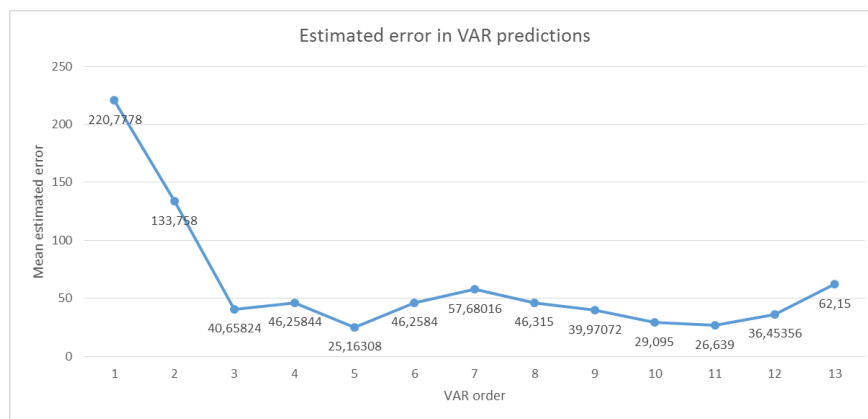


Fig. 2. Estimated mean absolute error over predictions of the terbium series.

The best option in the VAR models was found in VAR(5). Considering models of higher complexity, VAR(11) gave close results; nevertheless, in terms of both complexity and results, VAR(5) is the most appropriate model, with 25,163 US dollars as the mean absolute error.

However, in order to achieve better results, another approach was performed. To improve the performance, we consider that, due to the low quantity of available data, it could be inferred that both processes, prediction and modeling the multivariate relations, could be difficult to perform with only a model. With this aim, a hybrid approach is proposed, with time series processes and artificial intelligence techniques.

Univariate time series processes were used to model each of the first four series of prices. This also makes it possible to use more data for the modeling, as explained before. Therefore, the predictions for each single series are expected to be more accurate than those in the multivariate approach. Regarding the usual indicators, such as AIC and BIC, the best models with the simplest complexity for each series is found in Table 2.

Table 2. ARIMA processes for modeling each univariate series.

Series	1	2	3	4
ARIMA(p,d,q)	(0,1,0)	(0,1,0)	(1,1,0)	(1,1,0)

It is interesting to remark the fact that the multivariate approach required some level of complexity, whereas univariate modeling requires processes of low complexity, even when modeling all the series with just an order of differentiation, as is the case in the first two series. This led us to set an artificial intelligence model, the SVM, to model the multivariate relations.

Once univariate time series models were set to model each of the four series and predict their values, a SVM model was trained to predict the value of the fifth series, having as input the corresponding values predicted from the first four series. The method is detailed in Figure 3.

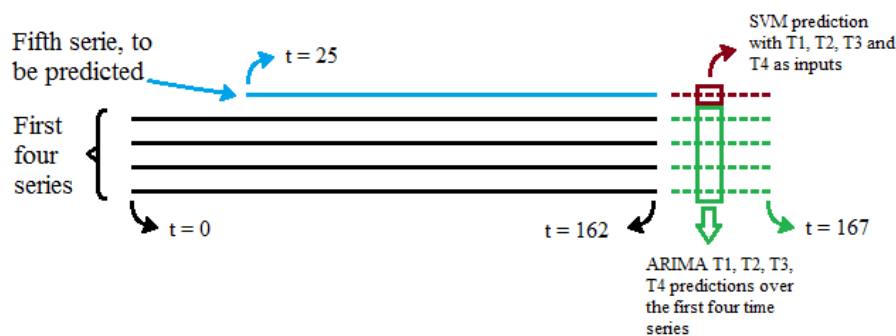


Fig. 3. Scheme of the method followed.

The SVM was trained to model the relation between series, predicting the value of the fifth series. In order to do so, the input data was the values of the first four series of the training set, and the values from the fifth series were the output. Data was taken from $t = 25$ to $t = 162$, where data from the fifth series was available.

To select the most adequate parameters of the SVM, a genetic algorithm was used, with a search in all the possible parameters and with four simultaneous branches, in the four possible types of kernels, since each kernel has a different number of parameters.

For the estimation of the performance, the mean absolute error in the fifth series validation data was used. The inputs of each of the SVM models were the predictions provided from each ARIMA model, and the outputs compared to the validation data of the fifth series.

The parameters obtained as result of the genetic algorithm determined the best SVM for the data considered, with a SVM of v -regression type, with $v = 1$. The chosen kernel

was polynomial, with degree 3, $\gamma = 2$ and independent parameter $\alpha_0 = 5$. Regarding the general parameters of the SVM, cost was set at $C = 1.12$ and $\varepsilon = 0.1$.

The model of SMV here presented, gives as result, a total of 9,843 dollars as average absolute error of the predictions from the fifth series, which represents 0.172% of the total variation of the price, improving significantly the forecasts from classical time series methods.

4 Conclusions

The prices of the rare earths were modeled with two methodologies. First, a multivariate time series approach with VARMA procedures was used. These techniques made it possible to model the time series and predict using the fitted model at the same time. This gave a mean absolute error of 24.16 dollars as result.

Despite being able to model and predict satisfactorily with the multivariate technique, another approach was performed, with a mixture of time series and artificial intelligence techniques.

Modeling and forecasting with ARIMA processes each of the first series individually made it possible to use more information for the modeling and consequently, to achieve more accurate forecasts. The SVM is trained to learn the relationships within the series, and to predict the value of the terbium series.

Using the predictions from the ARIMA processes as inputs for the SVM, the mean absolute error obtained for the predictions was 9,843 dollars. This represents a range of terbium prices of 0.172%, which is a significant improvement on the best predictions that can be obtained with the multivariate approaches.

The improvements that provide the stated hybrid approach, additionally to the good results presented, give the hints to the development of the method for its general use, by means of a deep study, considering computational cost and times, and the comparison with other methodologies applied in other fields, or deeply in this kind of economic studies.

5 References.

1. Kilbourn, B. T. A Lanthanide Lanthology: A Collection of Notes Concerning the Lanthanides and Related Elements. Molycorp. Inc., Mt. Pass, CA, USA **1993**.
2. Clapper, J. C. *Worldwide threat assessment of the US intelligence community*; 2013;
3. Chapman, A.; Arendorf, J.; Castella, T.; Thompson, P.; Willis, P.; Espinoza, L. T.; Klug, S.; Wichmann, E. Study on Critical Raw Materials at EU Level. *Oakdene Hollins Buckinghamshire, UK* **2013**.
4. British Geological Survey. Risk List 2015. **2015**.
5. Bauer, D.; Diamond, D.; Li, J.; Sandalow, D.; Telleen, P.; Wanner, B. US Department of Energy Critical Materials Strategy. **2011**.

6. ERECON (2014). Strengthening the European Rare Earths Supply Chain: Challenges and Policy Options. European Rare Earths Competency Network. Kooroshy, J.; Tiess, G.; Tukker, A.; Walton, A. (eds.).
7. Massari, S.; Ruberti, M. Rare earth elements as critical raw materials: Focus on international markets and future strategies. *Resour. Policy* **2013**, *38*, 36–43.
8. Binnemans, K.; Jones, P. T.; Van Acker, K.; Blanpain, B.; Mishra, B.; Apelian, D. Rare-earth economics: the balance problem. *Jom* **2013**, *65*, 846–848.
9. Barakos, G.; Gutzmer, J.; Mischo, H. Strategic evaluations and mining process optimization towards a strong global REE supply chain. *J. Sustain. Min.* **2016**, *15*, 26–35.
10. Brockwell, P. J.; Davis, R. A. *Time series: theory and methods*; Springer Science & Business Media, 2013;
11. Suárez Gómez, S. L.; Santos Rodríguez, J. D.; Iglesias Rodríguez, F. J.; de Cos Juez, F. J. Analysis of the Temporal Structure Evolution of Physical Systems with the Self-Organising Tree Algorithm (SOTA): Application for Validating Neural Network Systems on Adaptive Optics Data before On-Sky Implementation. *Entropy* **2017**, *19*, 103.
12. Baxter, M.; King, R. G. Measuring business cycles: approximate band-pass filters for economic time series. *Rev. Econ. Stat.* **1999**, *81*, 575–593.
13. Suárez Gómez, S. L. Técnicas estadísticas multivariantes de series temporales para la validación de un sistema reconstructor basado en redes neuronales. **2016**.
14. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
15. Vapnik, V. N.; Vapnik, V. *Statistical learning theory*; Wiley New York, 1998; Vol. 1;.
16. González-Gutiérrez, C.; Santos, J. D.; Martínez-Zarzuela, M.; Basden, A. G.; Osborn, J.; Díaz-Pernas, F. J.; De Cos Juez, F. J. Comparative Study of Neural Network Frameworks for the Next Generation of Adaptive Optics Systems. *Sensors* **2017**, *17*, 1263.
17. Osborn, J.; Guzman, D.; Juez, F. J. D. C.; Basden, A. G.; Morris, T. J.; Gendron, E.; Butterley, T.; Myers, R. M.; Guesalaga, A.; Lasheras, F. S.; Victoria, M. G.; Rodríguez, M. L. S.; Gratadour, D.; Rousset, G. Open-loop tomography with artificial neural networks on CANARY: On-sky results. *Mon. Not. R. Astron. Soc.* **2014**, *441*, 2508–2514, doi:10.1093/mnras/stu758.
18. Basden, A. G.; Atkinson, D.; Bharmal, N. A.; Bitenc, U.; Brangier, M.; Buey, T.; Butterley, T.; Cano, D.; Chemla, F.; Clark, P.; others Experience with wavefront sensor and deformable mirror interfaces for wide-field adaptive optics systems. *Mon. Not. R. Astron. Soc.* **2016**, *459*, 1350–1359.
19. Vilán, J. A. V.; Fernández, J. R. A.; Nieto, P. J. G.; Lasheras, F. S.; de Cos Juez, F. J.; Muñiz, C. D. Support vector machines and multilayer perceptron networks used to evaluate the cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain). *Water Resour. Manag.* **2013**, *27*, 3457–3476.
20. Lasheras, F. S.; Nieto, P. J. G.; de Cos Juez, F. J.; Bayón, R. M.; Suárez, V. M.

- G. A hybrid PCA-CART-MARS-based prognostic approach of the remaining useful life for aircraft engines. *Sensors* **2015**, *15*, 7062–7083.
21. Sánchez, A. S.; Iglesias-Rodríguez, F. J.; Fernández, P. R.; de Cos Juez, F. J. Applying the K-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders. *Int. J. Ind. Ergon.* **2016**, *52*, 92–99.
 22. Schölkopf, B.; Smola, A. J.; Williamson, R. C.; Bartlett, P. L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245.
 23. Lasheras, F. S.; Nieto, P. J. G.; de Cos Juez, F. J.; Vilán, J. A. V. Evolutionary support vector regression algorithm applied to the prediction of the thickness of the chromium layer in a hard chromium plating process. *Appl. Math. Comput.* **2014**, *227*, 164–170.
 24. Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. Misc Functions of the Department of Statistics (e1071), TU Wien. *R Packag. version* **2005**, 1–5.
 25. Holland, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*; MIT press, 1992;
 26. De Falco, I.; Della Cioppa, A.; Tarantino, E. Mutation-based genetic algorithm: performance evaluation. *Appl. Soft Comput.* **2002**, *1*, 285–299.
 27. Fernández, J. R. A.; Muñiz, C. D.; Nieto, P. J. G.; de Cos Juez, F. J.; Lasheras, F. S.; Roqueñí, M. N. Forecasting the cyanotoxins presence in fresh waters: A new model based on genetic algorithms combined with the MARS technique. *Ecol. Eng.* **2013**, *53*, 68–78.
 28. Ting, C.-K.; Su, C.-H.; Lee, C.-N. Multi-parent extension of partially mapped crossover for combinatorial optimization problems. *Expert Syst. Appl.* **2010**, *37*, 1879–1886.
 29. Galán, C. O.; Lasheras, F. S.; de Cos Juez, F. J.; Sánchez, A. B. Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. *J. Comput. Appl. Math.* **2017**, *311*, 704–717.

Predicting the financial status of companies using data balancing and classification methods

H. Jawazneh¹, A.M. Mora¹, and P.A. Castillo¹

Departamento de Arquitectura y Tecnología de Computadores.
Universidad de Granada (Spain)
huthaifa.jawazneh@gmail.com , {amorag,pacv}@ugr.es

Abstract. Bankruptcy prediction problem is not new. Over the past 50 years, researchers have become increasingly interested in this problem, as it is a critical issue to forecast for companies. Taking this into account, in this work an efficient solution to the bankruptcy prediction problem is presented. To this end, a real dataset from Spanish companies are used, dealing with the present unbalance on the data applying some data resampling techniques. Then, several well-known classifiers, i.e. J48, Random forest and Naïve Bayes, are compared in order to figure out which one obtains the highest prediction accuracy to be adopted as appropriate solution. The judgment about the most ideal and convenient classifier to do this job is based on several metrics, which are the outcomes of four experiments. Random forest classifier obtains the superior results according to the outcomes.

1 Introduction

Predicting financial status of companies is a very important problem, since there are numerous concerned stakeholders attentive to this information. For instance, the investors are usually very interested in the information about future investments. On the other hand, many firms and organizations care about the financial state prediction to do some studies. Also, many companies need a financial coverage from other companies or firms, and these cases the creditor company cares about whether the debtor company is a solvent or not.

Several techniques have been used before to solve the problem of predicting an incoming bankruptcy situation. Part of them was statistical, which made more effort in case of functional relations between dependent and independent variables [2]. In this paper, we use artificial intelligence methods, in order to predict the financial status of the companies considering a dataset of many companies in Spain.

This study comes as an improvement of our previous works, which applied a hybrid technique to solve the problem of detecting the financial state using Kohonen's Self-organizing Map + U-Matrix graph to determine the influence of the variables on the different clusters present in the data. This was combined respectively with Genetic Programming [2] and Support Vector Machines [15],

to classify the data regarding the book losses of the companies. Due to the constraints of the applied techniques, just numerical data were considered. That gives the current work a preference while the methods used also deal with categorical variables, related to financial and nonfinancial data. Thus, in this study, we have considered a wider dataset, enriched by more information, which might let to obtain better results. In addition, the use of decision trees will yield more interpretable outputs than those obtained in previous works, from the point of view of an expert who has to justify the decision of the system for a new company record.

J48, Random forest and Naïve Bayes classifiers addressed in this study in order to figure out the most convenient classifier to do this job. Four incremental experiment have been conducted, and their results analyzed and compared considering the values of several metrics such as the accuracy, sensitivity, specificity and false positive rate.

2 State of the art

Since 1966, many researchers have been interested in the financial status prediction problem. They attempted to innovate by applying useful tools or methods in order to be a solution.

In the case of using one prediction method or classifier for the same objective, several researchers used support vector machine (SVM) [7], which is well known artificial intelligence method invented to make a good prediction in various cases. It based on finding the separating hyperplane in the space between the classes with the largest margin. Thus, obtains a good accuracy and low generalization error. In this research line, Hui and Sun [8] depended on SVM to do an empirical study about the financial status of Chinese companies. Li and Sun [11] in their study improved the financial status prediction accuracy by using a straightforward wrapper approach in order to complement SVM. Also, Bose and Pal [3] made a study comparing several methods to forecast the financial status. They proved that neural networks accuracy is better than SVM for this aim.

On the other hand, many researches were focused on using more than one method simultaneously. Verikas et al. [18] proposed a solution for bankruptcy prediction problem relaying on hybrid and ensemble based soft computing techniques for prediction. Hybrid techniques constructed from combining different prediction methods in one, aims to enhance the accuracy of prediction. The ensembling technique achieved by combining the results of each prediction model according to majority voting, minimum, maximum and simple average to obtain optimized predictions.

Regarding the hybrid technique, numerous researchers focused on this technique as a robust and effective solution in order to predict the financial status. Some of them merge two classifiers, one of it assigned to select the features to the other classifiers. Yeh et al. [21] created two stage classifier by merging Rough set theory with SVM. Rough set used to eliminate the redundant attributes, then the bankruptcy state predicted by SVM. Lin et al. [12] eliminated the attributes

dimensionality using isometric feature-mapping algorithm; which is one of the most advanced dimensionality reduction techniques, then use SVM normally. Wu et al. [19] merged real valued genetic algorithm with SVM, so the attributes of SVM enhanced with real valued genetic algorithm, this classifier obtains better accuracy than pure SVM. Min et al. [13], also merge genetic algorithm with SVM to create a hybrid method with a different technique. They used the genetic algorithm to enhance the feature selection and SVM attributes, which gave better prediction ability to SVM classifier. Mora et al. [14] used the mutual-information criterion in order to select the variables in order to improve the capacity of SVM regarding to predicting the financial status, as an improvement of their previous work [15], which combining genetic programming, self-organizing map, evolutionary algorithms and neural networks in the aim of predicting the financial status too.

Also for ensembling technique, several researchers attended to use it as a convenient robust prediction solution. Jo and Han [9] used multivariate discriminant analysis, case-based reasoning and neural networks in order to reduce the bankruptcy prediction errors, the result of the three methods extracted by weighted sum scheme. Ravi et al. [17] also create a system for bankruptcy prediction by combining several classifiers results, which is probabilistic neural networks, SVM, decision trees and fuzzy rule.

As stated in the introduction, these data were also studied in previous works by the authors [2, 15], but the used variables were more limited, as they were just numerical and no data balancing methods were applied to the problem, which are highly recommended in order to obtain reliable results. Thus, in the present work, we have performed three data resampling approaches.

3 Problem description

In this study, we address the problem of predicting the financial status of the companies, transforming it into a classification one. We use a combination of the financial and non-financial data. Many previous studies based on some classifiers or predicting method were not effected with more than one type of data; numerical [14].

A dataset brought from the Infotel database has been used; it is a company in charge to gather information in several domains about companies in Spain. Data from 471 companies in Spain during six years sequentially (1998 to 2003) has been used. In this paper, several algorithms have been used in order to obtain good and accurate predictions about the financial state of the companies.

The dataset proposed in this work include particular domain attributes used in order to determine whether a firm succeed or fail. It includes 2859 instances, each one of them consists of 39 independent variables of different types (categorical and numerical). After removing meaningless variables (such as internal codes), we adopted 33 variables, 27 of them are numeric and the remaining are categorical. Some of these variables refer to financial information. Each record represents a company in one year, and have an attribute *Bankruptcy* to men-

tion the financial status for that firm. Table 1 shows the adopted independent variables in the dataset.

Table 1. Independent Variables

Financial Variables	Description	Type
Debt Structure	Long-Term Liabilities / Current Liabilities	Real
Debt Cost	Interest Cost / Total Liabilities	Real
Debt Paying Ability	Operating Cash Flow / Total Liabilities	Real
Debt Ratio	Total Assets / Total Liabilities	Real
Working Capital	Working Capital / Total Assets	Real
Warranty	Financial Warrant	Real
Operating Income Margin	Operating Income / Net Sales	Real
Return on Operating Assets	Operating Income / Average Operating Assets	Real
Return on Equity	Net Income / Average Total Equity	Real
Return on Assets	Net Income / Average Total Assets	Real
Stock Turnover	Cost of Sales / Average Inventory	Real
Asset Turnover	Net Sales / Average Total Assets	Real
Receivable Turnover	Net Sales / Average Receivables	Real
Asset Rotation	Asset allocation decisions	Real
Financial Solvency	Current Assets / Current Liabilities	Real
Acid Test	(Cash Equivalent + Marketable Securities + Net receivables) / Current Liabilities	Real
Non-financial Variables	Description	Type
Year	Corresponding to the sample	Integer
Size	Small—Medium—Large	Categorical
Number of employees		Integer
Age of the company		Integer
Type of company	Public Company—Limited Liability Company—Others	Categorical
Linked to a group	If the company is part of a group holding	Binary
Number of partners		Integer
Province code	Code of the location where the company is set	Categorical
Number of changes of location		Integer
Delay	If the company has submitted its annual accounts on time	Binary
Historic number of judicial incidences	Since the company was created	Integer
Number of judicial incidences	Last year	Integer
Historic amount of money spent on judicial incidences	Since the company was created	Real
Amount of money spent on judicial incidences	Last year	Real
Historic number of serious incidences	Such as strikes, accidents...	Integer
Audited	If the company has been audited	Binary
Auditor's opinion	Favourable—Exceptions—Unfavourable	Categorical

4 Methodology

In this paper, three classification algorithms have been used in order to predict the financial status of several companies.

J48 classifier is an improvement of C4.5 classifier [16], and both of them are extensions of ID3 decision tree algorithm. The main objective of J48 classifier is to implement the training dataset into a decision tree based on the number of attributes in it. While J48 classifier creates the decision tree it ignores all of

the missing values because they are valueless. The procedure of J48 predicting stands on the known attributes values [16], and handle the discrete and the contiguous data, then pruning the decision tree if there are some branches which do not help in order to reach the leaf node[10]. J48 algorithm runs according to particular steps. The first step is if all of the records in the dataset belong to the same class, so the tree is a leaf, this leaf will be labeled with the same class. The second step is calculating the information of the attributes given by applying a test on it depending on the probability of the attribute value in each record, then the information Gain calculation relaying on the information given by applying the tests. The last step is to select the best attributes regarding to the information gain calculated in the previous step. The final touch on the decision tree after the full creation of it, and before performing the classification, is to remove a discordant information, which is far away from the majority of data and adversely affect data classification. This process called pruning, it is very important to improve the accuracy of the prediction while many datasets may contain this type of useless data [10].

Naïve Bayes classifier is a probabilistic method used to assign the class of each record relaying on calculating the probability of each attribute independently from the other attributes. In other words, Naïve Bayes assume that the effect of each attribute value is detached from other attributes values on predicting the class. This classification method presents high accuracy and efficiency when applied on large training set by calculating the frequencies and combining the values to make a good decision about the predicted class for each records[20].

Also **Random forest classifier** used in this study, It is an ensemble classification method, which creates several individual random training subsets from the original dataset [1]. Training these subsets creates several decision trees constructing the random forest, each instances class in the test set predicts independently in each decision tree. The final results of the instances class relay on the majority voting of these decision trees. As an initial step in Random forest classifier bagging uses to split the original dataset into training and test set by taking a partition of data as a training and the remaining is the test set. In other words, some instances will be selected from the original dataset by sampling and replacement method to be a training dataset and the remaining instances that defined as Out-Of-Bag considered to be a test set. Creating a decision tree for each subset basically depend on C4.5 algorithm that stand on information Gain and Entropy. The last mission of Random Forest classifier is to gather the subtrees with each other to create the forest, the classification result is the average of class probabilities obtained from all the training trees [4].

5 Experiments and results

This section discusses the differences between J48, Random forest and Naïve Bayes classifiers, considering the obtained results in four different (and incremental) experiments, using Weka¹ (a machine learning software suite).

¹ <http://www.cs.waikato.ac.nz/ml/weka/index.html>

The dataset used in these experiments is extremely unbalanced, it contains 2797 records labeled with healthy companies class and 62 records labeled with bankrupt companies, which is around 98% / 2% out of the whole dataset. This situation creates a challenge for classifiers to work properly.

The solution to prepare such an extremely unbalanced dataset is to use data balancing (or data resampling) methods. That stands on changing the size of the original dataset to get the most proper and optimum dataset to evaluate and training the classifiers [5]. Three methods have been applied here to balance the data: the first one is oversampling method, it stands on creating a superset from the original dataset by replicating some random instances to achieve the desired distribution. The second one is undersampling method, it stands on creating a subset of the original dataset by removing some records randomly. The third is a hybrid approach, which combines the two previous methods. Thus, it stands on removing some valueless records and replicating another part of the training dataset to achieve a fairer distribution of classes in the samples [6].

Actually, four metrics adopted in order to make a judgment about the classifiers; the first one is the *accuracy*, which represents the ability of the classifier to assign the correct class to each instance. The second is the *sensitivity*, It represents the capacity of the classifier regarding to assign the company to the bankruptcy class (prediction) while it is actually bankrupt (real status). The third metrics is the *specificity*. It represents the capacity of the classifier to assign the companies to the succeed class (prediction) while it is actually that (real status). The last metric is the *false positive rate (FPR)*, which represents the failure of the classifier in assigning bankrupt companies to bankruptcy class (wrong prediction), while their actual class is bankruptcy (real status), this metric is a complement value of specificity, both of them have the same Standard deviation value. In other words, the superior classifier gives the maximum accuracy, sensitivity and specificity, and the minimum false positive rate.

The experiments addressed in this section considered in a certain sequence in order to figure out the most convenient to reliant regarding the dataset circumstances; unbalanced dataset, and achieve the required criteria.

The first experiment solves the problem of the balancing by partitioning the dataset to several equally subsets under the coverage of the balancing techniques, i.e. (undersampling and oversampling), to makes the subsets balanced (not exactly). in other hand, the cross validation creates a problem regarding to the reliability; in the case of existing mutual records in test and training folds. To avoid the problem of the reliability, the need of the next experiment arose, it's based on splitting the original dataset to training and test sets, but unfortunately the problem of the metrics values inconsistency appeared due to the balancing problem in the training set; the classifier selects unrequired procedure. To improve this technique the experiment 3 based on using the balancing techniques in the training set of the previous experiment in order to overcome the inconsistency problem. The performance of the classifiers improved as expected but the problem of the inconsistency still exist. The last experiment created as a combination of the 2 previous ways to solve to problem, it stands on merging

the dataset partitioning technique and splitting a test set from each partition, under the coverage of the balancing technique to make the training sets balanced after splitting the test sets. The new technique came as improvement of all the previous experiments techniques, while it solves the problem of the reliability and eliminates the metrics inconsistency problem.

5.1 Experiment 1: Using a balanced dataset

In this experiment, the dataset has been split in 9 equal subsets, ignoring 7 records as outliers in order to have an exact number of samples in order to avoid the problem of balancing. Thus, each subset contains 310 patterns labeled as healthy companies class (i.e. bankruptcy = 'NO'). On the other hand, and in order to obtain a more balanced amount of patterns labeled with class bankruptcy (YES), the 62 originally available records have been duplicated using an over-sampling technique to obtain 124 (30% of samples in every partial dataset). This technique aims to improve the performance of the classifiers. After the subsets created each classifier applied on each subset with 10-fold cross validation. For each classifier the average of the accuracy, sensitivity, specificity and false positive rate calculated. Table 2 shows the average values and the standard deviation for each metric.

Table 2. Experiment 1 results

Classifier	Accuracy	Accu. SD	Sensitivity	Se. SD	Specificity	FPR	Sp. & FPR SD
J48	91.6538%	± 0.0113	0.9058	± 0.0211	0.9207	0.0791	± 0.01411
Random Forest	96.7741%	± 0.00669	0.9766	± 0.0110	0.9641	0.0358	± 0.0068
Naïve Bayes	58.8069%	± 0.0825	0.9193	± 0.0237	0.4555	0.5443	± 0.1230

Random forest classifier gave the best results in this experiment, with the maximum values of sensitivity and specificity, and the minimum value of FPR. Thus, it obtains a high performance in predicting the healthy states of the companies while their in fact healthy. in other words, the rate of missing the healthy companies prediction is very low. In addition, it obtain also the maximum sensitivity comparing with the others classifiers. This means, the it obtains a high performance in case of predicting the bankrupt companies while their in fact bankrupt. Thus, the rate of missing the bankrupt companies is very low. Also, the minimum amount of accuracy, sensitivity and specificity standard deviation makes it the most stable classifier for all the subsets in this experiment; this values represents the oscillation of the metrics values obtained by the classifier for each subset test. J48 ranked as a second one with very good results, it have yielded not much less accuracy, sensitivity and specificity than Random forest, and not much more metrics standard deviation values, makes it also stable for all of the subsets. Also, the metrics values are consistent, that means J48 select the expected behavior to solve the problem; it distribute the effort on all of the classes, not just predict one class most the time. The lowest rank, as expected, assigned to Naïve Bayes classifier with the minimum accuracy and minimum

specificity, and maximum FPR. Thus, it obtains a low performance in predicting the healthy statuses of the companies while they are in fact healthy. In other words, the rate of missing the healthy companies prediction is very high. The value of sensitivity metric in Naïve Bayes is still high, which means it still predicts failed companies in a high rate on the expense of predicting the healthy companies (not optimum behavior). The values of metrics standard deviation is the biggest with Naïve Bayes, proving that it is obviously unstable.

In this experiment, all the records of bankruptcy class have been oversampled in each subset, in order to make them more balanced. In addition cross validation has been used to test the classifiers. This makes the results of all the classifier not very reliable, as some patterns could be potentially located in training folds and test folds at the same time, thus, ‘artificially’ improving the accuracy and other metrics.

5.2 Experiment 2: Training and test sets

In this experiment, a subset of the original dataset has been used for testing, while the remaining data has been used to train the classifiers. In this case, data balancing techniques have not been used. The test set contains a specific percentage of the original dataset: 20% of random companies’ records labeled as healthy (bankruptcy = ‘NO’), i.e. 559 samples. and 20% of random companies’ records labeled with bankruptcy class, i.e. 12 records samples.

Proposed classifiers have been used to classify the test set after being trained using a dataset containing 2238 records labeled with healthy class and 50 records labeled with bankruptcy class (training data).

Table 3. Experiment 2 results

Classifier	Accuracy	Sensitivity	Specificity	FPR
J48	94.2207%	0.1666	0.9588	0.0411
Random Forest	97.7233 %	0.1666	0.9946	0.0053
Naïve Bayes	19.6147 %	1	0.1788	0.8211

As shown in table 3 the accuracy of Random forest and J48 are reasonable, values of specificity and FPR metrics are considerable. As mentioned before, the rate of missing the healthy companies is low. The value of sensitivity contradicts the results of accuracy, the reason is because the training dataset and the test set are again extremely unbalanced. Therefore, in this experiment also J48 and Random forest unfortunately selects the easiest behavior, which is almost all of the time predicting healthy companies. Even if the accuracy is quite high the behavior of both classifiers make both of them inappropriate to solve the problem in this experiment circumstances. Naïve Bayes has a different situation; it predicts the bankruptcy case more than the reasonable, the rate of missing the bankrupt companies is 0 in the expense of predicting the healthy companies. This gave a poor accuracy for the same reason made the other classifiers confused: the unbalanced dataset.

5.3 Experiment 3: Training and test sets (applying oversampling)

As in previous experiment, a test set containing the 20% of the samples have been created and the remaining 80% have been used for training. However, in this experiment, a simple data balancing technique has been used on the training set. Thus, the records labeled with bankruptcy class have been replicated several times up to reach a 30% of records from bankrupt companies. Therefore, the training dataset contains totally 3197 records, 2238 records for healthy companies and 959 records for the failed ones.

J48, Random forest and Naïve Bayes classifiers have been applied on the test set, after being trained using the training dataset. Table 4 shows the obtained results for each classifier in this experiment.

Table 4. Experiment 3 results

Classifier	Accuracy	Sensitivity	Specificity	FPR
J48	92.2942 %	0.3333	0.9355	0.0644
Random Forest	96.8476 %	0.2500	0.9838	0.0161
Naïve Bayes	18.0385 %	1	0.1627	0.8372

The results of this experiment became worse than the previous one a little in the case of accuracy, specificity and FPR. On the other hand, sensitivity in this experiment is better than in the previous one, J48 classifier yield a value higher than it with Random forest, but the greatest one is given by Naïve Bayes classifier due to the same problem appears in all previous experiments; predicting the bankruptcy state more than reasonable. The improvement point in this experiment over the previous one is the distribution of the effort in each classifier, while each one improved the prediction of the bankrupt companies in the expense of the healthy companies. In general, the classifiers made more effort to predict the 2 cases of the financial status as expected, this makes the overall results relatively better than the previous experiment.

In addition, the main issue of previous experiments has been repeated in this one; the unbalanced dataset with huge amount of replications, which made the classifier confused.

As an outline, also in this experiment, all of the classifiers aren't completely appropriate to solve the problem.

5.4 Experiment 4: Several training and test subsets (using oversampling)

In the last experiment, the dataset has been splitted into nine equal subsets. Each of them contains 310 patterns labeled as healthy companies, and 62 labeled as bankrupt. After this step, every subset has been divided to create a test set containing 20% of each class; 12 records labeled as bankrupt and 62 records labeled as healthy. Thus, every training subset contains 248 records for succeed companies and 50 records for failed ones. Then, the oversampling method was

applied to all the training subsets to make them more balanced; 70% of its records for the healthy companies (248 records) and 30% for the failed ones (106 records). The three classifiers have been then used, and the obtained results are shown in Table 5.

Table 5. Experiment 4 results

Classifier	Accuracy	Accu. SD	Sensitivity	Se. SD	Specificity	FPR	Sp. & FPR SD
J48	84.9849 %	± 0.04479	0.5462	± 0.2119	0.9085	0.0913	± 0.0252
Random Forest	91.2912 %	± 0.0247	0.6018	± 0.2143	0.9659	0.0340	± 0.0221
Naïve Bayes	57.8078 %	± 0.1409	0.8703	± 0.1761	0.5214	0.4784	± 0.1994

In this experiment there are no mutual records in each training subset and its test set, also all the training subsets are balanced (70-30%). This gave strength to the obtained results and made the procedure more robust and reliable. In this case, all the classifiers gave considerable results with relatively low amount of standard deviation, being them equiponderant with regard to the consistency of all the metrics values. Comparing the classifiers results in this experiment, Random forest classifier results were the best with the maximum values of accuracy and specificity, and the lowest FPR value, and the lowest amounts of metrics standard deviation, which makes it the most appropriate classifier to solve the problem with the lowest missing the healthy and bankrupt companies rate, and the lowest oscillatory metrics value which makes it the most stable classifier. J48 ranked second with comparable results regarding to all the metrics, also it is convenient to solve the problem but the preference is for Random forest.

The lowest ranked classifier is Naïve Bayes with the minimum accuracy and specificity, and maximum FPR and standard deviation amount for each metric. As expected the values of sensitivity for Naïve Bayes was the biggest comparing with the others, but unfortunately, its values are lopsided; due to the same reason mentioned in all of the previous experiments, which means it is also inappropriate and unstable in this experiment also.

As shown in all of the previous experiments, Random forest achieved the most considerable results, makes it the ideal and the most stable classifier to do this job. This does not mean that J48 is not convenient or stable also. On the other hand, Naïve Bayes obtains a poor performance, which makes inappropriate classifier to solve the problem addressed in this study. Using the data balancing methods not always obtain good results regarding the reliability. In fact, it is perfect to improve the performance of the classifiers, but in the case of using cross validation the replication of some records in the dataset represents a major problem and makes the results imprecise.

6 Conclusions and future work

In this work, we compare three classification algorithms in order to predict the financial status (bankruptcy) of several companies in Spain. Weka software tool

was used to apply J48, Random forest and Naïve Bayes classifiers. In order to improve the efficiency of each classifier, data balancing techniques, namely undersampling, oversampling, and a mixture of them, have been used. Several experiments have been conducted using different datasets from the original dataset (extremely unbalanced) and considered in a certain sequence in order to figure out the convenient procedure to deal with the dataset to obtain the ideal results. It started from partitioning the dataset into several subsets and using the balancing techniques in order to solve the problem of the balancing, this technique presents unreliable outcomes while the cross validation is used also. Then, the procedure of solving the problem changed to splitting the original dataset in training and test set in order to avoid the reliability problem on the expense of the metrics values consistency; the classifiers select an inappropriate behavior to solve the problem while the training set is extremely unbalanced. Thus, the last experiment came with a perfect solution in order to avoid the disadvantages of the previous procedures, it stands on integrating both of procedure; partitioning the dataset to several equally subsets and split test set for each partition with using the balancing techniques to make the training dataset balanced. This integration yielded the best results regarding the reliability and the consistency of the metrics values and prove itself as the most proper style to solve the problem.

Taking into account the obtained results, Random Forest and J48 obtains considerable outcomes regarding the accuracy of the prediction, sensitivity (recall), specificity, and false positive rate metrics. The outcomes provided by Naïve Bayes are not as a required regarding to all of the metrics values. Random forest outcomes regarding to the financial status prediction are the best. This lead us to propose it as the most appropriate solution to predict the financial status of the companies.

As future work, it would be interesting to use other classification methods to solve the problem, and fine-tuning their parameters, or even combining them in ensembles. Also, in future studies, it would be convenient to study the development of methods that take into account type I and type II errors, instead of taking into account only the total error obtained.

Acknowledgments

This work has been supported in part by projects TIN2014-56494-C4-3-P and TEC2015-68752 (Spanish Ministry of Economy and Competitiveness and FEDER).

References

1. Hanady Abdulsalam, David B Skillicorn, and Patrick Martin. Streaming random forests. In *Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th International*, pages 225–232. IEEE, 2007.
2. E Alfaro-Cid, AM Mora, JJ Merelo, AI Esparcia-Alcázar, and K Sharman. Finding relevant variables in a financial distress prediction problem using genetic programming and self-organizing maps. In *Natural Computing in Computational Finance*, pages 31–49. Springer, 2009.

3. Indranil Bose and Raktim Pal. Predicting the survival or failure of click-and-mortar corporations: A knowledge discovery approach. *European Journal of Operational Research*, 174(2):959–982, 2006.
4. Jianguo Chen, Kenli Li, Zhuo Tang, Kashif Bilal, Shui Yu, Chuliang Weng, and Ke-qin Li. A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Transactions on Parallel and Distributed Systems*, 28(4):919–933, 2017.
5. MAH Farquard and Indranil Bose. Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, 53(1):226–233, 2012.
6. Alberto Fernández, Victoria López, Mikel Galar, María José Del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems*, 42:97–110, 2013.
7. Glenn Fung, Olvi L Mangasarian, and Jude W Shavlik. Knowledge-based support vector machine classifiers. In *NIPS*, pages 521–528, 2002.
8. Xiao-Feng Hui and Jie Sun. An application of support vector machine to companies financial distress prediction. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 274–282. Springer, 2006.
9. Hongkyu Jo and Ingoo Han. Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. *Expert Systems with applications*, 11(4):415–422, 1996.
10. Thales Sehn Korting. C4. 5 algorithm and multivariate decision trees. *Image Processing Division, National Institute for Space Research-INPE Sao Jose dos Campos-SP, Brazil*, 2006.
11. Hui Li and Jie Sun. Predicting business failure using support vector machines with straightforward wrapper: A re-sampling study. *Expert Systems with Applications*, 38(10):12747–12756, 2011.
12. Fengyi Lin, Ching-Chiang Yeh, and Meng-Yuan Lee. The use of hybrid manifold learning and support vector machines in the prediction of business failure. *Knowledge-Based Systems*, 24(1):95–101, 2011.
13. Sung-Hwan Min, Jumin Lee, and Ingoo Han. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert systems with applications*, 31(3):652–660, 2006.
14. Antonio Miguel Mora, Luis Javier Herrera, J Urquiza, Ignacio Rojas, and JJ Merelo. Applying support vector machines and mutual information to book losses prediction. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–7. IEEE, 2010.
15. Antonio M Mora García, Pedro A Castillo Valdivieso, Juan J Merelo Guervós, Eva Alfaro Cid, Anna I Esparcia-Alcázar, and Ken Sharman. Discovering causes of financial distress by combining evolutionary algorithms and artificial neural networks. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 1243–1250. ACM, 2008.
16. Tina R Patil and SS Sherekar. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2):256–261, 2013.
17. Vadlamani Ravi, H Kurniawan, Peter Nwee Kok Thai, and P Ravi Kumar. Soft computing system for bank performance prediction. *Applied soft computing*, 8(1):305–315, 2008.
18. Antanas Verikas, Zivile Kalsyte, Marija Bacauskiene, and Adas Gelzinis. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft Computing*, 14(9):995–1010, 2010.

19. Chih-Hung Wu, Gwo-Hshiung Tzeng, Yeong-Jia Goo, and Wen-Chang Fang. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert systems with applications*, 32(2):397–408, 2007.
20. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.
21. Ching-Chiang Yeh, Der-Jang Chi, and Ming-Fu Hsu. A hybrid approach of dea, rough set and support vector machines for business failure prediction. *Expert Systems with Applications*, 37(2):1535–1541, 2010.

Change Point Detection in Autoregression Without Variability Estimation

Barbora Peřtová^{1*} and Michal Peřta^{2**}

¹ The Czech Academy of Sciences, Institute of Computer Science, Czech Republic
`pestova@cs.cas.cz`

² Charles University, Faculty of Mathematics and Physics, Czech Republic
`michal.pesta@mff.cuni.cz`

Abstract. A sequence of time-ordered observations follows an autoregressive model of order one and its parameter is possibly subject to change at most once at some unknown time point. The aim is to test whether such an unknown change has occurred or not. A change point method presented here rely on a ratio type test statistic based on the maxima of cumulative sums. The main advantage of the proposed approach is that the variance of the observations neither has to be known nor estimated. Asymptotic distribution of the test statistic under the no change null hypothesis is derived. Moreover, we prove the consistency of the test under the alternative. The results are illustrated through a simulation study, which demonstrates computational efficiency of the procedure. A practical application to real data is presented as well.

Keywords: Change point, structural change, change in autoregression, hypothesis testing, ratio type statistic, variance estimation free test

1 Introduction and main goals

The focus lies on *autoregressive time series of order one*, i.e., AR(1) series. We try to detect a possible change of the scalar parameter from a stationary autoregressive model using the ratio type test statistic, which allows us to avoid estimating the unknown nuisance dispersion parameter of the time series.

The results are inspired by [6], where an autoregressive times series model of order p is taken into account and the whole vector of autoregression parameters is subject to change. The authors proposed to detect such change by computing partial sums of weighted residuals based on the maximum type CUSUM test statistics. The results were consequently extended by the bootstrap approach in [5]. The main disadvantage of these methods is that the *variance estimation is problematic*. To overcome such a dilemma, the ratio type test statistic is utilized in the change point detection.

* With institutional support RVO:67985807.

** Supported by the Czech Science Foundation project No. P402/12/G097.

The paper is structured as follows: Section 2 introduces a change point model for AR(1) series together with stochastic assumptions. The ratio type test statistic for the change point detection is proposed in Section 3. Consequently, the asymptotic behavior of the considered test statistic is derived, which covers the main theoretical contribution. Asymptotic critical values are calculated in Section 4 by Monte Carlo simulations. Section 5 contains a simulation study that illustrates performance of the asymptotic test. It numerically emphasizes the advantages and disadvantages of the proposed procedure. A practical application of the developed approach to a stock exchange index is presented in Section 6. Proofs are given in the Appendix.

2 Autoregressive model with possibly changed parameter

Let us consider the following time series model with a possible change in parameter β after an unknown time point τ :

$$Y_t = \beta Y_{t-1} + \delta Y_{t-1} \mathcal{I}\{t > \tau\} + \varepsilon_t, \quad t = 2, \dots, n, \quad (1)$$

where β and $\delta \neq 0$ are fixed (not depending on n) unknown parameters, $1 < \tau = \tau_n \leq n$ is the unknown change point, and $\varepsilon_2, \dots, \varepsilon_n$ are independent and identically distributed (iid) random errors satisfying further conditions specified later on. For the sake of convenience, we suppress the index n in the observations $Y_{t,n}$ as well as in the parameter τ_n whenever possible.

We are going to test the *null hypothesis* that the autoregression parameter remained constant for the whole observation period

$$H_0 : \tau = n \quad (2)$$

against the *alternative* that a change of the autoregression parameter occurred at some unknown time point τ prior to the latest observed time n , i.e.,

$$H_1 : \tau < n, \delta \neq 0. \quad (3)$$

3 Test statistic for change in autoregression

The ratio type test statistics for the simple change in mean were introduced in [2]. We utilize this idea and propose the following *ratio type test statistic* to detect the change in the autoregression of order one

$$\mathcal{V}_n = \max_{n\gamma \leq k \leq n-n\gamma} \frac{\max_{2 \leq i \leq k} \left| \sum_{j=1}^{i-1} Y_j (Y_{j+1} - \hat{\beta}_{1k} Y_j) \right|}{\max_{k+1 \leq i \leq n-1} \left| \sum_{j=i}^{n-1} Y_j (Y_{j+1} - \hat{\beta}_{2k} Y_j) \right|}, \quad (4)$$

where $0 < \gamma < 1/2$ is a given constant, $\hat{\beta}_{1k}$ is an ordinary least squares estimate of the parameter β based on the observations Y_1, \dots, Y_k and $\hat{\beta}_{2k}$ is an ordinary least

squares estimate of β based on the observations Y_{k+1}, \dots, Y_n . Being more formal, the estimate $\hat{\beta}_{1k}$ is obtained when regressing the vector of responses $\mathbf{y}_{1,k} := (Y_2, \dots, Y_k)^\top$ on the vector of covariates $\mathbf{x}_{1,k} := (Y_1, \dots, Y_{k-1})^\top$. Analogously, the estimate $\hat{\beta}_{2k}$ is obtained when regressing the vector of responses $\mathbf{y}_{k+1,n} := (Y_{k+2}, \dots, Y_n)^\top$ on the vector of regressors $\mathbf{x}_{k+1,n} := (Y_{k+1}, \dots, Y_{n-1})^\top$.

The motivation for constructing the ratio type test statistic \mathcal{V}_n comes from the linear regression setup (so-called normal equations). The estimate $\hat{\beta}_{1k}$ is a solution of

$$\mathbf{x}_{1,k}^\top (\mathbf{y}_{1,k} - \mathbf{x}_{1,k} b) = 0$$

with respect to $b \in \mathbb{R}$ and the estimate $\hat{\beta}_{2k}$ is a solution of

$$\mathbf{x}_{k+1,n}^\top (\mathbf{y}_{k+1,n} - \mathbf{x}_{k+1,n} b) = 0$$

with respect to $b \in \mathbb{R}$. Then, one may define partial sums of the weighted residuals as

$$\mathbf{x}_{1,i}^\top (\mathbf{y}_{1,i} - \mathbf{x}_{1,i} \hat{\beta}_{1k}), \quad i = 2, \dots, k$$

and

$$\mathbf{x}_{i,n}^\top (\mathbf{y}_{i,n} - \mathbf{x}_{i,n} \hat{\beta}_{2k}), \quad i = k+1, \dots, n.$$

Consequently, these partial sums can be used as basis for the maxima of partial sums in the numerator and the denominator of \mathcal{V}_n . Note that this approach—usage of the ratio type test statistics—can be generalized for the change of a vector autoregression parameter of the stationary autoregressive AR(p)-process, when $p \geq 2$, using the notation from [6].

Before deriving asymptotic properties of the ratio type test statistic, we formulate several stochastic assumptions on the time series model (1):

Assumption A1 $\beta \in (-1, 1) \setminus \{0\}$.

Assumption A2 $\beta + \delta \in (-1, 1) \setminus \{0\}$.

Assumption A3 $\{\varepsilon_i, i = 0, \pm 1, \dots\}$ are iid random variables having $E\varepsilon_i = 0$, $\text{Var } \varepsilon_i = \sigma^2 > 0$, and $E\varepsilon_i^4 < \infty$ for all i . Observation Y_1 is independent of $\{\varepsilon_2, \varepsilon_3, \dots\}$.

Assumptions A1, A2, and A3 mean that the time series is a stationary autoregressive sequence of order one (and not an iid sequence) before and even after the possible change point.

The limit behavior of the test statistic under the null hypothesis is characterized by the following theorem.

Theorem 1 (Under null). Suppose that Y_1, \dots, Y_n follow model (1), assume that Assumptions A1 and A3 hold. Then, under null hypothesis (2)

$$\mathcal{V}_n \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \sup_{\gamma \leq t \leq 1-\gamma} \frac{\sup_{0 \leq u \leq t} |\mathcal{W}(u) - u/t \mathcal{W}(t)|}{\sup_{t \leq u \leq 1} |\widetilde{\mathcal{W}}(u) - (1-u)/(1-t) \widetilde{\mathcal{W}}(t)|}, \quad (5)$$

where $\{\mathcal{W}(x), x \in [0, 1]\}$ is a standard Wiener process and $\widetilde{\mathcal{W}}(x) = \mathcal{W}(1) - \mathcal{W}(x)$.

The next theorem describes the test statistic's behavior under a fixed alternative.

Theorem 2 (Under alternative). *Suppose that Y_1, \dots, Y_n follow model (1), assume that alternative (3) holds for some fixed $\delta \neq 0$, and $\tau = \lfloor \zeta n \rfloor$ for some $\gamma < \zeta < 1 - \gamma$. Then, under Assumptions A1, A2, and A3*

$$\mathcal{V}_n \xrightarrow[n \rightarrow \infty]{P} \infty.$$

The previous theorem provides *consistency* of the studied test statistic under the given assumptions. The null hypothesis is rejected for large values of the ratio type statistic. Being more formal, we reject H_0 at significance level α if $\mathcal{V}_n > v_{1-\alpha, \gamma}$, where $v_{1-\alpha, \gamma}$ is the $(1 - \alpha)$ -quantile of the asymptotic distribution (5).

4 Asymptotic critical values

The explicit form of the limit distribution (5) is not known. The critical values may be determined by simulations from the limit distribution from Theorem 1. Theorem 2 ensures that we reject the null hypothesis for large values of the test statistic. We tried to simulate the asymptotic distribution (5) by *discretizing* the Wiener process and using the relationship of a random walk to the Wiener process. We considered 1000 as the number of discretization points within $[0, 1]$ interval and the number of simulation runs equals to 100000. In Table 1, we present several critical values for $\gamma = .1$ and $\gamma = .2$.

$100(1 - \alpha)\%$	90%	95%	97.5%	99%
$\gamma = .1$	6.298815	7.293031	8.283429	9.589896
$\gamma = .2$	4.117010	4.745884	5.368286	6.159252

Table 1. Simulated critical values corresponding to the asymptotic distribution of the test statistic \mathcal{V}_n under the null hypothesis.

Note that the numerator and denominator in the test statistic \mathcal{V}_n can be interchanged and such a modified test statistic can still be used for detection of the change in autoregression (but using different critical values).

A possible extension of the proposed methods, which will be part of the future research, is *bootstrapping*. Using the bootstrap techniques implemented similarly as in [3] for the change in means, one can obtain critical values in an alternative way compared to the presented asymptotic approach.

5 Simulation study

A simulation experiment was performed to study the *finite sample properties* of the asymptotic test for the change in the AR(1) parameter. In particular, the

interest lies in the *empirical size* of the proposed test under the null hypothesis and in the empirical *rejection rate* (power) under the alternative. Random samples (1000 each time) are generated from the time series change point model (1). The number of observations is set to $n = 200$, $n = 400$, and $n = 800$ in order to demonstrate the performance of the testing approach in case of different sample sizes. Two values of the autoregression parameter are taken into consideration, i.e., $\beta = -.6$ and $\beta = .2$ to represent stronger negative dependence and weaker positive dependence. The innovations are obtained as iid random variables from a standard normal $N(0, 1)$ or Student t_5 distribution. Simulation scenarios are produced as all possible combinations of the above mentioned settings. Parameter γ is set to .1.

To assess the theoretical results under H_0 numerically, Table 2 provides the empirical sizes (empirical probabilities of the type I error) of the test for change in the autoregression parameter, where the significance level is α .

α		.100		.050		.025		.010	
innovations		N(0, 1)	t_5	N(0, 1)	t_5	N(0, 1)	t_5	N(0, 1)	t_5
$n = 200$	$\beta = -.6$.258	.342	.172	.266	.126	.216	.088	.152
	$\beta = .2$.296	.400	.206	.318	.158	.234	.106	.176
$n = 400$	$\beta = -.6$.218	.238	.136	.160	.080	.108	.046	.072
	$\beta = .2$.186	.220	.124	.152	.086	.106	.044	.072
$n = 800$	$\beta = -.6$.157	.193	.098	.122	.059	.081	.022	.048
	$\beta = .2$.135	.187	.078	.115	.054	.073	.023	.049

Table 2. Empirical size of the test for change in autoregression under H_0 using the asymptotic critical values of \mathcal{V}_n with $\gamma = .1$, considering a significance level α . Innovations are iid having Student t_5 and standard normal $N(0, 1)$ distribution.

The proportion of rejecting the null hypothesis is getting closer to the theoretical significance level as the number of time series observations increases. Better performance of the test under the null hypothesis is observed, when the innovations have lighter tails (represented by $N(0, 1)$ distribution). Note that the test statistic \mathcal{V}_n is based on the L_2 regression approach. There is no visible direct effect of the value of the autoregression parameter on the empirical rejection rates based on this particular simulation study. Generally, the empirical sizes are higher than they should be, i.e., the test rejects the null hypothesis more often than one would expect.

The performance of the testing procedure under H_1 in terms of the empirical rejection rates is shown in Table 3, where the change point is set to $\tau = n/2$ or $\tau = n/3$. Parameter δ is chosen as $\delta = .5$.

We may conclude that the power of the test increases as the number of observations increases, which was expected. The test power drops when switching from a change point located in the middle of the time series to a change point closer to the beginning or the end of the time series. Innovations with heavier tails

α			.100		.050		.025		.010	
innovations			$N(0, 1)$	t_5	$N(0, 1)$	t_5	$N(0, 1)$	t_5	$N(0, 1)$	t_5
$n = 200$	$\beta = -.6$	$\tau = n/2$.924	.920	.888	.888	.834	.848	.774	.774
		$\tau = n/3$.930	.896	.894	.866	.838	.830	.772	.766
	$\beta = .2$	$\tau = n/2$.788	.828	.718	.766	.640	.694	.548	.602
		$\tau = n/3$.774	.784	.676	.698	.596	.606	.478	.508
$n = 400$	$\beta = -.6$	$\tau = n/2$.984	.984	.968	.958	.926	.924	.856	.888
		$\tau = n/3$.992	.982	.972	.962	.958	.944	.924	.926
	$\beta = .2$	$\tau = n/2$.948	.938	.906	.904	.864	.852	.792	.798
		$\tau = n/3$.898	.892	.826	.828	.752	.758	.634	.642
$n = 800$	$\beta = -.6$	$\tau = n/2$.999	.999	.996	.996	.992	.995	.980	.981
		$\tau = n/3$.999	.999	.996	.998	.995	.997	.987	.988
	$\beta = .2$	$\tau = n/2$.996	.989	.988	.978	.972	.963	.938	.931
		$\tau = n/3$.981	.980	.960	.959	.926	.929	.860	.866

Table 3. Empirical power of the test for change in autoregression under H_1 using the asymptotic critical values of \mathcal{V}_n with $\gamma = .1$, considering a significance level α and $\delta = .5$. Innovations are iid having Student t_5 and standard normal $N(0, 1)$ distribution.

(i.e., t_5) yield slightly smaller power than innovations with lighter tails. Negative dependence seems to give higher power of the test based on this simulation study.

In contrast to the slightly lower power in case of relatively small sample size and moderate change in the autoregression parameter, one may try to consider larger change in β from $-.8$ to $.8$ in case of $n = 150$. Here, the simulated power reaches .994 (for $\alpha = .05$). Hence, in case of a large change in autoregression, the test achieves high power.

To improve the computational performance of the test for detecting the change in autoregression, longer time series of observations are a general solution. Moreover, a suitable bootstrap extension of the developed procedure could be helpful from a numerical and computational point of view.

6 Application to stock exchange index

As an illustrative example of the proposed technique for detecting of the change in autoregression, we concentrate on the Prague Stock Exchange index called *PX Index* (formerly *PX50*). It is a capitalization-weighted index of major stocks that trade on the Prague Stock Exchange.

The starting exchange day for the Index PX50 was April 5, 1994. We consider a time series consisting of daily PX50 values starting from November 16, 1994 up to September 27, 2001. Only business days were taken into account, providing 1850 observations. The starting date of the observation period was chosen later than the starting day of the exchange, since only weekly (not daily) values of the PX50 records were available at the beginning. Moreover, the market after opening the exchange was not as stable as later on. The last observation date

was chosen in order to avoid effects of the attacks on September 11, 2001. The considered time series can be seen in Figure 1. The PX50 data can also be downloaded from [7].

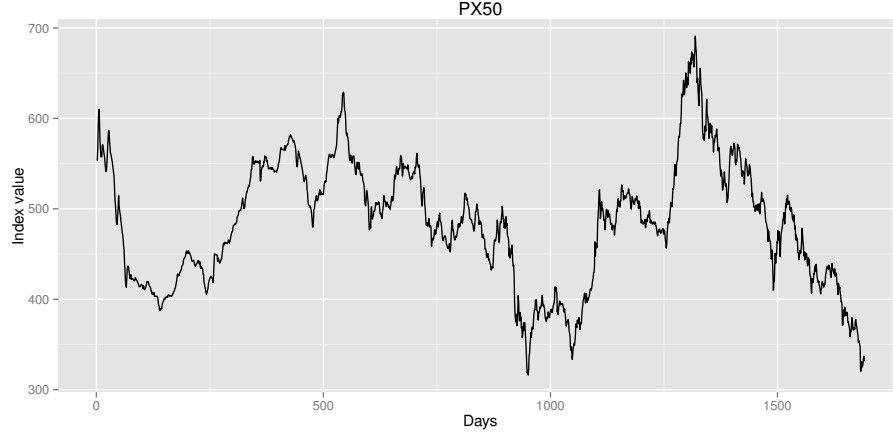


Fig. 1. Daily Prague Stock Exchange index (PX50) values from November 16, 1994 to September 27, 2001.

We denote the original data of the PX50 index as $\{X_t\}_t$. Firstly, we transform the PX50 index by taking into account the differences of logarithms, i.e., $Y_t = \log(X_t/X_{t-1})$. This transformation can be interpreted as considering logarithms of daily returns of the PX50 index. Besides that, using this approach stationary time series before and even after a possible change point are obtained. The transformed index values are shown in Figure 3.

Let us assume that Y_1, \dots, Y_n follow autoregressive change point model (1). We are going to decide whether the change in the AR(1) parameter occurred or not based on the proposed asymptotic test. The value of the test statistic \mathcal{V}_n for $\gamma = .1$ is 7.321143, which is larger than the 95%-critical value 7.293031 simulated from the limit distribution under the null hypothesis. Therefore, we reject the null hypothesis of no change in the autoregressive parameter. The progress of the ratio of the test statistic

$$Q_k = \frac{\max_{2 \leq i \leq k} \left| \sum_{j=1}^{i-1} Y_j (Y_{j+1} - \hat{\beta}_{1k} Y_j) \right|}{\max_{k+1 \leq i \leq n-1} \left| \sum_{j=i}^{n-1} Y_j (Y_{j+1} - \hat{\beta}_{2k} Y_j) \right|}, \quad n\gamma \leq k \leq n - n\gamma$$

is depicted in Figure 2.

We can *estimate* the unknown change point τ in a similar fashion as in [4]:

$$\hat{\tau} = \arg \max_{2 \leq k \leq n} \left| \sum_{j=1}^{k-1} Y_j (Y_{j+1} - \hat{\beta}_{1n} Y_j) \right|.$$

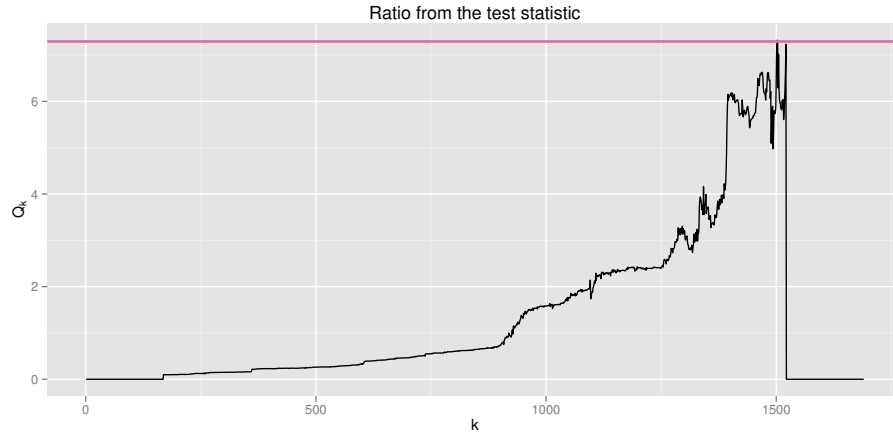


Fig. 2. The values of Q_k for the PX50 index data with $\gamma = .1$. The colored horizontal line represents the 95%-critical value.

This leads to $\hat{\tau} = 949$, which corresponds to October 7, 1998. The log returns of PX50 together with the depicted change point for the change in autoregression are displayed in Figure 3.

The explanation of the detected change in autoregression is possibly connected to the Russian financial crisis (also called Ruble crisis) that hit Russia on August 17, 1998. It resulted in the Russian government and the Russian Central Bank devaluing the ruble and defaulting on its debt. In 1998 influenced by Russian financial crisis, the index reached its historical bottom on October 8 with 316 points, which is the first day after the detected change in autoregression of the PX50 log returns.

Finally, we investigated eligibility of the model. The ACF (autocorrelation function) and PACF (partial autocorrelation function) plots of the time series before and after the estimated change point are employed. Both ACF plots go to zero at an exponential rate, while both PACF plots become zero immediately after the first lag. We applied the Ljung-Box test on the residuals of the fitted AR(1) models (before and after the change). The hypothesis that the residuals in each AR(1) model have no autocorrelation is rejected in both cases, which suggests that the two series are stationary.

7 Conclusions

A testing procedure for a possible change in the autoregression parameter is demonstrated. It detects whether the observed sequence is an AR(1) process, or the time series is an AR(1) process up to some unknown time point and it is again an AR(1) process after this unknown time point with a different autoregression parameter.

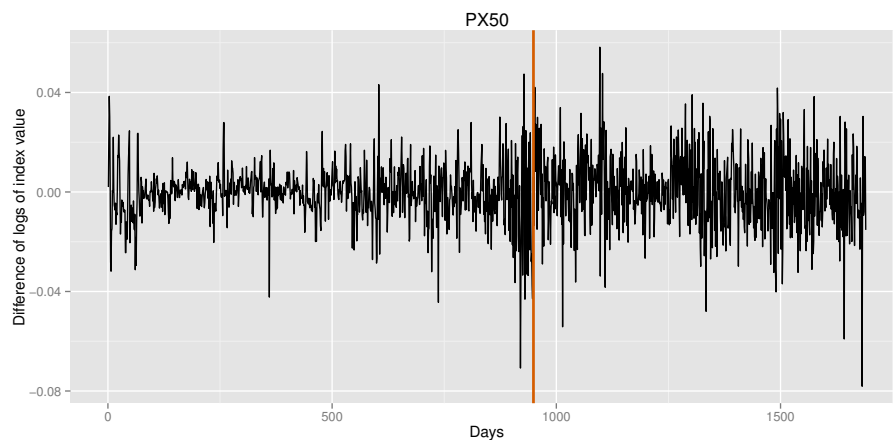


Fig. 3. Log returns of PX50.

The asymptotic behavior of the ratio type test statistic for the change in autoregression was investigated under the null hypothesis as well as under the alternative. The theoretical limiting distribution under the null hypothesis provided critical values for the test, which were obtained by simulation. The main advantage of the ratio type statistics in hypotheses testing is that they provide an alternative to the non-ratio type statistics mainly in situations, in which variance estimation is not straightforward. The simulations reveal that the method keeps the significance level under the null and provides reasonable powers under the alternatives. Finally, an application of the developed procedure on the stock exchange index data was performed.

Acknowledgments. Institutional support to Barbora Peřtová was provided by RVO:67985807. Michal Peřta was supported by the Czech Science Foundation project “DYME – Dynamic Models in Economics” No. P402/12/G097.

References

1. Davidson, J.: Stochastic Limit Theory: An Introduction For Econometricians. Oxford University Press, New York (1994)
2. Horváth, L., Horváth, Z., Huřková, M.: Ratio Tests for Change Point Detection. In: Balakrishnan, N., Peña, E.A., Silvapulle, M.J. (eds.) Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen, vol. 1, pp. 293–304. IMS Collections, Beachwood, Ohio (2009)
3. Peřtová, B., Peřta, M.: Testing Structural Changes in Panel Data With Small Fixed Panel Size and Bootstrap. *Metrika* 78(6), 665–689 (2015)
4. Peřtová, B., Peřta, M.: Change Point Estimation in Panel Data Without Boundary Issue. *Risks*, 5(1):7 (2017)

5. Huřková, M., Kirch, C., Prášková, Z., Steinebach, J.: On the Detection of Changes in Autoregressive Time Series, II. Resampling Procedures. *J. Stat. Plan. Infer.* 138(6):1697–1721 (2008)
6. Huřková, M., Prášková Z., Steinebach, J.: On the Detection of Changes in Autoregressive Time Series I. Asymptotics. *J. Stat. Plan. Infer.* 137(4):1243–1259 (2007).
7. Prague Stock Exchange: PX Index 2015. [Online; Available from <https://www.pse.cz/dokument.aspx?k=Burzovni-Indexy>; Updated April 30, 2015; Accessed April 30, 2015].

Appendix: Proofs

Proof (of Theorem 1). Let us consider an array

$$U_{n,i} = \frac{\sqrt{1-\beta^2}}{\sigma^2\sqrt{n-1}} Y_{i-1}\varepsilon_i, \quad i = 2, \dots, n$$

and a filtration $\mathcal{F}_{n,i} = \sigma\{\varepsilon_j, j \leq i\}$, $i = 2, \dots, n$ and $n \in \mathbb{N}$. Then, $\{U_{n,i}, \mathcal{F}_{n,i}\}$ is a martingale difference array such that

$$\mathbb{E}U_{n,i}^2 = \frac{1-\beta^2}{\sigma^4(n-1)} \mathbb{E}Y_{i-1}^2\varepsilon_i^2 = \frac{1}{n-1}.$$

Moreover,

$$\sum_{i=2}^n U_{n,i}^2 - \sum_{i=2}^n \mathbb{E}U_{n,i}^2 = \frac{1-\beta^2}{\sigma^4(n-1)} \sum_{i=2}^n (Y_{i-1}^2\varepsilon_i^2 - \mathbb{E}Y_{i-1}^2\varepsilon_i^2).$$

Furthermore,

$$\begin{aligned} & \frac{1}{n-1} \sum_{i=2}^n (Y_{i-1}^2\varepsilon_i^2 - \mathbb{E}Y_{i-1}^2\varepsilon_i^2) \\ &= \frac{1}{n-1} \sum_{i=2}^n [Y_{i-1}^2(\varepsilon_i^2 - \sigma^2)] + \frac{1}{n-1} \sum_{i=2}^n (Y_{i-1}^2 - \mathbb{E}Y_{i-1}^2)\sigma^2. \end{aligned}$$

Since $\{Y_{i-1}^2(\varepsilon_i^2 - \sigma^2)\}$ is a martingale difference array again with respect to $\mathcal{F}_{n,i}$, we have under Assumption A3 from the Chebyshev's inequality that

$$\frac{1}{n-1} \sum_{i=2}^n [Y_{i-1}^2(\varepsilon_i^2 - \sigma^2)] \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0.$$

Similarly, as a consequence of Lemma 4.2 in [6],

$$\frac{1}{n-1} \sum_{i=2}^n (Y_{i-1}^2 - \mathbb{E}Y_{i-1}^2) \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 0.$$

Thus,

$$\sum_{i=2}^n U_{n,i}^2 \xrightarrow[n \rightarrow \infty]{\mathbb{P}} 1. \tag{6}$$

Next, for any $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(\max_{2 \leq i \leq n} U_{n,i}^2 > \epsilon\right) &\leq \sum_{i=2}^n \mathbb{P}\left(\frac{1-\beta^2}{\sigma^4(n-1)} Y_{i-1}^2 \varepsilon_i^2 > \epsilon\right) \\ &\leq \frac{(1-\beta^2)^2}{\epsilon^2 \sigma^8 (n-1)^2} \sum_{i=2}^n \mathbb{E} Y_{i-1}^4 \mathbb{E} \varepsilon_i^4 \xrightarrow{n \rightarrow \infty} 0. \end{aligned} \quad (7)$$

Additionally,

$$\lim_{n \rightarrow \infty} \sum_{i=2}^{[nt]} \mathbb{E} U_{n,i}^2 = \lim_{n \rightarrow \infty} \frac{[nt] - 1}{n - 1} = t \quad (8)$$

for all $t \in [0, 1]$.

According to Theorem 27.14 from [1] for the martingale difference array $\{U_{n,i}, \mathcal{F}_{n,i}\}$, where the assumptions of this theorem are satisfied due to (6), (7), and (8), we get

$$\sum_{i=2}^{[nt]} U_{n,i} \xrightarrow[n \rightarrow \infty]{\mathcal{D}[0,1]} \mathcal{W}(t).$$

Therefore,

$$\frac{1}{\sqrt{n-1}} \left(\sum_{i=2}^{[nt]} Y_{i-1} \varepsilon_i, \sum_{i=[nt]+2}^n Y_{i-1} \varepsilon_i \right) \xrightarrow[n \rightarrow \infty]{\mathcal{D}[0,1]} \frac{\sigma^2}{\sqrt{1-\beta^2}} \left(\mathcal{W}(t), \widetilde{\mathcal{W}}(t) \right), \quad (9)$$

where $\widetilde{\mathcal{W}}(t) = \mathcal{W}(1) - \mathcal{W}(t)$.

Let us define $\mathbf{Y}_{j,l} = (Y_j, \dots, Y_l)^\top$ and $\boldsymbol{\varepsilon}_{j,l} = (\varepsilon_j, \dots, \varepsilon_l)^\top$. Hence, for the expression in the numerator of \mathcal{V}_n it holds

$$\begin{aligned} \sum_{j=1}^{i-1} Y_j (Y_{j+1} - \hat{\beta}_{1k} Y_j) &= \mathbf{Y}_{1,i-1}^\top \left(\mathbf{Y}_{2,i} - \mathbf{Y}_{1,i-1} \hat{\beta}_{1k} \right) \\ &= \mathbf{Y}_{1,i-1}^\top \left(\mathbf{Y}_{1,i-1} \beta + \boldsymbol{\varepsilon}_{2,i} - \mathbf{Y}_{1,i-1} \beta - \mathbf{Y}_{1,i-1} (\mathbf{Y}_{1,k-1}^\top \mathbf{Y}_{1,k-1})^{-1} \mathbf{Y}_{1,k-1}^\top \boldsymbol{\varepsilon}_{2,k} \right) \\ &= \mathbf{Y}_{1,i-1}^\top \boldsymbol{\varepsilon}_{2,i} - \mathbf{Y}_{1,i-1}^\top \mathbf{Y}_{1,i-1} (\mathbf{Y}_{1,k-1}^\top \mathbf{Y}_{1,k-1})^{-1} \mathbf{Y}_{1,k-1}^\top \boldsymbol{\varepsilon}_{2,k}. \end{aligned} \quad (10)$$

Similarly for the expression in the denominator of \mathcal{V}_n

$$\begin{aligned} \sum_{j=i}^{n-1} Y_j (Y_{j+1} - \hat{\beta}_{2k} Y_j) \\ = \mathbf{Y}_{i,n-1}^\top \boldsymbol{\varepsilon}_{i+1,n} - \mathbf{Y}_{i,n-1}^\top \mathbf{Y}_{i,n-1} (\mathbf{Y}_{k+1,n-1}^\top \mathbf{Y}_{k+1,n-1})^{-1} \mathbf{Y}_{k+1,n-1}^\top \boldsymbol{\varepsilon}_{k+2,n}. \end{aligned} \quad (11)$$

Lemma 4.2 in [6] gives

$$\sup_{\gamma \leq t < 1} \frac{1}{[nt]} \left| \sum_{s=1}^{[nt]} (Y_s^2 - \mathbb{E} Y_s^2) \right| = o_{\mathbb{P}}(1) \quad (12)$$

and

$$\sup_{0 < t \leq 1-\gamma} \frac{1}{[n(1-t)]} \left| \sum_{s=[nt]+1}^{n-1} (Y_s^2 - \mathbb{E}Y_s^2) \right| = o_{\mathbb{P}}(1). \quad (13)$$

Finally, (9) together with (10), (11), (12), and (13) implies

$$\begin{aligned} & \frac{1}{\sqrt{n-1}} \left(\sup_{0 \leq u \leq t} \left| \sum_{j=1}^{[nu]-1} Y_j(Y_{j+1} - \hat{\beta}_{1[nt]}Y_j) \right| \right) \\ & \sup_{t \leq u \leq 1} \left| \sum_{j=[nu]+1}^{n-1} Y_j(Y_{j+1} - \hat{\beta}_{2[nt]}Y_j) \right| \Bigg) \\ & \xrightarrow[n \rightarrow \infty]{\mathcal{O}^2[\gamma, 1-\gamma]} \frac{\sigma^2}{\sqrt{1-\beta^2}} \left(\sup_{0 \leq u \leq t} |\mathcal{W}(u) - u/t\mathcal{W}(t)| \right. \\ & \left. \sup_{t \leq u \leq 1} |\widetilde{\mathcal{W}}(u) - (1-u)/(1-t)\widetilde{\mathcal{W}}(t)| \right). \end{aligned}$$

Then, the assertion of the theorem directly follows. \square

Proof (of Theorem 2). Let us take $k = \tau + 2$, $k = [\xi n]$ for some $\zeta < \xi < 1 - \gamma$ and $i = \tau + 1$. Then,

$$\begin{aligned} & \sum_{j=1}^{\tau} Y_j(Y_{j+1} - \hat{\beta}_{1(\tau+2)}Y_j) \\ & = \mathbf{Y}_{1,\tau}^{\top} \boldsymbol{\varepsilon}_{2,\tau+1} - \mathbf{Y}_{1,\tau}^{\top} \mathbf{Y}_{1,\tau} (\mathbf{Y}_{1,\tau+1}^{\top} \mathbf{Y}_{1,\tau+1})^{-1} \mathbf{Y}_{1,\tau+1}^{\top} \boldsymbol{\varepsilon}_{2,\tau+2} - \mathbf{Y}_{1,\tau}^{\top} \mathbf{Y}_{1,\tau} \delta. \end{aligned}$$

According to the proof of Theorem 1, as $n \rightarrow \infty$,

$$\frac{1}{\sqrt{n-1}} \left(\mathbf{Y}_{1,\tau}^{\top} \boldsymbol{\varepsilon}_{2,\tau+1} - \mathbf{Y}_{1,\tau}^{\top} \mathbf{Y}_{1,\tau} (\mathbf{Y}_{1,\tau+1}^{\top} \mathbf{Y}_{1,\tau+1})^{-1} \mathbf{Y}_{1,\tau+1}^{\top} \boldsymbol{\varepsilon}_{2,\tau+2} \right) = \mathcal{O}_{\mathbb{P}}(1).$$

Lemma 4.2 from [6] gives

$$\frac{1}{\sqrt{n-1}} |\mathbf{Y}_{1,\tau}^{\top} \mathbf{Y}_{1,\tau} \delta| \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \infty.$$

Now,

$$\frac{1}{\sqrt{n-1}} \max_{2 \leq i \leq k} \left| \sum_{j=1}^{i-1} Y_j(Y_{j+1} - \hat{\beta}_{1k}Y_j) \right| \xrightarrow[n \rightarrow \infty]{\mathbb{P}} \infty.$$

For $\tau < k = [\xi n]$, the denominator in (4) divided by $\sqrt{n-1}$ has the same distribution as under the null hypothesis and it is, therefore, bounded in probability. It follows that the maximum of the ratio has to tend in probability to infinity as well, while $n \rightarrow \infty$. \square

Distance Between VAR Models and its Application to Spatial Differences Analysis in the Relationship GDP - Unemployment Growth Rate in Europe

Francesca Di Iorio^a and Umberto Triacca^b

^aUniversità di Napoli Federico II,

^bUniversità dell'Aquila

{fdiiorio}@unina.it

{umberto.triacca}@ec.univaq.it

Abstract. In this paper a novel distance measure for evaluating the closeness of two vector autoregressive models is presented and its main properties are discussed. The proposed distance is used to investigate the presence of spatial differences in the dynamic link between unemployment rate variation and GDP growth in some European Union countries.

Keywords: AR metric; Distance; Unemployment; GDP; VAR models

1 Introduction

Vector autoregressive (VAR) models, popularized by [6], are a class of models that are designed to capture joint movements and dynamic patterns in an array of multiple variables. These models have been applied in various research fields. Their success is mainly based upon the fact that they are considered to be data-driven, i.e., the underlying structure in the estimated model is determined by the data. However, there is no canonical way to measure the dissimilarity between two different VARs. The need for such a distance measure arises in both clustering and classification of multivariate time series. In this paper, we propose a distance measure for evaluating the closeness of two vector VAR models and we use such notion of distance to investigate the presence of spatial differences in the dynamic link between unemployment rate variation and economic growth in some European Union economies.

The rest of paper is organized as follows. Section 2 introduces a distance measure between pairs of VAR models. We start by giving a formal definition of distance between VAR models. After that, its main properties are discussed. In section 3, we present the application. Finally, we conclude in section 4.

2 A distance measure between VAR models

In this section, we introduce a new distance measure between VAR models. Let us remind that a k -dimensional process $\mathbf{y} = \{\mathbf{y}_t = (y_{1t}, \dots, y_{kt})'; t \in \mathbb{Z}\}$ is a

VAR(p) process if it can be represented as

$$\mathbf{A}(\mathbf{L})\mathbf{y}_t = \mathbf{u}_t \quad (1)$$

where $\{\mathbf{u}_t = (u_{1t}, \dots, u_{kt})'; t \in \mathbb{Z}\}$ is a k -variate white-noise process with zero mean vector and nonsingular covariance matrix $\Sigma_{\mathbf{u}}$. The $(k \times k)$ matrix $\mathbf{A}(L)$ has finite polynomial elements in lag operator L and is assumed to be of full rank. It can be expressed as

$$\mathbf{A}(L) = \mathbf{I} - \mathbf{A}_1 L - \dots - \mathbf{A}_p L^p$$

where \mathbf{I} is the $(k \times k)$ identity matrix and $\{\mathbf{A}_i\}$ are matrices of parameters.

Process (1) is stationary if the roots of the determinant equation $\det[\mathbf{A}(z)] = 0$ are outside the unit circle.

In this paper, we assume that

$$\det[\mathbf{A}(z)] \neq 0, \quad |z| < 1 \quad \text{for } z \in \mathbb{C} \quad (2)$$

It is important to note that the condition (2) allows for nonstationarity. However, it excludes explosive processes from our consideration. Now, following [8], we show that a VAR process implies a particular specification of its individual elements in terms of univariate ARMA processes.

We first observe that

$$\text{adj}[\mathbf{A}(L)]\mathbf{A}(L) = \det[\mathbf{A}(L)]\mathbf{I}$$

where $\text{adj}[\mathbf{A}(L)]$ is the adjoint of the matrix $\mathbf{A}(L)$. Then, premultiplying both sides of (1) by $\text{adj}[\mathbf{A}(L)]$, we obtain the autoregressive final form,

$$\det[\mathbf{A}(L)]\mathbf{y}_t = \text{adj}[\mathbf{A}(L)]\mathbf{u}_t.$$

Consequently, the marginal model for the i th element of \mathbf{y}_t is given by

$$\det[\mathbf{A}(L)]y_{it} = \text{adj}_i[\mathbf{A}(L)]\mathbf{u}_t \quad (3)$$

where $\text{adj}_i[\mathbf{A}(L)]$ denotes the i th row of the matrix $\text{adj}[\mathbf{A}(L)]$. As the right-hand side of (3) is the sum of k finite moving averages, it can also be represented as a finite moving average $\theta_i(L)\epsilon_{it}$, where ϵ_{it} is a white noise process, such that

$$\theta_i(L)\epsilon_{it} = \text{adj}_i[\mathbf{A}(L)]\mathbf{u}_t \quad (4)$$

The coefficients of the polynomial $\theta_i(L)$ are found by equating the autocovariances in the two representations. The invertibility condition ensures a unique solution. Considering (3) and (4), the univariate ARMA models implied by (1) are given by

$$\det[\mathbf{A}(L)]y_{it} = \theta_i(L)\epsilon_{it}.$$

Thus we have

$$y_{it} \sim \text{ARMA}(p^*, q^*) \quad i = 1, \dots, k,$$

where it is well known that $p^* \leq kp$ and $q^* \leq (k-1)p$. We note that the innovations in the different ARMA models are correlated and, in absence of any cancellation, all the univariate models have identical autoregressive parts. Further, we observe that one unit root in the VAR model implies a unit root in each of the univariate models.

Summarizing, we have that a set $A_{\mathbf{y}}$ of k invertible univariate ARMA processes corresponds to every k -variate VAR process, \mathbf{y} . We note that any univariate process $y_i \in A_{\mathbf{y}}$ has an (possibly infinite order) AR representation,

$$y_{it} = \sum_{l=1}^{\infty} \pi_{il} y_{it-l} + \epsilon_{it}.$$

Given two invertible ARMA processes, x, y , following [4], we consider the quantity

$$d(x, y) = \left[\sum_{l=1}^{\infty} (\pi_{xl} - \pi_{yl})^2 \right]^{\frac{1}{2}}.$$

as a measure of distance between the two invertible ARMA processes. Thus in the class of the k -variate VAR processes, V_k , could seem natural to consider the following function as measure of dissimilarity between two VAR processes. Then:

Definition 1. Let \mathbf{x} and \mathbf{y} two VAR models in V_k ; then their distance $L(\mathbf{x}, \mathbf{y})$ is given by the sum of the distances between the implied ARMA models component by component:

$$L(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^k d(x_i, y_i), \quad \mathbf{x}, \mathbf{y} \in V_k \quad (5)$$

where x_i and y_i ($i = 1, \dots, k$) are the univariate invertible ARMA processes implied by k -variate VAR processes \mathbf{x} and \mathbf{y} , respectively. ■

Using the proposed distance (5) we can introduce the notion of *norm* of a VAR process.

Definition 2. Let \mathbf{y} a VAR model in V_k ; their norm is given by: $\|\mathbf{y}\| = L(\mathbf{y}, \mathbf{u})$. ■

Example 1. Consider the following VAR(1)

$$\begin{bmatrix} 1 - 0.5L & 0.66L \\ 0.5L & 1 + 0.3L \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \epsilon_{x_t} \\ \epsilon_{y_t} \end{bmatrix} \quad (6)$$

where $(\epsilon_{x_t}, \epsilon_{y_t})'$ is a bivariate white noise with covariance matrix

$$E \left(\begin{bmatrix} \epsilon_{x_t} \\ \epsilon_{y_t} \end{bmatrix} \begin{bmatrix} \epsilon_{x_t} & \epsilon_{y_t} \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This example is presented in [2](p.438). The norm of this process is equal to 1.05. ■

The norm can also be seen as a "measure" of stochastic dependence structure in a VAR process. To describe this point we recall that the innovation vector

$$\mathbf{u} = \{\mathbf{y}_t = (u_{1t}, \dots, u_{kt})'; t \in \mathbb{Z}\}$$

is a k -variate VAR(0) process; thus the norm of the VAR(p) process $\mathbf{y} \in V_k$ can be defined as the distance between the VAR process and its innovations. Furthermore, we observe that the norm of a k -variate process $\mathbf{y} \in V_k$ depends on the sequences $\{\pi_{1i}\}, \dots, \{\pi_{ki}\}$. Since these sequences contain all information about the stochastic dependence structure of the process \mathbf{y} , we can interpret the norm of a VAR process like a measure of the stochastic dependence structure of the process. To illustrate this point, we consider the following Example 2.

Example 2.

$$\begin{bmatrix} 1 - 0.5L & 0 \\ 0 & 1 + 0.3L \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \epsilon_{x_t} \\ \epsilon_{y_t} \end{bmatrix} \quad (7)$$

The norm of process (7) is 0.8, less than the norm of the process (6), then we can say that in process (7) there is less "structure" than in the process (6). ■

The next proposition provides some main properties of the distance L .

Proposition 1. Let V_k be the class of the k -variate VAR processes. The function $L : V_k \times V_k \rightarrow \mathbb{R}$ defined as

$$L(\mathbf{v}_1, \mathbf{v}_2) = \sum_{i=1}^k d(v_{1i}, v_{2i}) \quad \mathbf{v}_1, \mathbf{v}_2 \in V_k,$$

satisfies the following properties:

- i. Non-negativity: $d(\mathbf{v}_1, \mathbf{v}_2) \geq 0 \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in V_k$;
- ii. Symmetry: $d(\mathbf{v}_1, \mathbf{v}_2) = d(\mathbf{v}_2, \mathbf{v}_1) \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in V_k$;
- iii. Triangularity: $d(\mathbf{v}_1, \mathbf{v}_2) \leq d(\mathbf{v}_1, \mathbf{v}_3) + d(\mathbf{v}_3, \mathbf{v}_2) \quad \forall \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V_k$;
- iv. $\mathbf{v}_1 = \mathbf{v}_2$ implies $d(\mathbf{v}_1, \mathbf{v}_2) = 0 \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in V_k$.

Proof. Evidently, $L(\mathbf{v}_1, \mathbf{v}_2)$ is a nonnegative function. Further, since $d(v_{1i}, v_{2i}) = d(v_{2i}, v_{1i})$ for $i = 1, \dots, k$, we have that

$$L(\mathbf{v}_1, \mathbf{v}_2) = \sum_{i=1}^k d(v_{1i}, v_{2i}) = \sum_{i=1}^k d(v_{2i}, v_{1i}) = L(\mathbf{v}_2, \mathbf{v}_1) \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in V_k.$$

and hence $L(\mathbf{v}_1, \mathbf{v}_2)$ is a symmetric function. In order to show the triangle inequality, we first note that

$$d(v_{1i}, v_{2i}) \leq d(v_{1i}, v_{3i}) + d(v_{3i}, v_{2i}),$$

where v_{3i} is the i^{th} univariate invertible ARMA component implied by the k -variate VAR processes v_3 . Thus

$$\begin{aligned} L(\mathbf{v}_1, \mathbf{v}_2) &\leq \sum_{i=1}^k [d(v_{1i}, v_{3i}) + d(v_{3i}, v_{2i})] \\ &= \sum_{i=1}^k d(v_{1i}, v_{3i}) + \sum_{i=1}^k d(v_{3i}, v_{2i}) \\ &= L(\mathbf{v}_1, \mathbf{v}_3) + L(\mathbf{v}_3, \mathbf{v}_2). \end{aligned}$$

Finally, it is clear that if $\mathbf{v}_1 = \mathbf{v}_2$ we have that $v_{1i} = v_{2i}$ for $i = 1, \dots, k$ and hence $d(v_{1i}, v_{2i}) = 0$ for $i = 1, \dots, k$. It follows that $L(\mathbf{v}_1, \mathbf{v}_2) = 0$. ■

It is important to underline that the distance between two VAR processes is

allowed to be null even if they are generated by different white noise processes. This implies that $L(\mathbf{x}, \mathbf{y})$ is a pseudometric.

In order to obtain an estimate of $L(\mathbf{x}, \mathbf{y})$ we use the following procedure:

Procedure 1

1. Estimate on the observed data the VAR(p) models for the processes \mathbf{x} and \mathbf{y} ;
2. using the estimated parameters from step 1 obtain the implied univariate ARMA models for the univariate processes x_i and y_i ($i = 1, \dots, k$);
3. evaluate the AR(∞) representation truncated a some suitable lag p^* of the ARMA models in step 2;
4. using the the AR(∞) representation from step 3 calculate the estimate distance $\hat{d}(x_i, y_i)$ ($i = 1, \dots, k$);
5. use the formula $\hat{L}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^k \hat{d}(x_i, y_i)$ as estimate the VAR distance $L(\mathbf{x}, \mathbf{y})$.

3 Spatial variability of the relationship between unemployment and GDP

The linkage between the rate of change in GDP and change in unemployment (the Okuns Law) is one of the most studied issue of empirical macroeconomics (see [1], [7], and [5], among others).

The aim of this section is to investigate the presence of spatial differences in the dynamic linkage between unemployment (U) and Gross Domestic Product (GDP) in thirteen European Union economies: Belgium, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Netherlands, Austria, Portugal, Finland and UK. The used quarterly data, from Eurostat database, are the Gross Domestic Product at market prices, chain linked volumes index with 2010 = 100, and the Total Unemployment rates. The data are seasonally and calendar adjusted

and the sample period is first quarter 1998 - fourth quarter of 2016 (1998Q1 - 2016Q4). Figure 3 and figure 3 describe the general behavior of the GDP and the Unemployment rates. The considered economies share more or less the same story before and after the 2009 crisis: a decreasing path for Unemployment until 2008 then an increasing path, an increasing path for GDP until 2008 then a growth slowdown, or a fall or a sharp fall for example in Greece.

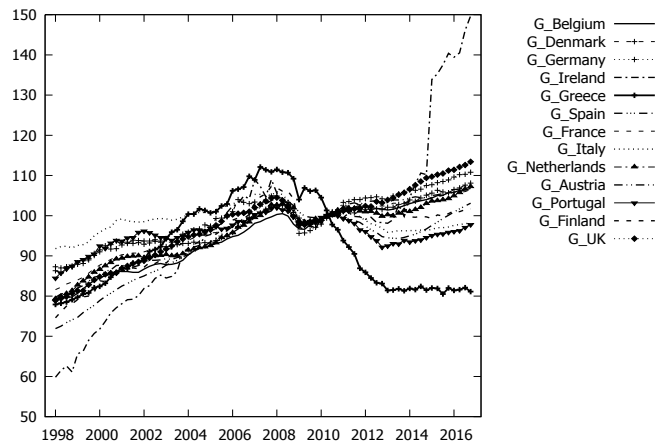


Fig. 1. Gross Domestic Product at market prices, chain linked volumes index 2010 = 100

There are some missing values for the Unemployment rates at the beginning of the period for some countries, as reported in Table 1.

	missing data
Be	1998: Q1, Q3 Q4
Dk	1998: Q1, Q3 Q4
Ge	1998, 1999, 2000, 2001, 2002, 2003, 2004: Q1, Q3 Q4
Ei	1998: Q1, Q3 Q4; 1999: Q1
Fr	1998, 1999, 2000, 2001, 2002: Q1, Q3 Q4
Nl	1998, 1999 : Q1, Q3 Q4
Au	1998: Q1, Q3 Q4
Uk	1998: Q1, Q3 Q4

Table 1. Missing data

The missing data are imputed using backcasting. The usual preliminary analysis shows some of the series have outliers that have been corrected using

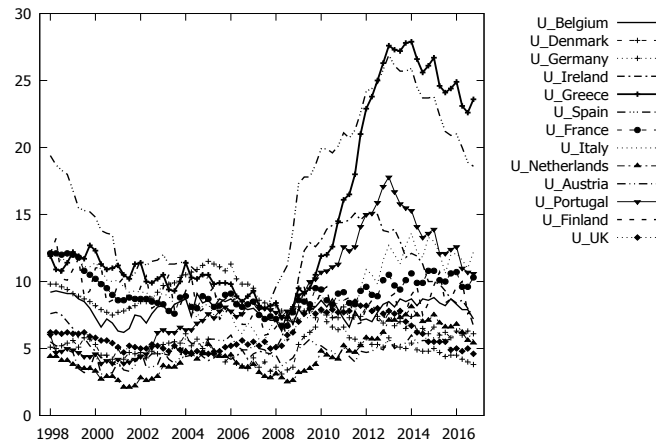


Fig. 2. Total Unemployment rates

TRAMO. More, using ADF unit root test, with lags selection according the Bai-Ng test, we verified that all the series are $I(1)$. Then we decided to consider the relationship between the rates of growth of unemployment rate and GDP. Lags selection procedure for the VARs, based on BIC criteria, choose just one lag for all countries. The relationship is thus analyzed through bivariate VAR(1) model for the variables $\Delta \log(U)$ and $\Delta \log(GDP)$. We apply for each countries the Procedure 1 described above evaluating proposed distance (5) between any pair of models, setting $p^* = 15$, obtaining the matrix of distances reported in Table 2.¹

A useful way to visualizing the information contained in a distance matrix between units is to conduct a MultiDimensional Scaling (MDS) analysis.² Given the matrix of distances among VARs presented in Table 2, classical MDS produces the map reported in Figure 3.

The analysis by MDS of the distance matrix shows that we can identify the following clouds of similarity:

1. France, Germany, Portugal;
2. Italy, Spain;
3. Austria, Denmark

¹ Full results concerning the estimation of the VAR models are available from the authors on request.

² See [3]. From a non-technical point of view, the purpose of MDS is to provide a visual representation of the pattern of distances among a set of objects. Given a matrix of distances between various objects, MDS plots the objects on a map such that those objects that are very similar to each other are placed near each other on the map, and those objects that are very different from each other are placed far away from each other on the map.

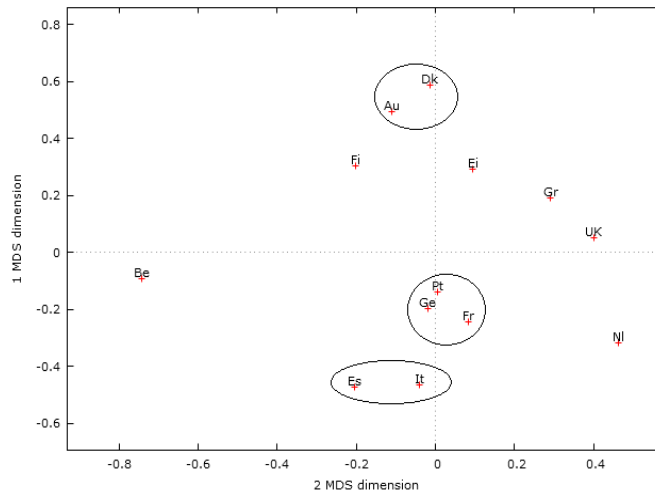


Fig. 3. Two-dimensional scatterplot of countries obtained by multidimensional scaling.

The other countries reveal peculiar paths. This is true, in particular, for Belgium and Netherlands. The relative proximity between Italy and Spain seems to be overwhelmed by the recent dynamics of both the labor market and GDP resulting from the crisis that has hit Europe since 2009. In the same way the well known situation in Greece explains its position in the figure. It seems more difficult to explain, however, the position of Portugal near to Germany and France, even if a graphical inspection reveals a similar general dynamic in the GDP growth rate before 2009. Belgium and Netherlands seem to be very far each other; also in this case an explanation can be found in the peculiar political situation experimented by Belgium when the political parties were not able to form a government for at least 2 year starting from June 2010. The overall conclusion is that, despite the ongoing integration within the EU, there are still significant differences among countries regarding the dynamic link between unemployment rate variation and economic growth.

4 Conclusions

There are many circumstances in which is important to compute a distance measure for multivariate time series. In this paper, a novel notion of distance measure between pairs of VAR models has been introduced and its main properties have been discussed. We have used such notion to investigate the presence of spatial differences in the dynamic linkage between unemployment rate variation and economic growth in 13 European Union economies. The analysis reveals that, despite the ongoing integration within the EU, many countries have special positions of their own.

	Be	Dk	Ge	Ei	Gr	Es	Fr	It	Nl	Au	Pt	Fi	UK
Be	0.000	0.967	0.682	0.925	1.112	0.707	0.787	0.667	1.216	0.751	0.715	0.569	1.155
Dk		0.000	0.787	0.346	0.411	1.098	0.840	1.050	1.098	0.217	0.727	0.479	0.654
Ge			0.000	0.496	0.463	0.386	0.133	0.285	0.549	0.696	0.062	0.568	0.480
Ei				0.000	0.259	0.840	0.572	0.779	0.794	0.295	0.435	0.370	0.401
Gr					0.000	0.784	0.432	0.644	0.690	0.371	0.424	0.587	0.244
Es						0.000	0.458	0.285	0.885	1.010	0.430	0.903	0.827
Fr							0.000	0.212	0.472	0.752	0.152	0.647	0.370
It								0.000	0.608	0.962	0.344	0.850	0.543
Nl									0.000	1.019	0.531	0.892	0.455
Au										0.000	0.636	0.296	0.567
Pt											0.000	0.508	0.447
Fi												0.000	0.637
Uk													0.000

Table 2. Distance matrix

Bibliography

- [1] Lee, J. , The Robustness of Okun's Law: Evidence From OECD Countries, *Journal of Macroeconomics*, 331-56 (2000)
- [2] Lütkepohl, H. *New introduction to multiple time series analysis*, Springer, Berlin, (2005).
- [3] Mardia, K. V., J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press (1979)
- [4] Piccolo, D., A distance measure for classifying ARIMA models. *Journal of time Series Analysis*, 11, 153-164 (1990)
- [5] Prachowny, M. F. J. , Okuns Law: Theoretical Foundations and Revised Estimates, *The Review of Economics and Statistics*, 75(2), 331-336 (1993)
- [6] Sims C., Macroeconomics and Reality. *Econometrica* 48, 1-48 (1980)
- [7] Watts, M., and W. Mitchell , Alleged Instability of the Okun's Law relationship in Australia: An Empirical Analysis, *Applied Economics*, 1829-1838 (1991)
- [8] Zellner, A., and Palm, F., Time series analysis and simultaneous equation econometric models. *Journal of Econometrics* 2, 17-54 (1974)

A Least-Squares Approach to Estimate the Impulse-Response Function of a General Linear Model

Miguel Jerez and Alfredo Garcia-Hiernaux
mjerezme@ucm.es agarciah@ucm.es

Universidad Complutense de Madrid

Abstract. We discuss how to estimate the impulse-response function of a general linear model by regressing its endogenous variables on the lagged residuals derived from a general linear model. Therefore, it amounts to estimating a truncated Wold form by least-squares. In comparison with the standard VAR approach, our method accommodates any formulation for the data model and only requires a least-squares routine for implementation. On the other hand, its least-squares foundation is an advantage by itself, as it provides easy solutions to many difficult problems related with IR analysis, such as computing analytical standard errors for the response coefficients or testing for causality. Besides describing the procedure in detail, we provide several examples and discuss its pros and cons in comparison with the alternatives.

Keywords. Impulse-response analysis; VAR models; Least-squares.

1 Introduction

This paper describes a procedure to estimate the impulse-response (IR) function of a general linear model by regressing its endogenous variables on the lagged residuals derived from a general linear model. Therefore, it amounts to estimating a truncated Wold [1] form by least-squares (LS).

The standard approach to IR analysis typically inverts a VAR model to obtain the corresponding Wold form. In comparison, our direct estimation approach has important advantages as it accommodates any formulation for the data model and only requires a LS routine for implementation. On the other hand its LS foundation is an advantage by itself, as it provides easy and well-tested solutions to many difficult problems related with IR analysis, such as computing analytical standard errors for the response coefficients or testing for causality.

The structure of the paper is as follows: Section 2 defines the basic notation and summarizes some previous results. Section 3 describes the basic methods and Section 4 discusses some extensions, such as models with orthogonal errors

or exogenous inputs. Last, Section 5 discusses the pros and cons of the proposed approach in comparison with its alternatives.

2 Notation and previous results

2.1 The impulse-response form of a stochastic process

Let $\mathbf{z}_t \in \mathbb{R}^m$ be a random vector of *endogenous variables* or *outputs* at time t , which has been decomposed using a previously fitted data model as:

$$\mathbf{z}_t = \hat{\mathbf{E}}(\mathbf{z}_t | \boldsymbol{\Omega}_{t-1}) + \hat{\mathbf{a}}_t, \quad (1)$$

where $\boldsymbol{\Omega}_{t-1}$ is the information set containing all the information available up to time $t-1$, $\hat{\mathbf{E}}(\cdot | \cdot)$ denotes the (estimated) expected value of the first argument, conditional to the information set in the second argument, and $\hat{\mathbf{a}}_t$ is a vector of zero-mean white noise residuals.

As it is well known, if the stochastic process underlying the decomposition (1) is zero-mean and covariance-stationary, \mathbf{z}_t can be alternatively represented by the corresponding Wold [1] form:

$$\mathbf{z}_t = \hat{\mathbf{a}}_t + \hat{\boldsymbol{\Psi}}_1 \hat{\mathbf{a}}_{t-1} + \hat{\boldsymbol{\Psi}}_2 \hat{\mathbf{a}}_{t-2} + \dots, \quad (2)$$

where the weights $\boldsymbol{\Psi}_i$ ($i=1,2,\dots$) must be square summable to assure the stability of the stochastic process, see Hamilton [2].

The coefficient matrices $\hat{\boldsymbol{\Psi}}_i$ ($i=1,2,\dots$) in (2) provide the expected response of \mathbf{z}_t to a unit impulse in $\hat{\mathbf{a}}_{t-i}$, so determining the IR function in practice reduces to computing these matrices.

2.2 The VAR approach to estimate the impulse-response function

In applied macroeconometrics, the standard approach to estimate the IR coefficients consists in: (a) fitting a vector autoregressive (VAR) model to \mathbf{z}_t :

$$\left(\mathbf{I} - \hat{\boldsymbol{\Pi}}_1 \mathbf{B} - \hat{\boldsymbol{\Pi}}_2 \mathbf{B}^2 - \dots - \hat{\boldsymbol{\Pi}}_p \mathbf{B}^p \right) \mathbf{z}_t = \hat{\mathbf{a}}_t, \quad (3)$$

where \mathbf{B} denotes the backshift operator, such that for any sequence ω_t : $\mathbf{B}^i \omega_t = \omega_{t-i}$, $i = 0, \pm 1, \pm 2, \dots$, (b) computing the sequence $\hat{\boldsymbol{\Psi}}_i$ ($i=1,2,\dots$) by inverting the polynomial in the left-hand-side of (3):

$$\mathbf{z}_t = \left[\mathbf{I} - \hat{\boldsymbol{\Pi}}_1 \mathbf{B} - \hat{\boldsymbol{\Pi}}_2 \mathbf{B}^2 - \dots - \hat{\boldsymbol{\Pi}}_p \mathbf{B}^p \right]^{-1} \hat{\mathbf{a}}_t, \quad (4)$$

and then, if required, (c) obtaining standard errors for the IR coefficients resulting from (4) either by an asymptotic approximation or bootstrapping, see Lütkepohl [3] or Hamilton [2].

3 Basic methods

3.1 Least-squares estimation of the impulse-response function

Point estimation The first L parameter matrices in (2) can be estimated consistently by applying LS to the truncated Wold form:

$$\mathbf{z}_t - \hat{\mathbf{a}}_t = \Psi_1 \hat{\mathbf{a}}_{t-1} + \Psi_2 \hat{\mathbf{a}}_{t-2} + \dots + \Psi_L \hat{\mathbf{a}}_{t-L} + \varepsilon_t, \quad (5)$$

where the residuals $\hat{\mathbf{a}}_{t-k}$ ($k = 0, 1, \dots, L$) in (5) are obtained from (1) and the term ε_t represents the terms omitted in the right-hand-side of (5), i.e., $\Psi_{L+1} \hat{\mathbf{a}}_{t-L-1} + \Psi_{L+2} \hat{\mathbf{a}}_{t-L-2} + \dots$. Because of this, we will refer to ε_t as the “approximation error”. Note that:

1. The stochastic properties of ε_t depend on those of the omitted residuals $\hat{\mathbf{a}}_{t-L-1}, \hat{\mathbf{a}}_{t-L-2}, \dots$
2. In particular, if these residuals are zero-mean, homoscedastic, non-autocorrelated and Gaussian, the approximation error will have the same properties.
3. Lack of autocorrelation of $\hat{\mathbf{a}}_{t-L-1}$ is very important, as it implies that:
 - (a) ...the regressors in the right-hand-side of (5) are asymptotically independent among themselves and from the approximation errors, so that
 - (b) ...the degree of collinearity between the regressors in (5) will be low, and
 - (c) ...the omitted regressors will be asymptotically independent from those included in (5), so LS estimates will be consistent and robust to the truncation lag.
4. The determination coefficient obtained when estimating model (5) measures the approximation achieved because, as the truncation lag L increases, the determination coefficient converges to unity and, accordingly, the covariance of the approximation error converges to zero.

Precedents The idea of using residuals to estimate a MA structure by LS is not new. It can be traced back to the seminal paper by Durbin [4], where the estimated residuals from a first-stage long autoregression were used to replace the unobservable innovations in the MA term. The same idea was later applied in different frameworks by Hannan and Rissanen [5, 6], Spliid [7], Hannan and Kavalieris [8], Koreisha and Pukkila [9] or Flores and Serrano [10]. All these papers concentrate in model estimation and conclude that the estimates obtained are very precise. On the other hand, they do not investigate the potential of this approach to estimate IR functions.

Another branch of the literature explored the idea of estimating the IR function by LS methods, with emphasis in estimating “model-free” responses. In this line, Jorda [11] proposed an IR estimator computed by running a sequence of

predictive regressions of the variable of interest on a structural shock, for different prediction horizons. The IR is then given by the sequence of LS coefficients of the shock. An alternative and equivalent procedure by Chang and Sakata [12] combines a long first-stage autoregression with a second-stage MA fit.

The work of Chang and Sakata [12] is therefore close to our proposal with some differences, as we allow for a general first-stage model and, in the next sections, take advantage of LS results to solve nontrivial problems such as, e.g., computing analytical standard errors for the IR coefficients.

Example #1: Precision of IR point estimates Table 1 shows four ARIMA models with autoregressive and moving average structure. We simulated a sample of 200 observations of the stochastic process in the column “True model” and then estimated it using Gretl [13] exact maximum likelihood algorithm. The results obtained are summarized in the column “Estimated model”.

Table 1. Univariate models. The estimates shown in the last column have been calculated with Gretl 2016d exact maximum likelihood procedure.

Series	True model	Estimated model
A	$(1 - .7B)z_t = a_t$; $\sigma_a^2 = 1$	$(1 - .722B)z_t = \hat{a}_t$; $\hat{\sigma}_a^2 = 1.003$ (.049)
B	$z_t = (1 - .8B)a_t$; $\sigma_a^2 = 1$	$z_t = (1 - .808B)\hat{a}_t$; $\hat{\sigma}_a^2 = 1.016$ (.040)
C	$(1 - .7B + .6B^2)z_t = a_t$; $\sigma_a^2 = 1$	$(1 - .721B + .577B^2)z_t = \hat{a}_t$; $\hat{\sigma}_a^2 = .940$ (.058) (.057)
D	$z_t = (1 - .8B + .6B^2)a_t$; $\sigma_a^2 = 1$	$z_t = (1 - .843B + .628B)\hat{a}_t$; $\hat{\sigma}_a^2 = .927$ (.051) (.059)

Building on these models, we have computed: (a) the “true” IR function, derived analytically from the true parameter values, as well as (b) ten IR values corresponding to “Estimated model” using both, the standard method described in Subsection 2.2¹ and the LS procedure proposed in previous sections. Table 2 displays the Root Mean Squared Errors (RMSEs) corresponding to both approaches, being both remarkably small and similar.

Example #2: Robustness of IRs to the truncation lag As noted before, the regressors in (5) are (supposedly) white noise residuals. If so, the variables omitted by truncation in (5) and accumulated in the approximation error, ε_t , are asymptotically independent from those included in the model, so the IR estimates resulting from our procedure should be robust to the truncation lag.

¹ In the case of MA structures, the standard IR values have been set to zero when the lag is higher than the order of the stochastic process. This criterion provides the standard approach with a deliberate advantage, as most IR lags are computed without error.

Table 2. RMSE of the IR estimates computed with the standard and alternative approaches. The smallest RMSE in each row is underlined.

	Model	Standard approach	LS approach	Standard - LS
Series A	AR(1)	<u>2.414E-02</u>	2.506E-02	-9.281E-04
Series B	MA(1)	2.649E-03	<u>2.632E-03</u>	1.699E-05
Series C	AR(2)	<u>3.223E-02</u>	3.282E-02	-5.884E-04
Series D	MA(2)	1.626E-02	<u>1.617E-02</u>	9.011E-05

The Table 3 shows the first 10 IR weights estimated for series A, computed with $L = 10, 15, 20$. Note that, for this sample size ($N=200$): (a) changes in IR estimates typically occur in the 2^{nd} or 3^{rd} decimal place, and (b) the accumulated response estimated with the first 10 coefficients is very robust.

Table 3. Robustness of IR estimates to the specification of the truncation lag. The table shows the first 10 point estimates for the IR coefficients computed with $L=10$, $L=15$ and $L=20$. The last rows display respectively the sum of the coefficients displayed, which would be the 10 lag accumulated response or “gain”, and the determination coefficient of the corresponding regression.

Regressor	$L=10$	$L=15$	$L=20$
\hat{a}_{t-1}	0.7173	0.7226	0.7219
\hat{a}_{t-2}	0.5171	0.5220	0.5213
\hat{a}_{t-3}	0.3757	0.3767	0.3763
\hat{a}_{t-4}	0.2722	0.2719	0.2716
\hat{a}_{t-5}	0.2001	0.1960	0.1962
\hat{a}_{t-6}	0.1463	0.1405	0.1417
\hat{a}_{t-7}	0.1069	0.1012	0.1023
\hat{a}_{t-8}	0.0776	0.0735	0.0738
\hat{a}_{t-9}	0.0550	0.0532	0.0533
\hat{a}_{t-10}	0.0380	0.0392	0.0384
Sum (10-lag gain)	2.5062	2.4967	2.4968
R^2	0.9987	0.9999	1.0000

3.2 Confidence intervals

Calculation of asymptotic standard errors The standard errors for the parameters in (5) provided by a standard LS routine are not adequate. To see this, note that they would be computed with the sample covariance of the approximation errors ε_t , $\hat{\Sigma}_\varepsilon$, thus ignoring the uncertainty in $\hat{\mathbf{a}}_t$.

It is easy to compute asymptotical standard errors by replacing in the LS formulas the covariance $\hat{\Sigma}_\varepsilon$ by any of the following alternatives: $\hat{\Sigma}_a$ or $\hat{\Sigma}_\varepsilon + \hat{\Sigma}_a$, being $\hat{\Sigma}_a$ the residual in the decomposition (1).

As noted before, $\hat{\Sigma}_\varepsilon$ converges to zero as the truncation lag L increases. This implies that:

1. ...estimating expression (5) produces consistent estimates for the corresponding IR coefficients,
2. ...when L tends to infinity the only source of uncertainty in the model is a_t , which justifies replacing $\hat{\Sigma}_\varepsilon$ by $\hat{\Sigma}_a$,
3. ...on the other hand, one may want to compute a more conservative (i.e. larger) standard error, by using $\hat{\Sigma}_\varepsilon + \hat{\Sigma}_a$ as replacement matrix, bearing in mind that,
4. ...both, $\hat{\Sigma}_a$ and $\hat{\Sigma}_\varepsilon + \hat{\Sigma}_a$ converge to the same limit, Σ_a , as L and T tend to infinity.

Example #3: Confidence intervals for IR coefficients and comparison with standard software Figure 1 shows the IRs and the LS 95% confidence interval obtained for series A, see the example in subsection 3.1. Note that the interval is symmetric and has a width which increases with the lag.

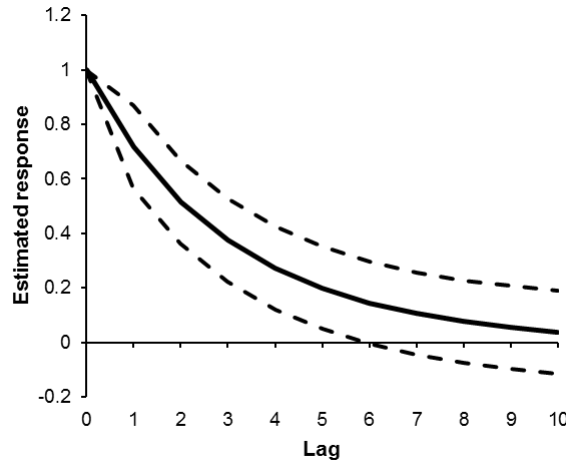


Fig. 1. Impulse response function and LS 95% confidence interval obtained for series A using the LS method proposed. Standard errors provided by the LS formulae have been rescaled using $\hat{\Sigma}_\varepsilon + \hat{\Sigma}_a$ as replacement matrix.

Figure 2 displays the IR analysis result calculated with Gretl [13]. In particular, it is the response to one sigma innovation with a 95% confidence interval calculated by bootstrap. These results are somewhat surprising because: (a) 0-lag response seems to have some uncertainty, (b) the IR confidence interval is not asymmetric, and (c) it displays a “banana shape”, meaning that its width decreases for lags greater than 4.

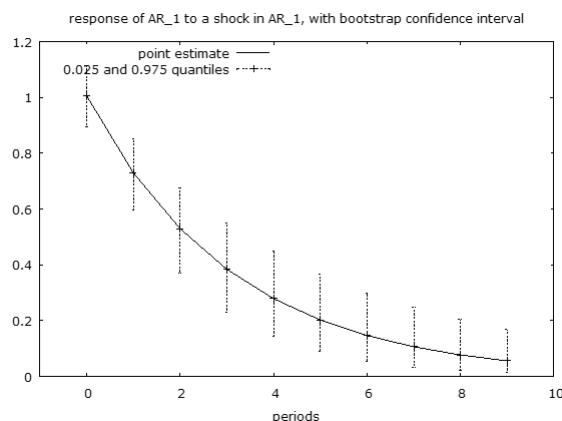


Fig. 2. Impulse response function and bootstrap 95% confidence interval obtained for series A using Gretl 2016d. The figure displayed is a direct copypaste of Gretl’s output.

On the other hand, Figure 3 shows the analogous results computed with Eviews [14]. Specifically, it is the response to a unit innovation with a 95% confidence interval calculated by a method which Eviews describes as “asymptotic s.e.”. In this case: (a) the 0-lag response has no uncertainty, as it could be expected, (b) the confidence interval is symmetric, and (c) it also displays the aforementioned “banana shape” configuration, as its width decreases for lags greater than 5.

Note that the differences in the confidence intervals may affect the conclusions of inference, as the response becomes non-significant around lags 6-7 for the results computed with our approach and Eviews, while it is significant for any lag with the response computed by Gretl.

4 Extensions

4.1 Response to orthogonal shocks

IR functions are interpreted under the assumption that a single error is pulsed while all the others remain constant. However this assumption is not

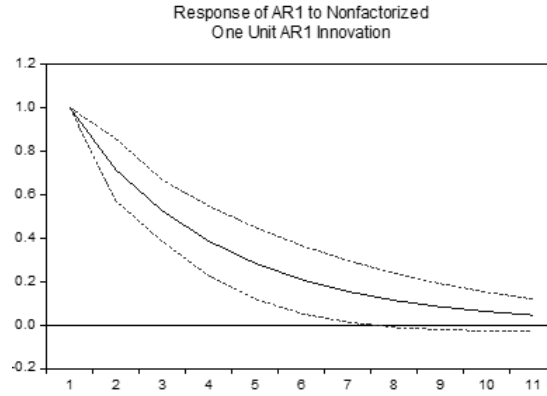


Fig. 3. Impulse response function and asymptotic 95% confidence interval obtained for series A using EViews 8.1. The figure displayed is a direct cypypaste of EViews' output.

realistic because the errors in a model are not orthogonal in most cases. Because of this, most IR analysis is carried out after transforming the data model, which is given by equation (1) in our case, to a structural representation excited by orthogonal shocks.

To this end, we define the change of variables $\hat{\mathbf{a}}_t = \mathbf{U}\hat{\mathbf{a}}_t^*$, where the covariance $E(\hat{\mathbf{a}}_t\hat{\mathbf{a}}_t^T) = \Sigma$ is a general positive semi-definite matrix, $E(\hat{\mathbf{a}}_t^*\hat{\mathbf{a}}_t^{*T}) = \Sigma^*$ is diagonal and the transformation matrix \mathbf{U} is nonsingular². Applying this change of variables to the model (5) we obtain:

$$\mathbf{z}_t - \mathbf{U}\hat{\mathbf{a}}_t^* = \Psi_1\mathbf{U}\hat{\mathbf{a}}_{t-1}^* + \Psi_2\mathbf{U}\hat{\mathbf{a}}_{t-2}^* + \dots + \Psi_L\mathbf{U}\hat{\mathbf{a}}_{t-L}^* + \varepsilon_t, \quad (6)$$

and the response function to the orthogonal shocks $\hat{\mathbf{a}}_t^*$ would be given by: $\Psi_0^* = \mathbf{U}$, $\Psi_1^* = \Psi_1\mathbf{U}$, $\Psi_2^* = \Psi_2\mathbf{U}$, ...

Therefore, after choosing a matrix \mathbf{U} and computing the corresponding structural residuals $\hat{\mathbf{a}}_t^* = \mathbf{U}^{-1}\hat{\mathbf{a}}_t$, one can estimate the structural coefficient matrices and their standard errors by applying LS to the model:

$$\mathbf{z}_t = \Psi_0^*\hat{\mathbf{a}}_t^* + \Psi_1^*\hat{\mathbf{a}}_{t-1}^* + \Psi_2^*\hat{\mathbf{a}}_{t-2}^* + \dots + \Psi_L^*\hat{\mathbf{a}}_{t-L}^* + \varepsilon_t, \quad (7)$$

² As it is well known, the matrix \mathbf{U} is not identified in general and the literature proposes many procedures to determine it. Many works concentrate in choosing \mathbf{U} so that the covariance of the transformed errors has a triangular or block-triangular causal structure, see e.g. Sims [15] and Bernanke and Blinder [16], respectively. Another approach concentrates in imposing meaningful constraints on the IR, either on the long-term responses, see Blanchard and Quah [17], or over the sign of some IR values, see Uhlig [18].

4.2 Response to pulsed inputs and testing for exogeneity

Wold's [1] representation theorem states that every covariance-stationary time series \mathbf{z}_t can be written as the sum of two time series, one deterministic and one stochastic. This result points to a straightforward generalization of model (5) to accommodate exogenous inputs:

$$\mathbf{z}_t - \hat{\mathbf{a}}_t = \mathbf{\Gamma}_0 \mathbf{u}_t + \mathbf{\Gamma}_1 \mathbf{u}_{t-1} + \mathbf{\Gamma}_2 \mathbf{u}_{t-2} + \dots + \mathbf{\Gamma}_{L_u} \mathbf{u}_{t-L_u} + \mathbf{\Psi}_1 \hat{\mathbf{a}}_{t-1} + \mathbf{\Psi}_2 \hat{\mathbf{a}}_{t-2} + \dots + \mathbf{\Psi}_{L_a} \hat{\mathbf{a}}_{t-L_a} + \varepsilon_t, \quad (8)$$

so \mathbf{z}_t receives shocks from both, the model inputs and the errors, and L_u and L_a denote, respectively, the truncation lags for the input and error-related IR weights.

Expression (8) suggests flexible and easy ways to test for exogeneity, for example by computing a LR test comparing the Gaussian likelihood values corresponding to the unconstrained model (8), and a constrained version of (8), including as many exclusion restrictions as required by the null to be tested.

5 Pros and cons of the proposed approach

In comparison with the standard approach summarized in Subsection 2.2, the procedure proposed in this paper has the following pros:

1. Its implementation only requires a standard LS routine and data for the output and input variables, as well as a time series of uncorrelated residuals.
2. Therefore, it is not constrained to a specific model formulation and, in particular, can be applied directly to any linear model, including VAR, VARMA or VARMAX.
3. It estimates the IR coefficients as regression parameters, so analytical standard errors can be computed from standard LS results, see Subsection 3.2.
4. Last, its LS foundations allows one to apply immediately a wealth of associated results and techniques such as, e.g., combining the IR estimates with heteroscedasticity and autocorrelation-consistent standard errors, see e.g. White [19] and Newey and West [20].

On the other hand, it also has some cons in comparison with standard approaches because:

1. ...estimating the regression (5) consumes degrees of freedom, and
2. ...the resulting IRs are somewhat "ragged", as the coefficients are estimated directly. On the other hand, smoothness can be easily achieved by fitting a high-degree polynomial to the sequence of estimated IR values.

Our method may have other advantages which have not been explored in this paper. In particular, estimating directly the IR form probably simplifies sophisticated IR analyses such as, e.g., imposing constraints or applying a Bayesian analysis of the IR. Last, in the nonlinear case our procedure may offer an approximation to the system response by a linear IR function. The precision of such approximation would obviously depend on the type of nonlinearity.

Acknowledgments. Support from Instituto Complutense de Analisis Economico (ICAE) and grant PR26/16-20270 from *Programa de ayudas a proyectos de investigacion Santander-UCM* is gratefully acknowledged.

References

1. Wold, H.O.A.: A Study in the Analysis of Stationary Time Series. Almqvist and Wiksell, Uppsala (1964)
2. Hamilton, H.D.: Time Series Analysis. Princeton University Press (1993)
3. Lutkepohl, H.: New Introduction to Multiple Time Series Analysis. Springer-Verlag, Berlin (2005)
4. Durbin, J.: The fitting of time-series models. *Revue de l' Institut International de Statistique* 28, 233-244 (1960)
5. Hannan, E.J., Rissanen, J.: Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, 69, 819-824 (1982.)
6. Hannan, E., Rissanen, J.: Amendments and corrections: recursive estimation of mixed autoregressive-moving average order. *Biometrika* 70, 30 (1983)
7. Spliid, H.: A fast estimation method for the vector autoregressive moving average model with exogenous variables. *Journal of the American Statistical Association*, 78, 384, 843-849 (1983)
8. Hannan, E., Kavalieris, L.: A method for autoregressive-moving average estimation. *Biometrika*, 71, 273-80 (1984)
9. Koreisha, S.G., Pukkila, T.H.: Fast linear estimation methods for vector autoregressive moving average models. *Journal of Time Series Analysis*, 10, 325-339 (1989)
10. Flores, R., Serrano, G.: A generalized least squares estimation method for VARMA models. *Statistics*, 36, 4, 303-316 (2002)
11. Jorda, O.: Estimation and inference of impulse responses by local projections. *American Economic Review* 95, 161-182 (2005)
12. Chang, P., Sakata, S.: Estimation of impulse response functions using long autoregression. *Econometrics Journal*, 10, 453-469 (2007)
13. Baiochi, G., Distaso, W.: Gretl: Econometric software for the GNU generation. *Journal of Applied Econometrics*, 18, 105-110 (2003)
14. Ouliaris, A., Pagan, A., Restrepo, J.: Quantitative Macroeconomic Modeling with Structural Vector Autoregressions An EViews Implementation. This book can be freely downloaded from <http://www.eviews.com/StructVAR/structvar.html> (2016)
15. Sims, C.A.: Macroeconomics and reality. *Econometrica*, 48, 1-47 (1980)
16. Bernanke, B.S., Blinder, A.: The federal funds rate and the channels of monetary transmission. *American Economic Review*, 82, 901-921. (1992)

17. Blanchard, O.J., Quah, D.T.: The dynamic effects of aggregate demand and supply disturbances. *American Economic Review*, 79, 655-73 (1989)
18. Uhlig, H.: What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics*, 52, 381-419 (2005)
19. White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 4, 817-838 (1980)
20. Newey, W.K., West, K.D.: A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55, 3, 703-708 (1987)

Recovering the background noise of a Lévy-driven CARMA process using an SDDE approach

Mikkel Slot Nielsen and Victor Rohde*

Aarhus University, Department of Mathematics,
Ny Munkegade 118, 8000, Aarhus C, Denmark
`{mikkel,victor}@math.au.dk`

Abstract Based on a vast amount of literature on continuous-time ARMA processes, the so-called CARMA processes, we exploit their relation to stochastic delay differential equations (SDDEs) and provide a simple and transparent way of estimating the background driving noise. An estimation technique for CARMA processes, which is particularly tailored for the SDDE specification, is given along with an alternative and (for the purpose) suitable state-space representation. Through a simulation study of the celebrated CARMA(2, 1) process we check the ability of the approach to recover the distribution.

Keywords: continuous-time ARMA process; Lévy processes; noise estimation; stochastic volatility

1 Introduction

Continuous-time ARMA processes, specifically the class of CARMA processes, have been studied extensively and found several applications. The most basic CARMA process is the CAR(1) process, which corresponds to the Ornstein-Uhlenbeck process. This process serves as the building block in stochastic modeling, e.g., Barndorff-Nielsen and Shephard [1] use it as the stochastic volatility component in option pricing modeling and Schwartz [14] models (log) spot price of many different commodities through an Ornstein-Uhlenbeck specification. More recently, several researchers have paid attention to higher order CARMA processes. To give a few examples, Brockwell et al. [8] model turbulent wind speed data as a CAR(2) process, García et al. [11] and Benth et al. [3] fit a CARMA(2, 1) process to electricity spot prices, and Benth et al. [4] find a good fit of the CAR(3) to daily temperature observations (and thus, suggests a suitable model for the OTC market for temperature derivatives). In addition, as for the CAR(1) process, several studies have concerned the use of CARMA processes in the modeling of stochastic volatility (see, e.g., [7, 15, 17]).

From a statistical point of view, as noted in the above references, the ability to recover the underlying noise of the CARMA process is important. However,

* This work was supported by the Danish Council for Independent Research (Grant DFF - 4002 - 00003)

while it is possible to recover the driving noise process, it is a subtle task. Due to the non-trivial nature of the typical algorithm, see [7], implementation is not straightforward and approximation errors may be difficult to locate. The recent study of Basse-O'Connor et al. [2] on processes of ARMA structure relates CARMA processes to certain stochastic (delay) differential equations, and this leads to an alternative way of backing out the noise from the observed process which is transparent and easy to implement. The contribution of this paper is exploiting this result to get a simple way to recover the driving noise. The study both relies and supports the related work of Brockwell et al. [7].

Section 2 recalls a few central definitions and gives a dynamic interpretation of CARMA processes by relating them to solutions of stochastic differential equations. Section 3 briefly discusses how to do (consistent) estimation and inference in the dynamic model and, finally, in Section 4 we investigate through a simulation study the ability of the approach to recover the distribution of the underlying noise for two sample frequencies.

2 CARMA processes and their dynamic SDDE representation

Recall that a Lévy process is interpreted as the continuous-time analogue to the (discrete-time) random walk. More precisely, a (one-sided) Lévy process $(L_t)_{t \geq 0}$, $L_0 = 0$, is a stochastic process having stationary independent increments and càdlàg sample paths. From these properties it follows that the distribution of L_1 is infinitely divisible, and the distribution of $(L_t)_{t \geq 0}$ is determined by the one of L_1 according to the relation

$$\mathbb{E}[e^{iyL_t}] = \mathbb{E}[e^{iyL_1}]^t$$

for $y \in \mathbb{R}$ and $t \geq 0$. The definition is extended to a two-sided Lévy process $(L_t)_{t \in \mathbb{R}}$, $L_0 = 0$, which can be constructed from a one-sided Lévy process $(L_t^1)_{t \geq 0}$ by taking an independent copy $(L_t^2)_{t \geq 0}$ and setting $L_t = L_t^1$ if $t \geq 0$ and $L_t = -L_{(-t)-}^2$ if $t < 0$. Throughout, $(L_t)_{t \in \mathbb{R}}$ denotes a two-sided Lévy process, which is assumed to be square integrable.

Next, we will give a brief recap of Lévy-driven CARMA processes. (For an extensive treatment, see [5, 7, 9].) Let $p \in \mathbb{N}$ and set

$$P(z) = z^p + a_1 z^{p-1} + \cdots + a_p \quad \text{and} \quad Q(z) = b_0 + b_1 z + \cdots + b_{p-1} z^{p-1} \quad (2.1)$$

for $z \in \mathbb{C}$ and $a_1, \dots, a_p, b_0, \dots, b_{p-1} \in \mathbb{R}$. Define

$$\tilde{A}_p = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_p & -a_{p-1} & -a_{p-2} & \cdots & -a_1 \end{bmatrix},$$

$e_p = [0 \ 0 \ \cdots \ 0 \ 1]' \in \mathbb{R}^p$, and $b = [b_0 \ b_1 \ \cdots \ b_{p-2} \ b_{p-1}]'$. In order to ensure the existence of a casual CARMA process we will assume that the eigenvalues of \tilde{A}_p or, equivalently, the zeroes of P all have negative real parts. Then there is a unique (strictly) stationary \mathbb{R}^p -valued process $(X_t)_{t \in \mathbb{R}}$ satisfying

$$dX_t = \tilde{A}_p X_t dt + e_p dL_t, \quad (2.2)$$

and it is explicitly given by $X_t = \int_{-\infty}^t e^{\tilde{A}_p(t-u)} e_p dL_u$ for $t \in \mathbb{R}$. For a given $q \in \mathbb{N}_0$ with $q < p$, we set $b_q = 1$ and $b_j = 0$ for $q < j < p$. A CARMA(p, q) process $(Y_t)_{t \in \mathbb{R}}$ is then the strictly stationary process defined by

$$Y_t = b' X_t \quad (2.3)$$

for $t \in \mathbb{R}$. This is the state-space representation of the formal stochastic differential equation

$$P(D)Y_t = Q(D)DL_t, \quad (2.4)$$

where D denotes differentiation with respect to time. One says that $(Y_t)_{t \in \mathbb{R}}$ is causal, since Y_t is independent of $(L_s - L_t)_{s > t}$ for all $t \in \mathbb{R}$. We will say that $(Y_t)_{t \in \mathbb{R}}$ is invertible if all the zeroes of Q have negative real parts. The word "invertible" is justified by Theorem 1 below and the fact that this is the assumption imposed in [7] in order to make the recovery of the increments of the Lévy process possible. In Figure 1 we have simulated a CARMA(2, 1) process driven by a gamma (Lévy) process and by a Brownian motion, respectively.

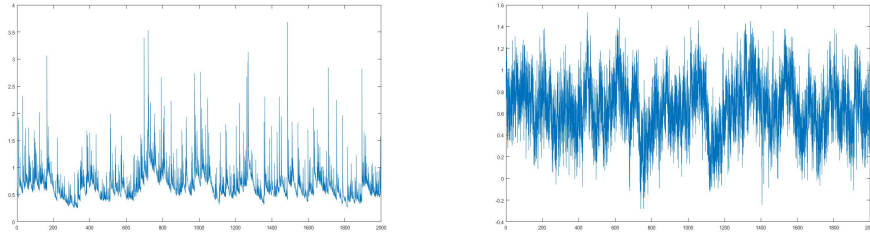


Figure 1. A simulation of a CARMA(2, 1) process with parameters $a_1 = 1.3619$, $a_2 = 0.0443$, and $b_0 = 0.2061$. It is driven by a gamma (Lévy) process with parameters $\lambda = 0.2488$ and $\xi = 0.5792$ on the left and a Brownian motion with mean $\mu = 0.1441$ and standard deviation $\sigma = 0.2889$ on the right.

For a given finite (signed) measure η concentrated on $[0, \infty)$ we will adopt a definition from [2] and say that an integrable measurable process $(Y_t)_{t \in \mathbb{R}}$ is a solution to the associated Lévy-driven stochastic delay differential equation (SDDE) if it is stationary and satisfies

$$dY_t = \int_{[0, \infty)} Y_{t-v} \eta(dv) dt + dL_t. \quad (2.5)$$

In the formulation of the next result we denote by δ_0 the Dirac measure at 0 and use the convention $\prod_{\emptyset} = 1$ and $\sum_{\emptyset} = 0$. Furthermore, we introduce the finite measure $\eta_{\beta} = \mathbb{1}_{[0,\infty)}(v)e^{\beta v} dv$ for $\beta \in \mathbb{C}$ with $\operatorname{Re}(\beta) < 0$, and let $\eta_0 = \delta_0$ and $\eta_j = \eta_{j-1} * \eta_{\beta_j}$ for $j = 1, \dots, p-1$. By relying on [2, Theorem 3.12] we get the following dynamic SDDE representation of an invertible CARMA($p, p-1$) process:

Theorem 1. *Let $(Y_t)_{t \in \mathbb{R}}$ be an invertible CARMA($p, p-1$) process and let $\beta_1, \dots, \beta_{p-1}$ be the roots of Q . Then $(Y_t)_{t \in \mathbb{R}}$ is the (up to modification) unique stationary solution to (2.5) with the real-valued measure η given by*

$$\eta = \sum_{j=0}^{p-1} \alpha_j \eta_j, \quad (2.6)$$

where $\alpha_0, \dots, \alpha_{p-1} \in \mathbb{C}$ are chosen such that the relation

$$P(z) = z \prod_{k=1}^{p-1} (z - \beta_k) - \sum_{j=0}^{p-1} \alpha_j \prod_{k=j+1}^{p-1} (z - \beta_k) \quad (2.7)$$

holds for all $z \in \mathbb{C}$. In particular, if $\beta_1, \dots, \beta_{p-1}$ are distinct,

$$\eta(dv) = \gamma_0 \delta_0(dv) + \left(\mathbb{1}_{[0,\infty)}(v) \sum_{i=1}^{p-1} \gamma_i e^{\beta_i v} \right) dv \quad (2.8)$$

where

$$\gamma_0 = -\left(a_1 + \sum_{j=1}^{p-1} \beta_j\right) \quad \text{and} \quad \gamma_i = -\frac{P(\beta_i)}{Q'(\beta_i)} \quad \text{for } i = 1, \dots, p-1.$$

Proof. It follows immediately from [2, Theorem 3.12] that $(Y_t)_{t \in \mathbb{R}}$ is the unique stationary solution to (2.5) with η given by (2.6). Assume now that the roots of Q are distinct. Then relation (2.7) implies in particular that $\gamma_0 = \alpha_0 = -(a_1 + \sum_{j=1}^{p-1} \beta_j)$. Moreover, an induction argument shows that

$$\eta_j(dv) = \mathbb{1}_{[0,\infty)}(v) \sum_{i=1}^j e^{\beta_i v} \prod_{k=1, k \neq i}^j (\beta_i - \beta_k)^{-1} dv,$$

from which it follows that

$$\begin{aligned} \eta(dv) - \alpha_0 \delta_0(dv) &= \sum_{j=1}^{p-1} \alpha_j \left(\mathbb{1}_{[0,\infty)}(v) \sum_{i=1}^j e^{\beta_i v} \prod_{k=1, k \neq i}^j (\beta_i - \beta_k)^{-1} dv \right) \\ &= \mathbb{1}_{[0,\infty)}(v) \sum_{i=1}^{p-1} e^{\beta_i v} \sum_{j=i}^{p-1} \alpha_j \prod_{k=1, k \neq i}^j (\beta_i - \beta_k)^{-1} dv. \end{aligned}$$

Finally, observe that the definition of $\alpha_0, \alpha_1, \dots, \alpha_{p-1}$ implies that

$$\gamma_i = \frac{\sum_{j=i}^{p-1} \alpha_j \prod_{k=j+1}^{p-1} (\beta_i - \beta_k)}{\prod_{k=1, k \neq i}^{p-1} (\beta_i - \beta_k)} = \sum_{j=i}^{p-1} \alpha_j \prod_{k=1, k \neq i}^j (\beta_i - \beta_k)^{-1}, \quad i = 1, \dots, p-1,$$

which concludes the proof.

Remark 1. In Brockwell et al. [7] they assume that the roots of P are distinct. This makes it possible to write $(Y_t)_{t \in \mathbb{R}}$ as a sum of dependent Ornstein-Uhlenbeck processes, which they in turn use to recover the driving Lévy process. In Theorem 1 above we invert the CARMA process by using that it is a solution to an SDDE and thereby circumvent the assumption of distinct roots. On the other hand, when $q \geq 2$, the roots of Q may complex-valued and this would make an estimation procedure that is parametrized by these roots (such as the one given in Section 3) more complicated in practice.

Theorem 1 gives an insightful intuition about inverting CARMA processes as well. Let \mathcal{F} be the Fourier transform where $\mathcal{F}[f](y) = \int_{\mathbb{R}} e^{iyx} f(x) dx$ for $f \in L^1$. If we then heuristically take this Fourier transform on both sides of (2.4) we get

$$P(-iy)\mathcal{F}[Y](y) = Q(-iy)\mathcal{F}[DL](y).$$

For $\gamma_0 \in \mathbb{R}$, this can be rewritten as

$$\mathcal{F}[DL](y) = \left(\frac{P(-iy) + (iy + \gamma_0)Q(-iy)}{Q(-iy)} - \gamma_0 \right) \mathcal{F}[Y](y) + \mathcal{F}[DY](y).$$

If we let $\gamma_0 = -(a_1 + \sum_{j=1}^{p-1} \beta_j)$ then

$$y \mapsto \frac{P(-iy) + (iy + \gamma_0)Q(-iy)}{Q(-iy)} \in L^2,$$

and consequently, there exists $f \in L^2$ such that

$$\left(\frac{P(-iy) + (iy + \gamma_0)Q(-iy)}{Q(-iy)} - \gamma_0 \right) \mathcal{F}[Y](y) = \mathcal{F}[-f * Y - \gamma_0 Y](y).$$

We conclude that $(Y_t)_{t \in \mathbb{R}}$ satisfy the formal equation $DY_t = f * Y_t + \gamma_0 Y_t + DL_t$. By integrating this equation we get an equation of the form (2.5), and in the case where Q has distinct roots, contour integration and Cauchy's residue theorem imply that

$$f(v) = \mathbf{1}_{[0, \infty)}(v) \sum_{i=1}^{p-1} -\frac{P(\beta_i)}{Q'(\beta_i)} e^{\beta_i v}$$

in line with Theorem 1.

The simplest example beyond the (Lévy-driven) Ornstein-Uhlenbeck process is the invertible CARMA(2, 1) process:

Example 1. Suppose that $a_0, a_1 \in \mathbb{R}$ are chosen such that the zeroes of $P(z) = z^2 + a_1 z + a_2$ have negative real parts and let $b_0 > 0$ so that the same holds for $Q(z) = b_0 + z$. Then there exists an associated invertible CARMA(2, 1) process $(Y_t)_{t \in \mathbb{R}}$, and Theorem 1 implies that

$$dY_t = \alpha_0 Y_t dt + \alpha_1 \int_0^\infty e^{\beta_1 v} Y_{t-v} dv dt + dL_t,$$

where $\beta_1 = -b_0$, $\alpha_0 = b_0 - a_1$, and $\alpha_1 = (a_1 - b_0)b_0 - a_2$. Note that, in this particular case, we have $\gamma_0 = \alpha_0$ and $\gamma_1 = \alpha_1$.

We end this section by giving the mean and the autocovariance function of the invertible CARMA($p, p-1$) process. To formulate the result we introduce the $p \times p$ -matrix

$$A_p = \begin{bmatrix} \beta_1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & \beta_2 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \beta_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \beta_{p-1} & 0 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_{p-2} & \alpha_{p-1} & \alpha_0 \end{bmatrix}, \quad (2.9)$$

where $\alpha_0, \alpha_1, \beta_1, \dots, \alpha_{p-1}, \beta_{p-1} \in \mathbb{C}$ are given as in Theorem 1. In case $p = 1$, respectively $p = 2$, the matrix in (2.9) reduces to $A_1 = \alpha_0$, respectively

$$A_2 = \begin{bmatrix} \beta_1 & 1 \\ \alpha_1 & \alpha_0 \end{bmatrix}.$$

Proposition 1. *Let $(Y_t)_{t \in \mathbb{R}}$ be an invertible CARMA($p, p-1$) process and let η be the associated measure introduced in Theorem 1. Then*

$$\mathbb{E}[Y_t] = -\frac{\mu}{\eta([0, \infty))} \quad \text{and} \quad \gamma(t) := \text{Cov}(Y_t, Y_0) = \sigma^2 e_p' e^{A_p |t|} \Sigma e_p, \quad t \in \mathbb{R},$$

where

$$\mu = \mathbb{E}[L_1], \quad \sigma^2 = \text{Var}(L_1), \quad \text{and} \quad \Sigma = \int_0^\infty e^{A_p y} e_p e_p' e^{A_p' y} dy.$$

In particular, $(Y_t)_{t \in \mathbb{R}}$ is centered if and only if $(L_t)_{t \in \mathbb{R}}$ is centered.

Proof. The mean of Y_t is obtained from (2.5) using the stationarity of $(Y_t)_{t \in \mathbb{R}}$. The autocovariance of $(Y_t)_{t \in \mathbb{R}}$ function is given in [2, p. 14].

3 Estimation of the SDDE parameters

Fix $\Delta > 0$ and $n \in \mathbb{N}$, and suppose that we have $n+1$ equidistant observations $Y_0, Y_\Delta, Y_{2\Delta}, \dots, Y_{n\Delta}$ of an invertible CARMA($p, p-1$) process $(Y_t)_{t \in \mathbb{R}}$. Our interest will be on estimating the vector of parameters

$$\theta_0 = (\alpha_0, \alpha_1, \beta_1, \alpha_2, \beta_2, \dots, \alpha_{p-1}, \beta_{p-1})'$$

of η in (2.6). We will restrict our attention to the case where $\theta_0 \in \mathbb{R}^{2p-1}$. For simplicity, we will also assume that $(Y_t)_{t \in \mathbb{R}}$ or, equivalently, $(L_t)_{t \in \mathbb{R}}$ is centered. For any given θ let $P_{k-1}(Y_{k\Delta} \mid \theta)$ be the $L^2(\mathbb{P}_\theta)$ -projection of $Y_{k\Delta}$ onto the linear span of $Y_0, Y_\Delta, Y_{2\Delta}, \dots, Y_{(k-1)\Delta}$ and set $\epsilon_k(\theta) = Y_{k\Delta} - P_{k-1}(Y_{k\Delta} \mid \theta)$. Then the least squares estimator $\hat{\theta}_n$ of θ_0 is the point that minimizes

$$\theta \mapsto \sum_{k=1}^n \epsilon_k(\theta)^2.$$

In practice, the projections $P_{k-1}(Y_{k\Delta} \mid \theta)$, $k = 1, \dots, n$, can be computed using the Kalman recursions (see, e.g., [6, Proposition 12.2.2]) together with the state-space representation given in Proposition 2 below. We stress that one can compute the projections without a state-space representation, e.g., using the Durbin-Levinson algorithm (see [6, p. 169]), but this approach will be very time-consuming for large n and a cut-off is necessary in practice. (This technique is used by [12] in the SDDE framework (2.5) when η is compactly supported and $(L_t)_{t \in \mathbb{R}}$ is a Brownian motion.) Under weak regularity assumptions, following the arguments in [7, Proposition 4-5] that rely on [10], one can show that the estimator $\hat{\theta}_n$ of θ_0 is (strongly) consistent and asymptotically normal.

Proposition 2 provides a convenient state-space representation of $(Y_{k\Delta})_{k \in \mathbb{N}_0}$ in terms of $\alpha_0, \alpha_1, \beta_1, \dots, \alpha_{p-1}, \beta_{p-1}$ (rather than the one from the definition of $(Y_t)_{t \in \mathbb{R}}$ in terms of the coefficients of P and Q).

Proposition 2. *Let the setup be as above and let A_p be the matrix given in (2.9). Then $(Y_{k\Delta})_{k \in \mathbb{N}_0}$ has the state-space representation $Y_{k\Delta} = e'_p Z_k$, $k \in \mathbb{N}_0$, with $(Z_k)_{k \in \mathbb{N}_0}$ satisfying the state-equation*

$$Z_k = e^{A_p \Delta} Z_{k-1} + U_k, \quad k \in \mathbb{N},$$

where $(U_k)_{k \in \mathbb{N}}$ is a sequence of i.i.d. random vectors with mean 0 and covariance matrix $\int_0^\Delta e^{A_p u} e_p e'_p e^{A_p' u} du$.

Proof. It follows by [2, Proposition 3.13] that $Y_t = e'_p \tilde{Z}_t$, $t \in \mathbb{R}$, where $(\tilde{Z}_t)_{t \in \mathbb{R}}$ is the \mathbb{R}^p -valued Ornstein-Uhlenbeck process given by

$$\tilde{Z}_t = \int_{-\infty}^t e^{A_p(t-u)} e_p dL_u$$

for $t \in \mathbb{R}$. Thus, by defining $Z_k = \tilde{Z}_{k\Delta}$ so that $Y_{k\Delta} = e'_p Z_k$, $k \in \mathbb{N}_0$, and observing that

$$Z_k = \int_{-\infty}^{(k-1)\Delta} e^{A_p(k\Delta-u)} e_p dL_u + \int_{(k-1)\Delta}^{k\Delta} e^{A_p(k\Delta-u)} e_p dL_u = e^{A_p \Delta} Z_{k-1} + U_k$$

with $U_k := \int_{(k-1)\Delta}^{k\Delta} e^{A_p(k\Delta-u)} e_p dL_u$ for $k \in \mathbb{N}$, the result is immediate.

4 A simulation study, $p = 2$

The simulation of the invertible CARMA(2, 1) is done in a straightforward manner by the (defining) state-space representation of $(Y_t)_{t \in \mathbb{R}}$ and an Euler discretization of (2.2). In order to ensure that X_0 is a realization of the stationary distribution we take 20,000 steps of size 0.01 before time 0. Given X_0 the simulation is based on 200,000 steps each of size 0.01, and then it is assumed that we have $n + 1 = 2,000$, respectively $n + 1 = 20,000$, observations of the process $Y_0, Y_\Delta, Y_{2\Delta}, \dots, Y_{(n-1)\Delta}$ on a grid with distance $\Delta = 1$, respectively $\Delta = 0.1$, between adjacent points. We will be considering the case where the background noise $(L_t)_{t \in \mathbb{R}}$ is a gamma (Lévy) process with shape parameter $\lambda > 0$ and scale parameter $\xi > 0$. Recall that the gamma process with parameters λ and ξ is a pure jump process with infinite activity, and the density f (at time 1) is given by

$$f(x) = \frac{1}{\Gamma(\lambda)\xi^\lambda} x^{\lambda-1} e^{-\frac{x}{\xi}}, \quad x > 0,$$

where Γ is the gamma function. In line with [7] we will choose the parameters to be $\lambda = 0.2488$ and $\xi = 0.5792$. For comparison we will also study the situation where $(L_t)_{t \in \mathbb{R}}$ is Brownian motion with mean $\mu = \lambda\xi = 0.1441$ and standard deviation $\sigma = \xi\sqrt{\lambda} = 0.2889$ (these parameters are chosen so that the Brownian motion matches the mean and standard deviation of the gamma process). After subtracting the sample mean $\bar{Y}_n = n^{-1} \sum_{k=0}^{n-1} Y_{k\Delta}$ from the observations, the vector of true parameters $\theta_0 = (\alpha_0, \alpha_1, \beta_1)$ is estimated as outlined in Section 3. We will choose $\theta_0 = (-1.1558, 0.1939, -0.2061)$ as in [7] (this choice corresponds to $a_1 = 1.3619$, $a_2 = 0.0443$, and $b_0 = 0.2061$, which are certain estimated values of a stochastic volatility model by [16]). We repeat the experiment 100 times and the estimated parameters are given in Table 1.

It appears that the (absolute value of the) bias of $(\alpha_0, \alpha_1, \beta_1)$ is very small when $\Delta = 0.1$. The general picture is that the bias is largest for α_0 , and it is also consistently negative. This observations should, however, be seen in light of the relative size of α_0 compared to α_1 and β_1 .

Once we have estimated θ_0 we can estimate the driving Lévy process by exploiting the relation presented in Theorem 1 and using the trapezoidal rule. Note that, as in the estimation, we use the relation in Theorem 1 on the demeaned data so that we in turn recover the centered version of the Lévy process. Finally, to obtain an estimate of the true Lévy process we estimate $\mu = \mathbb{E}[L_1]$ using Proposition 1. In order to get a proper approximation of the integral $\int_0^\infty e^{\beta_1 v} (Y_{t-v} - \mathbb{E}_{\theta_0}[Y_0]) dv$ we will only be estimating $L_{k\Delta} - L_{(k-1)\Delta}$ for $m := 50\Delta^{-1} \leq k \leq n$. If one is interested in estimating the entire path $L_{(m+1)\Delta} - L_{m\Delta}, L_{(m+2)\Delta} - L_{m\Delta}, \dots, L_{n\Delta} - L_{m\Delta}$, one will need data observed at a high frequency, that is, small Δ , since the approximation errors accumulate over time. Typically, one is more interested in estimating the distribution of L_1 , which is less sensitive to these approximation errors, and this is our focus in the following. For this reason, we have in Figure 2 plotted five estimations of the

Noise	Spacing	Parameter	Sample mean	Bias std	Sample deviation
Gamma	$\Delta = 1$	α_0	-1.2075	-0.0517	0.1155
		α_1	0.2157	0.0218	0.0501
		β_1	-0.2190	-0.0129	0.0366
	$\Delta = 0.1$	α_0	-1.1688	-0.0130	0.0466
		α_1	0.1934	-0.0005	0.0315
		β_1	-0.2053	0.0008	0.0296
Gaussian	$\Delta = 1$	α_0	-1.1967	-0.0409	0.1147
		α_1	0.2117	0.0178	0.0524
		β_1	-0.2201	-0.0140	0.0358
	$\Delta = 0.1$	α_0	-1.1653	-0.0095	0.0469
		α_1	0.2002	0.0062	0.0353
		β_1	-0.2121	-0.0060	0.0324

Table 1. Estimated SDDE parameters based on 100 simulations of the CARMA(2, 1) process on $[0, 2000]$ with true parameters $\alpha_0 = -1.1558$, $\alpha_1 = 0.1939$, and $\beta_1 = -0.2061$.

distribution function of L_1 in dashed lines against the true distribution function (represented by a solid line) in the low frequency case ($\Delta = 1$). The left, respectively right, figure corresponds to the gamma, respectively Gaussian, case. Due to the above conventions, each estimated distribution function is based on 1,950 estimated realizations of L_1 . Generally, the estimated distribution functions in the figures seem to capture the true structure and give a fairly precise estimate, however, there is a slight tendency to over-estimate small values and under-estimate large values.

Due to the high degree of precision of the estimated distribution functions, we plot an associated histogram, based on 1,950 realizations of L_1 and a sampling frequency of $\Delta = 1$, against the theoretical probability density function in order to detect potential (smaller) biases. We compare this to a histogram of the actual increments. For simplicity, we have restricted ourselves to the Gaussian case as the gamma case is difficult to analyze close to zero (specifically, this will require more observations). The plots are found in Figure 3. We see that the two histograms have very similar appearances, but the histogram based on estimated parameters has a slightly smaller mean.

5 Conclusion and future research

In this paper we have studied the ability to recover the underlying Lévy process from an observed invertible CARMA process using the SDDE relation presented in Theorem 1. In particular, after discussing the theoretical foundations, we did a simulation study similar to the one in the classical approach presented in [7] and estimated the underlying Lévy noise. Our findings supported the theory and it seemed possible to (visually) detect the distribution of the underlying Lévy process.

Future research could include a further study of the performance of the presented SDDE inversion technique compared to the classical approach in [7]. Specifically, in light of Remark 1, a suggestion could be to consider a situation where P has a root of multiplicity strictly greater than one or where $q \geq 2$ and some of the roots of Q are not real numbers. Such situations may complicate the analysis in one approach relative to the other. Furthermore, it may be interesting to study inversion formulas for invertible CARMA(p, q) processes when $p > q + 1$. In particular, a manipulation of the equation (2.4) yields

$$dL_t = \left(\frac{P(D)}{Q(D)} Y_t \right) dt. \quad (5.1)$$

The content of Theorem 1 is that the right-hand side of (5.1) is meaningful when $p = q + 1$ and it should be interpreted as $dY_t - \int_{[0, \infty)} Y_{t-v} \eta(dv) dt$. It seems that this statement continues to hold when $p > q + 1$ as well when dY_t is replaced by a suitable linear combination of $dY_t, dY_t^{(1)}, \dots, dY_t^{(p-q-1)}$.

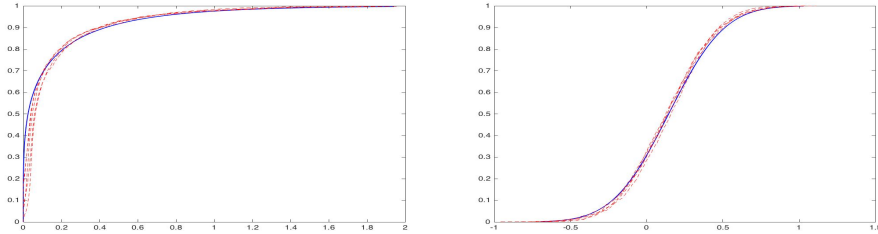


Figure 2. Five estimations of the distribution function of L_1 , based on estimates of α_0 , α_1 , and β_1 , plotted against the true distribution function for a sampling frequency of $\Delta = 1$. The left corresponds to gamma noise and the right to Gaussian noise.

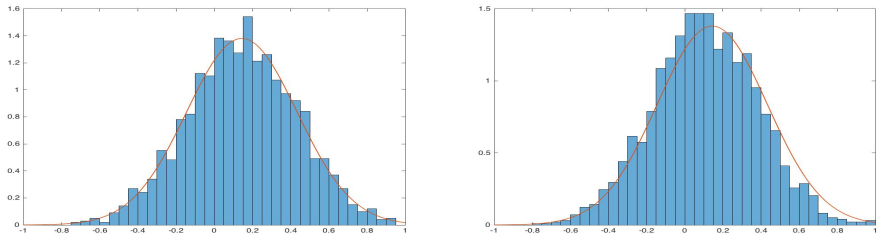


Figure 3. Histograms of the true increments on the left and estimated increments, based on estimates of α_0 , α_1 , and β_1 for a sampling frequency of $\Delta = 1$, on the right plotted against the theoretical (Gaussian) probability density function.

Bibliography

- [1] Barndorff-Nielsen, O. E. and N. Shephard (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 63(2), 167–241.
- [2] Basse-O'Connor, A., M. S. Nielsen, J. Pedersen, and V. Rohde (2017). A continuous-time framework for ARMA processes. *arXiv preprint arXiv:1704.08574*.
- [3] Benth, F., C. Klüppelberg, G. Müller, and L. Vos (2011). Futures pricing in electricity markets based on stable carma spot models, submitted manuscript. *arXiv preprint arXiv:1201.1151*.
- [4] Benth, F. E., J. v. S. Benth, and S. Koekebakker (2007). Putting a price on temperature. *Scand. J. Statist.* 34(4), 746–767.
- [5] Brockwell, P. J. (2001). Lévy-driven CARMA processes. *Ann. Inst. Statist. Math.* 53(1), 113–124. Nonlinear non-Gaussian models and related filtering methods (Tokyo, 2000).
- [6] Brockwell, P. J. and R. A. Davis (2006). *Time series: theory and methods*. Springer Series in Statistics. Springer, New York. Reprint of the second (1991) edition.
- [7] Brockwell, P. J., R. A. Davis, and Y. Yang (2011). Estimation for non-negative Lévy-driven CARMA processes. *J. Bus. Econom. Statist.* 29(2), 250–259.
- [8] Brockwell, P. J., V. Ferrazzano, and C. Klüppelberg (2013). High-frequency sampling and kernel estimation for continuous-time moving average processes. *J. Time Series Anal.* 34(3), 385–404.
- [9] Brockwell, P. J. and A. Lindner (2009). Existence and uniqueness of stationary Lévy-driven CARMA processes. *Stochastic Process. Appl.* 119(8), 2660–2681.
- [10] Francq, C. and J.-M. Zakoïan (1998). Estimating linear representations of nonlinear processes. *J. Statist. Plann. Inference* 68(1), 145–165.
- [11] García, I., C. Klüppelberg, and G. Müller (2011). Estimation of stable carma models with an application to electricity spot prices. *Statistical Modelling* 11(5), 447–470.
- [12] Küchler, U., M. Sørensen, et al. (2013). Statistical inference for discrete-time samples from affine stochastic delay differential equations. *Bernoulli* 19(2), 409–425.
- [13] Marquardt, T. and R. Stelzer (2007). Multivariate CARMA processes. *Stochastic Process. Appl.* 117(1), 96–120.
- [14] Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of Finance* 52(3), 923–973.
- [15] Todorov, V. (2009). Estimation of continuous-time stochastic volatility models with jumps using high-frequency data. *Journal of Econometrics* 148(2), 131–148.

- [16] Todorov, V. (2011). Econometric analysis of jump-driven stochastic volatility models. *J. Econometrics* 160(1), 12–21.
- [17] Todorov, V. and G. Tauchen (2006). Simulation methods for Lévy-driven continuous-time autoregressive moving average (CARMA) stochastic volatility models. *J. Bus. Econom. Statist.* 24(4), 455–469.
- [18] Whitney, H. (1972). *Complex analytic varieties*. Addison-Wesley Pub. Co.

Fuel Consumption Estimation for Climbing Phase

JingJie Chen¹, YongPing Zhang²

Civil Aviation University of China, Tianjin, China jjchen@cauc.edu.cn

Abstract. Aiming at the problem of the civil aviation carbon emission, the purpose of this paper is to pre-sent a simplified method to estimate aircraft fuel consumption using an adaptive Genetic Algorithm-Back Propagation (GA-BP) Strong prediction network.

This paper gives a brief overview of the modelling approach, and describes efforts to validate and analyze the initial results of this project. The parameters of fuel consumption are analyzed by using QAR flight data, and two kinds of fuel consumption prediction model are proposed. It is the BP prediction model and the adaptive GA-BP (Genetic Algorithm-Back Propagation) Strong prediction model. The crossover and mutation probability of GA-BP Strong prediction model can be adaptive adjustment network parameters, according to the characteristics of data. The BP neural network as a weak predictor, after the limited number of iterations, it can realize error optimization adjustment and solve the complicated nonlinear problem. Results of the simulation indicated the two models have obvious advantages in nonlinear prediction, and the prediction accuracy and the degree of fitting are good. The results of this study illustrate that the two neural networks, it with nonlinear transfer functions can accurately represent complex aircraft fuel consumption functions for climb phases of flight, and so the two models are feasible in the field of fuel consumption prediction. The methodology can be extended to cruise and descent phases of flight.

Keywords: Flight data, Adaptive GA-BP-Adaboost network, Fuel consumption, Prediction

1 INTRODUCTION

Air Transport industry acts as a catalyst to the economic and social development of a nation. But the development of air transport industry is faced with major issues like high fuel consumption [1 2]. Furthermore, according to Henderson RP (2012) research that aircraft fuel burn is proportional to CO₂ emission [3]. From the perspective of civil aviation, reducing fuel oil can not only reduce operating costs, but also reduce carbon emissions and ease the pressure on the environment [4]. Consequently, to keep sustainable and stable development, accurate forecast of energy consumption is essential, and the development of rational forecast model especially is urgent and necessary [5]. For this challenge, many researchers have studied about different models on aircraft fuel flow prediction. Thus, to obtain rational forecast results, various prediction models are put out. Such as Chang R C (2014) presents a fuzzy logic system [6], Zhang HF (2015) presents a support Vector Regression (SVR) model [7], Cavcar Aydan analyzed the influence of aircraft performance difference on fuel consumption in air traffic environment, which reflected the influence of air traffic control on flight fuel economy [8]. Based on the principle of energy conservation, a model was established to estimate the fuel consumption by Bella P [9]. Based on the flight process, Ralf H presented an exponential relationship model which is to establish the relationship between the fuel flow and the height in the fall and climb phases [10]. Based on the genetic algorithm, the Baklacioglu Tolga researched a fuel consumption model which used to study the change of the fuel with the air speed and altitude at different time [11]. Bak-

This work was supported by the Supported by the Energy saving and emission reduction projects of CAAC of China under Grant No. DPDSR0010.

This work was supported by the Supported by the National Key Technology R&D Program of China under Grant No. 2012BAC20B03.

J. J. Chen is a professor, Civil Aviation University of China, Tianjin, 300300, China (e-mail: 94238045@qq.com).

Y. Y. Zhang is a Graduate student, Civil Aviation University of China, Tianjin, 300300, China (e-mail: 2468134855@qq.com).

lacioglu.T used a thrust model to optimize the flight trajectory [12], Matthias Bartel also proposed the relationship between thrust and fuel consumption model to achieve fuel saving purposes [13]. However, the mentioned above have some limitations for the study of fuel consumption. Most of them were based on the level of time series with a single flight as the research object to analyze the characteristics of flight, and then built a model to predict the fuel consumption at different times in the flight process. It is lack of practical application.

Using neural network, complex nonlinear function can be easily handled, and the network has an advantage that there is no need to reveal the mathematical equation describing the input-output mapping before the network training. The output of the network function is close to the actual output value, so as to achieve a more accurate prediction effect. The genetic algorithm can realize global optimization and optimize the initial state of the network. Back propagation neural network prediction model and a back propagation based on genetic algorithm (GA-BP) neural network prediction model have many successful applications [14-15], in areas such as prediction of electricity demand and stock price etc[16-17]. This is mainly because they have very good approximation capabilities and self-learning and adaptive ability.

Based on this premise, this paper attempts to develop a suitable method to estimate aircraft fuel consumption using neural network approach to deal with relation between energy consumption and its influence factors. To study single aircraft fuel consumption, it can forecast the total fuel consumption of fixed flight and the fuel flow. Harshad Khadilkar (2012) demonstrates that flight data record the main parameters used for the analysis of fuel consumption [18]. In the study, through the analysis of QAR data, we select the main factors that affect the fuel consumption, but only for the climbing condition.

An adaptive GA-BP neural network fuel consumption forecasting model is established, and a strong prediction is joined in the model. It will enable more accurate climb fuel predictions. In this paper, the fitting degree of the model output, the average value of the relative error and the mean square error are chosen the benchmark of the feasibility of the model. Furthermore, as an improvement to the existing models is compared with the BP neural network prediction model, and the simulation results are analyzed. It is proved that these two models have good prediction effect.

2 FUEL CONSUMPTION INFLUENCING FACTORS

Flight data QAR records the most flight parameters of the aircraft from takeoff to landing, these parameters can reflect the impact of engine performance degradation on fuel consumption, which provides a good basis for the analysis of fuel consumption [19]. So, in this paper, the QAR data is selected as the data set of model training and prediction.

Based on those flight control parameters, a prediction model can be made to estimate fuel loading. There is required to do data filtering. Sometimes, when data is missed out, interpolation is also required if the filtering frequency is fixed. Similarly, in this paper, the idea of stepwise linear regression is used to screen and eliminate the multiple co linear variables in the QAR data [20]. We can screen effect parameters of aircraft climb phase. In this study, we take the initial weight of climbing segment, the climbing distance, the rate of climb, and the force of the wind in the nose and the total temperature of the atmosphere as the input of the model 5 factors. We use the direct factor of fuel consumption and analyze the contribution of the minor changes of each factor to fuel consumption. The outputs of the model are relative easy to determine according to our modeling objective. They are fuel consumption.

After that, the model input data are normalized to the same dimension. We define x_{\max} and x_{\min} as the maximum and minimum value in the sample and x_k as the sample normalization value. And the function can be written as

$$X_k = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

3 Improved Model Algorithm

3.1 Basic Procedures and Ideas.

We propose a genetic optimization BP neural network. The whole idea of network structure is using the improved adaptive genetic algorithm to obtain the optimal individual, and used to optimize the initial weights and threshold of the neural network. After training the BP network learning to get the fuel consumption forecast output, in order to improve the prediction accuracy, we BP network as a weak predictor, after a finite number of iterations and outputs the result of strong prediction results. Structure diagram as shown in Figure 1.

3.2 GA-BP-Adaboost Parameters Setup

In this section, we propose a fuel consumption prediction model based on GA-BP-Adaboost neural network. We take the 280 sets of QAR data from different flights in the same course, and take them as the training data set and test data set of the fuel consumption model. The input is the starting weight of aircraft, the climbing distance, the rate of climb, the wind speed, the total temperature of the atmosphere, set X as input vector, $X=[x_1, x_2, x_3, x_4, x_5]$, Aircraft fuel flow as the output, set Y as the output vector. Genetic algorithm for optimizing network weight needs to design the main parameters.

Step 1. An initial population is generated. A population of W individuals were randomly generated. W is chosen 30.

Step 2. Coding for neural network. In order to get the high precision weight and threshold value, the real number coding method is adopted. Individual coding length is L . There L is equal to $N * M + M + M * K + K$, and the Input node number N take 5, the number of hidden nodes M take 10, the number of output nodes K take 1.

Step 3. Fitness Function. Training overall error is as small as possible and genetic algorithm uses the minimum objective function value as the fitness function. We calculate square of fuel consumption error sum in the training sample and chromosome adaptive value is

$$f = k \sum_{i=0}^n abs(y_i - o_i) \quad (2)$$

$n = 200$ is the training samples number, o_i is the expected output of BP network, y_i is the forecasted output of BP network, $k = 10$ is the coefficient of function.

Step 4. Selecting operation. Training overall error is as small as possible and genetic algorithm uses the minimum objective function value as the fitness function. We calculate square of components response time error sum in the training sample and chromosome adaptive value is F_i selective probability is P_i

$$F_i = k / f_i \quad (3)$$

$$P_i = \frac{F_i}{\sum_{i=0}^P F_i} \quad (4)$$

P is the reciprocal of fitness, $P = 30$ is population size.

Step 5. Crossover Operator and Crossover Probability. We randomly select two chromosomes and choose their weights and thresholds with crossover probability to form two new individuals a_{ki}, a_{lj} . We use arithmetic crossover operation to generate two new individuals:

$$a_{kj} = a_{ki}(1 - a) + a_{lj}a \quad (5)$$

$$a_{lj} = a_{lj}(1 - a) + a_{ki}a \quad (6)$$

$a \in (0, 1)$ is a parameter. The crossover probability is usually take the value of the $[0.5, 0.98]$.

The crossover probability of this study is obtained by the following formula:

$$P_c = m + \frac{m(f_{av} - f_i)}{10(f_i - f_{\min})}, f_i \geq f_{av} \quad (7)$$

$$P_c = m - \frac{m(f_{av} - f_i)}{10(f_i - f_{\min})}, f_i < f_{av} \quad (8)$$

$m = 0.7$, f_i is the smaller value of the fitness value of the two selected individuals, f_{av} is the average value of all fitness values, f_{\min} is the minimum value in all fitness values.

Step 6. Mutation Operator and Mutation Probability. We randomly select a chromosome, and mutation probability P_m selects weight threshold to achieve mutation. We use arithmetic mutation operation to generate two new individuals:

$$a_{ij} = a_{ij} + (a_{ij} - a_{\max}) * f(g), r \geq 0.5 \quad (9)$$

$$a_{ij} = (a_{ij} + a_{\min} - a_{ij}) * f(g), r < 0.5 \quad (10)$$

$$f(g) = r_2(1 - \frac{g}{G_{\max}}) \quad (11)$$

Mutation point is a_{ij} and the value range is $[a_{\min}, a_{\max}]$. Random number is r_2 the number of iterations is g . The max number of evolution is G_{\max} . The value range of r is $[0, 1]$.

$$P_m = m_1 - \frac{m_1(f'_i + f_{av})}{10(f_i - f_{\min})}, f'_i < f_{av} \quad (12)$$

$$P_m = m_1 - \frac{m_1(f'_i - f_{av})}{10(f_i - f_{\min})}, f'_i > f_{av} \quad (13)$$

$m_1 = 0.04$, f'_i is The value of the smaller adaptation for the current two variants.

Step7. Initialization of BP weights and thresholds. The obtained optimal individual is decomposed into the initialization weight and threshold value of the BP network. In this paper, we use three layers of BP neural network, the input neurons are 5, the hidden neurons is 10.

Step8. Training model. After optimized, the composite model of BP and Adaboost is established and began to start training.

In the 5) and 6) step the formula mentioned, we can know that the crossover probability and muta-

tion probability follow the change of fitness value, which can avoid the divergence of GA algorithm, fall into local minimum and speed up the convergence. In step 8) we can be obtained the optimization of initial weights and thresholds. Then began to train the BP network, if the predicted output error does not meet the conditions, the iterative adjustment of program can reduce the output error, and prediction model precision improved, and the practical application of the prediction is enhanced.

4 Model Validations

4.1 Model feasibility evaluation criteria

In this paper, mean relative error (MRE), sum of relative error absolute values and mean square error of relative error (MSE) were introduced as metrics of the modelling accuracy. Goodness of fit (R) is used to quantify model accuracy. For a data set of n measured outputs O_i and predicted outputs Y_i , Relative error, MRE and MSE is calculated as

$$erro = \frac{O_i - Y_i}{Y_i} \quad (14)$$

$$MRE = \frac{1}{N} \sum_{i=1}^N \frac{O_i - Y_i}{Y_i} \quad (15)$$

$$MSE = \frac{\sum_{i=1}^N \left(\frac{O_i - Y_i}{Y_i} \right)^2}{N} \quad (16)$$

$$R = \frac{\sum_{i=1}^N (O_i - \hat{O}_i)^2}{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2} \quad (17)$$

4.2 Model testing and evaluation

We use Matlab software platform to build model. It is the BP neural network model and the GA-BP-Adaboost model. There were 200 training samples and 80 of the samples were verification samples. The best stability of ELM was analyzed by comparison of [21]. In this case, the optimal sample set is selected by the cross-validation model as the input sample set of the network described below. The parameters of the BP neural network model are set as follows: the number of training is 30, the training target is 0.002, and the learning rate is 0.1. The parameters of the GA-BP-Adaboost neural network model are set as follows: the population size is 30, and the evolutionary algebra is 30.

Neural network training index of output value are as follows: the mean relative error value is 0.0035 and the sum of absolute value of relative error is 2.0225, the relative error of standard deviation is 0.0260, the goodness of fit is 0.9801, the mean square error of the relative error is 6.8235e-04. According to evaluation index shows that the training model good can output prediction.

Figure 2 are the contrast diagram of the two models of the prediction of the output and the error. Figure 3 is the neural network curve fitting, which can show the validity of the prediction function. With the different prediction samples 40, 60, 80, table 1 gives the relative mean error, mean square error and fitting degree of two kinds of prediction models.

Factors affecting fuel consumption is not limited to the factors considered in this paper, also affected by other factors, such as weather conditions, the same flight with different routes and airports and other factors such as route congestion. And considering the security requirements of actual flight, the plane will usually carry 45 minutes' alternate or return flight fuel, which may lead to "Fuel oil consumption" happen. So the prediction value and the expected value of the fuel consumption are allowed to error exist.

It is concluded that adaptive GA-BP-Adaboost prediction model has a small increase in the prediction accuracy and nonlinear fitting ability, fault tolerance capability. Moreover, when the training data set is less, the prediction precision and dynamic quality of the model are still kept. By analyzing the index value of the simulation and evaluation model, and considering the existence of the actual error, it can be draw the conclusion that the results indicate that these two methods used to forecast fuel consumption is feasible, effective and convenient for practical applications.

4.3 Analysis of simulation experiment

With the increase of the genetic algebra, the fitness value is reduced after 20 generations, and the network fitness is optimal. So the weights and thresholds of the model are optimized which can improve the prediction accuracy of the model. In Figure 3, the relative error values of the two prediction models are between [-0.1 0.1].

When the prediction sample number is 80, after 3 iterations the best validation parameters of the improved adaptive GA-BP-Adaboost prediction model is 0.0025504. Similarly, after 4 iterations the best validation parameters of BP model is 0.0035213. Which it indicates that mean square error of GA-BP strong prediction is smaller than BP prediction. So, its output values closer to the predicted target and has a stronger adaptability.

Table 1 gives the average relative error, mean square error and fitting value of two prediction models at different prediction samples, through the table we can analysis the accuracy of the prediction models. In the application of fuel consumption prediction, along with the increase of the number of prediction samples, the two-forecast model can be effective convergence and the predictive accuracy of adaptive GA-BP-Adaboost prediction model increases slightly which is converted to the number level of the actual project can achieve a more significant improvement.

The fuzzy logic model of fuel consumption proposed by Chang R C and the SVR model of fuel consumption proposed by HF Zhang are using the same data as the above experimental data to forecast, the prediction sample number is 80. Results in the following table 2. Analyzing the data in the table 2, it is found that the Improved GA - BP - Adaboost model than other model has a good capability of nonlinear fitting, and the average relative error is small than SVR model and BP model .However, the average relative error of fuzzy logic model proposed by Chang R C is small than the average relative error of Improved GA - BP - Adaboost model, but it's sum of relative error is 2.4319, higher than the Improved GA - BP - Adaboost model error sum 0.5489.it shows that fuzzy logic model is at the cost of model performance to reduce the output error ,so that dynamic stability of fuzzy logic model become lower.

Factors affecting fuel consumption is not limited to the factors considered in this paper, also affected by other factors, such as weather conditions, the same flight with different routes and airports and other factors such as route congestion. And considering the security requirements of actual flight, the plane will usually carry 45 minutes' alternate or return flight fuel, which may lead to "Fuel oil consumption"

happen. So, the prediction value and the expected value of the fuel consumption are allowed to error exist.

It is concluded that adaptive GA-BP-Adaboost prediction model has a small increase in the prediction accuracy and nonlinear fitting ability, fault tolerance capability. Moreover, when the training data set is less, the prediction precision and dynamic quality of the model are still kept. By analyzing the index value of the simulation and evaluation model, and considering the existence of the actual error, it can be draw the conclusion that the results indicate that these two methods used to forecast fuel consumption is feasible, effective and convenient for practical applications.

In the premise of not affect flight safety, it is in order to effectively reduce the fuel consumption, as much as possible to improve the utilization rate of fuel resources. We use the model proposed in this paper to analysis single factor influence on fuel consumption when defining other influence factors.

Prediction of fuel consumption, we can get the single factor variable interval after we get most economical fuel consumption costs. It can provide a reference for the flight plan, so that flight to achieve the best fuel saving flight state.

5 Conclusion

Accurate forecasts of fuel consumption are vital when demand grows faster, it can guide Civil Aviation energy policies effective implementation and reduce flight carbon emissions. Energy consumption forecast is a complex problem due to interactive factors. In this study, we in order to reduce the amount of carbon emissions from flights and improve the fuel utilization rate, by analysing flight data, a fuel consumption forecast model based on GA-BP-Adaboost is presented. The model is verified by flight climbing stage.

By comparing the BP prediction model, the proposed model in the prediction accuracy, nonlinear fitting ability and fault-tolerant ability are increased. And there have small training data set in the study, the model can still maintain prediction accuracy and dynamic performance index. Analyzing the Simulation and considering the actual causes of errors, it can be concluded that two models have practical applications. Also, through the analysis of QAR data, we can aim at the different stages of the same voyage flight to establishment two kinds of prediction models is presented in the paper, they used to predict aircraft in various stages of the fuel consumption. Similarly, in the case of network training well, if the forecast result of test set is found to have a large deviation from the actual value and then analyze the error, we can even determine this flight if there is a fault. Therefore, it is necessary for us to study the prediction model of fuel oil.

In future work, the enhancement of the model presented here is the extension to estimate thrust associated with a fuel burn flight condition parameter such as Thrust Specific Fuel Consumption (TSFC). Preliminary results obtained in research indicate that thrust and TSFC can also be easily characterized using genetic algorithm. To build trajectory predictions of transport aircraft based on time series, this is a good research direction.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. Y Fan, Y Xia (2012) Exploring energy consumption and demand in China. *Energy*, vol 40, pp 23–30
2. C Yuan, S Liu, Z Fang, N Xie (2010) The relation between Chinese economic development and energy consumption in the different periods. *Energy Policy*, vol 38, pp 5189–5198.
3. R. P. Henderson, J. R. R. A. Martins, R. E. Perez(2012) Aircraft conceptual design for optimal environmental performance [J], *Aeronautical Journal*, vol 116,pp 1-22.
4. Chen Jingjie, Xiao Guanping (2014) Analysis of aircraft segment fuel consumption interval estimates the minimum sample size. *Computer Engineering and Design*, vol 12, pp 4356-4359+4364.
5. [5] Y. C. He, K. Liu, X. Y. Shen (2015) Simulation research of the aircraft fuel consumption estimation model [J]. *Computer Simulation*, vol 105, pp 33-36.
6. Ray C. Chang (2015) Examination of excessive fuel consumption for transport jet aircraft based on fuzzy-logic models of flight data [J]. *Fuzzy Sets and Systems*, vol 269, pp 115-134.
7. H. F. Zhang, X. H. Wang, X. F. Chen (2015) Support Vector with ROC Optimization Method Based Fuel Consumption Modeling for Civil Aircraft . *Procedia Engineering*, vol 99, pp 296-300.
8. Cavcar Aydan (2004) Impact of aircraft performance differences on fuel consumption of aircraft in air of management environment. *Aircraft Engineering and Aerospace Technology*, vol 76, pp 502-515.
9. Bella P (1982) Collins Estimation of Aircraft Fuel Consumption, *J AIRCR*,vol 19, pp 969-975.
10. Ralf H, Mayer (2012) Change-oriented aircraft fuel burn and emissions assessment methodologies, *ICNS 2012: Bridging CNS and ATM - Conference Proceedings*, pp N51-N515.
11. Baklacioglu Tolga (2016) Modeling the fuel flow-rate of transport aircraft during flight phases using genetic algorithm-optimized neural networks. *Aerospace Science and Technology*, vol 49, pp 52-62.
12. Baklacioglu, T (2014) Aero-propulsive modelling for climb and descent trajectory prediction of transport aircraft using genetic algorithms .*Aeronautical Journal*, vol 118, pp 65-79.
13. M atthias Bartel, Trevor M Young (2008) Simplified Thrust and Fuel Consumption Models for Modern Two-Shaft Turbofan Engines.*JOURNAL OF AIRCRAFT*, vol 45, pp 1450-1456.
14. T. Baklacioglu (2015) Fuel flow-rate modelling of transport aircraft for the climb flight using genetic algorithms. *Aeronautical Journal*, vol 199, pp 173-183.
15. Ueda Yoshiaki, Horio Keiichi,Kubota Ryo (2014) A modified real-coded genetic algorithm considering with fitness-based variability [J]. *International Journal of Innovative Computing, Information and Control*, vol 10, pp. 1509-1518.
16. J. J. Chen, Y. Yan, J. X Liu (2015) Fuel consumption estimation of cruise phase based on BP network and influence structure analysis. *Computer Measurement & Control*, vol 06, pp 2135-2138.
17. S. Saravanan, S. Karman, Amosedinakaran S, Thangaraj C (2014) India's electricity demand estimation using Genetic Algorithm. *International Conference on Circuits. Power and Computing Technologies*, pp 97-101.
18. Harshad Khadilkar, Hamsa Balakrishnan (2012) Estimation of aircraft taxi fuel burn using flight data. *Transport and Environment*,vol 17, pp 532-537.
19. H. Yang, C. L. Zhao (2014) QAR data fault detection research based on the FP – Tree. *Computer applications and software*, vol 10, pp 41-44.
20. L.X. Liu (2015) Selection of independent variables and the stepwise regression method in the linear regression model. *Statistics and Decision*, vol 21, pp 80-82.

21. QIU J L, HE C. Stability of neural networks of cross validation model [J]. Computer Engineering and Applications | Comput Eng Appl, 2010, (34):43-45(in Chinese).

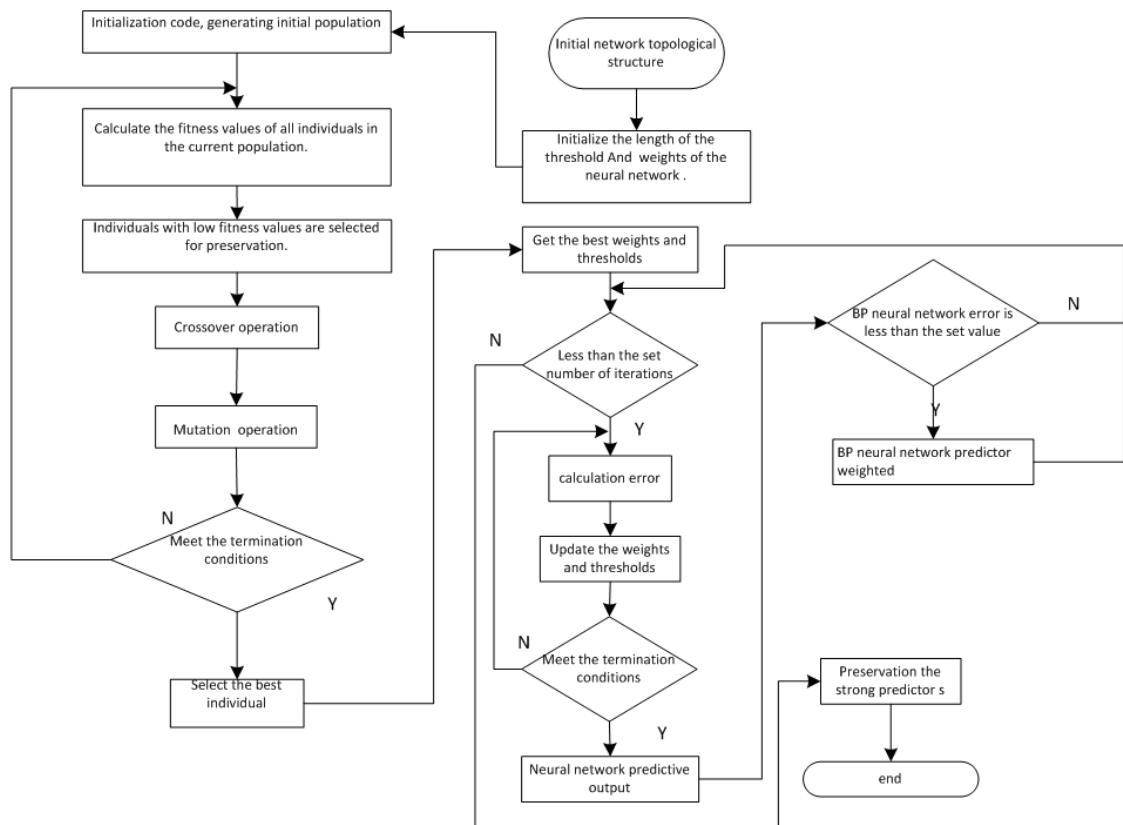


Fig. 1. Based on improved GA-BP strong prediction model

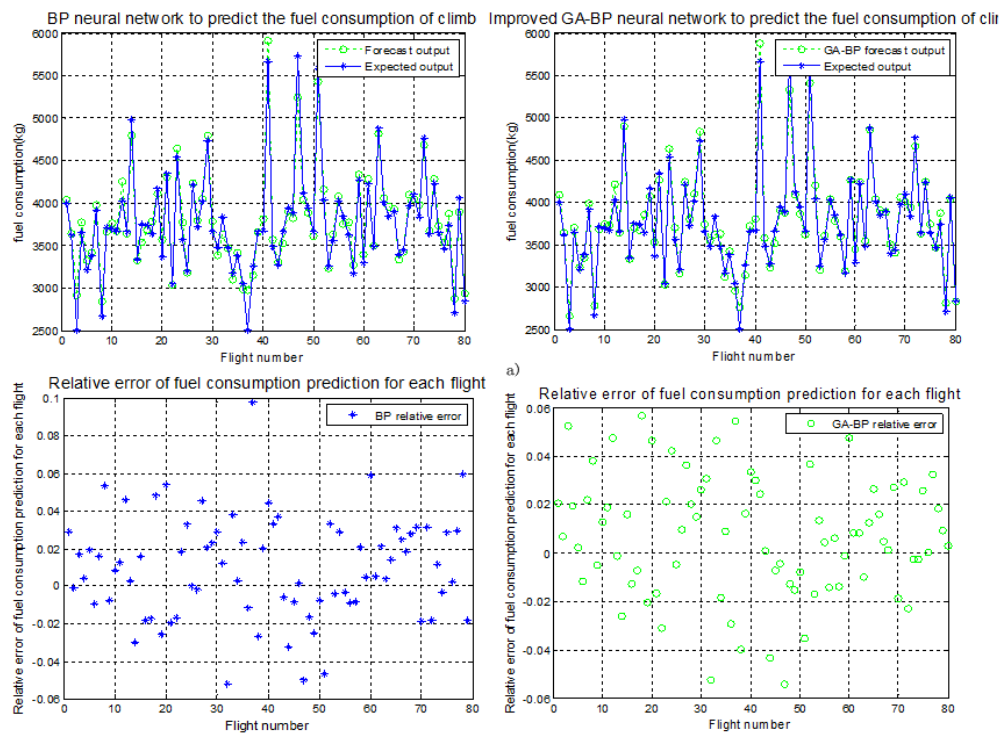


Fig. 2. Output and relative error of model prediction

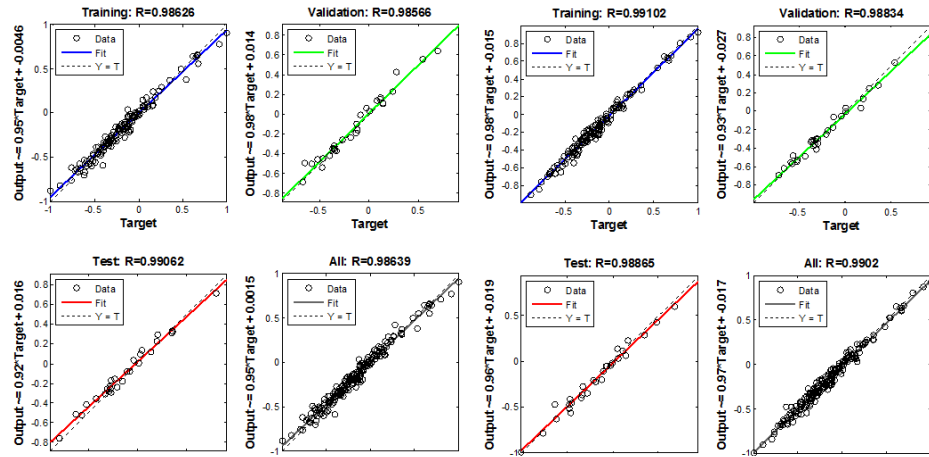


Fig. 3. Analysis of neural network fitting curve

Table 1. Model index value

	BP model			Improved GA-BP-Adaboost model		
	40	60	80	40	60	80
Forecast sample number						
MRE	0.0100	0.0177	0.0127	0.0185	0.0149	0.0119
SRE	1.0125	1.5365	1.8855	1.1026	1.3550	1.8830
MSE	0.0011	0.0011	8.8011e-04	0.0011	7.6700e-04	0.0011
R	0.9558	0.9041	0.9043	0.9347	0.9440	0.9633

- a. MRE-Mean Relative Error.
- b. SRE-Sum of Relative Error Absolute Values.
- c. MSE- Mean Square Error of Relative Error.
- d. R-Goodness of Fit.

Table 2. Model comparison

Forecast sample number	BP model	FUZZY model	SVR model	Improved GA-BP-Adaboost model
80				
MRE	0.0177	0.0065	0.01945	0.0119
R	0.9043	0.9049	0.9070	0.9633

Energy Prediction of Access Points in Wi-Fi Networks Using Time Series Modeling

David Rodriguez-Lozano, Juan A. Gomez-Pulido ^{*}, and
Arturo Duran-Dominguez

School of Technology, University of Extremadura,
Campus Universitario s/n, 10003 Caceres, Spain
{drlozano, jangomez, arduran}@unex.es
<http://arco.unex.es/jangomez>

Abstract. One of the most important elements in Wi-Fi networks is the access point. The number of sessions and data traffic established in the access points has a direct impact on the energy consumption, which can be considered as a certain measure of the users' behavior. Moreover, the energy is a parameter to be taken into account when we plan many maintenance tasks of the network infrastructure. Therefore, knowing the energy consumption in a determined access point in advance can be useful to make decisions about the maintenance works. In this work, we present an energy prediction methodology based on system identification applied to time series. Ten time series model the energy patterns of the access points of a Wi-Fi network in an academic environment during five weeks. The identified models of these series were used to predict next energy consumption in the access points, with reasonably good results.

Keywords: Wi-Fi networks; access point; energy consumption; time series; system identification; prediction.

1 Introduction

The Access Point (AP) is a device that supports the data traffic and the sessions requests in a Wi-Fi infrastructure. The energy use in the AP comes mainly from the demand of network access by users, although there are other factors with energy impact, such as usual operations, location, physical characteristics, etc.

The energy levels in the APs are optimization objectives in many research works [1] [2], and they give us useful information about the users' behavior. From this knowledge, we can plan several maintenance tasks of the network infrastructure, with regard to the deployment, device replacing, etc. In this sense, we should know the energy impact of any maintenance task before doing it, predicting the energy in the APs according to the past energy patterns.

In this work we predict the energy in the APs from time series modeling, using a three-step methodology. First, we collect data of the energy from the

^{*} Corresponding author.

network usage during a determined time period where the users had a regular activity. Next, we build as many time series as access points the network has, where each series draws the daily energy level in the corresponding AP. Next, we model the time series applying system identification: auto-regressive modeling and recursive least squares identification. Finally, the models obtained are used to predict the next energy data.

We have not found works about predicting energy in the APs using time series, but other focus with regard to predict the density of neighbouring APs and the corresponding data traffic [3], or the data quota [4], for example. Other aspects of the wireless networks, as users' location [5], data traffic [6], mobility [7] and [8] application workloads were studied for prediction purposes.

2 System Identification of Time Series

A Time Series (TS) is a signal $y(k)$ sampled by a period T that describes the behaviour of a dynamic system. System Identification (SI) [9] tries to find a parametric mathematical model of the TS from the measures of $y(k)$.

The parametric polynomial description is usual in SI. The ARMAX (Moving-Average Auto-Regressive) model [10] is a well-known option to model a discrete system. If q is the delay unit, and q^{-d} the time delay $(k-d)$, ARMAX is described by $A(q)y(k) = 0$, where $A(q) = 1 + a_1q^{-1} + \dots + a_{na}q^{-na}$, being na the model size. In a polynomial description, the ARMAX model is given by (1).

$$y(k) + a_1y(k-1) + \dots + a_{na}y(k-na) = 0 \quad (1)$$

The identification of the time series consists in determining the values of a_i from the observation of the signal y . From this model, we can calculate the estimated signal $y_e(k)$ (2), and compare it with the real signal $y(k)$ in order to determine the error done (3). Hence, the time series in k is given by (4).

$$y_e(k) = [-a_1y(k-1) - \dots - a_{na}y(k-na)] = \varphi^T(k)\theta \quad (2)$$

$$\text{where } \theta = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_{na} \end{pmatrix} \quad \varphi = \begin{pmatrix} -y(k-1) \\ -y(k-2) \\ \dots \\ -y(k-na) \end{pmatrix}$$

$$err(k) = y(k) - y_e(k) = y(k) + [a_1y(k-1) + \dots + a_{na}y(k-na)] = 0 \quad (3)$$

$$y(k) = \varphi(k)\theta + err(k) \quad (4)$$

There are two possibilities to perform the identification (obtain θ): batch or recursive. We choose the recursive parametric estimation, which estimates and updates θ along the time, in a way that, for each time k , we obtain an ARMAX

model. Obviously, the more past samples we have, the more accurate model we obtain, because we have more information about the behaviour of the system.

There are several algorithms for the recursive identification: Kalman Filter, RLS (Recursive Least Squares) and LMS (Least Mean Squares). We choose RLS because its goodness and accuracy. It starts from the initial conditions: $\theta(p) = 0$ and $P(p) = 10,000I$, where I is the identity matrix and p is the initial time for the recursive algorithm, so that $p > na$. Next, RLS processes iteratively five steps: build the data matrix $\varphi(k)$, calculation of the estimated signal (5), calculation of the error made (3), calculation of an intermediate matrix (6), updating matrix P (7), and updating parameter matrix θ (8).

$$y_e(k) = \varphi^T(k)\theta(k-1) \quad (5)$$

$$K = \frac{P(k-1)\varphi(k)}{\lambda + \varphi^T(k)P(k-1)\varphi(k)} \quad (6)$$

$$P(k) = y \frac{P(k-1) - K\varphi^T(k)P(k-1)}{\lambda} \quad (7)$$

$$\theta(k) = \theta(k-1) + K\text{err}(k) \quad (8)$$

This algorithm considers the forgetting factor λ , which value is chosen in the interval 0.97 to 0.995 [10].

3 Energy Prediction in Access Points

The prediction of the future behaviour of the TS is possible if it is previously identified in order to know its behaviour by a mathematical description. The recursive estimation of the ARMAX model allows us to obtain this description. From this approach, the prediction can improve when the identification advances in the time, because we suppose more accurate models.

3.1 Prediction Approach

In order to perform the prediction, we choose the time k_s from which we model the system based on the past history. In practice, k_s will be the last known value of the TS. Figure 1 shows this approach, where y is the real TS, y_e is the estimated value from the ARMAX model, and y_s is the predicted TS. We have real values of y until $k = k_s$, so the last estimated value will be $y_e(k_s + 1)$, since the estimated value in $k_s + 1$ is calculated from the model built with the real values up to previous time, k_s . From $k_s + 1$, we predict by RLS assuming $y_s(k) = y_e(k)$. Therefore, $y_s(k_s + 1) = y_e(k_s + 1)$, and we apply RLS successively.

Figure 2 shows an example of TS predicted with this approach. The top plot shows the identification of the full TS ($na = 3$ and $\lambda = 0.98$), where y_e is the TS estimated from the model generated with all the data in the TS, and $NM = 39$ is the number of samples. Now, let's suppose we only know the TS up to $k_s = 20$. From this time, we generate the predicted signal y_s , calculated in this way:

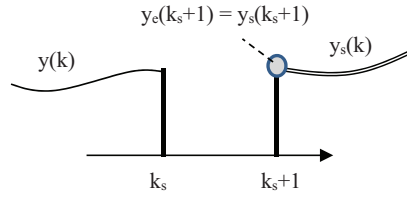


Fig. 1. The predicted signal is the estimated one from the next time to the last known.

- In the next time ($k_s + 1$) to the last known value $y(k_s)$, the predicted value y_s is the estimated one y_e according to the model ARMAX-RLS built considering the previous known y values.
- In $k_s + 2$, we perform the identification taking the real value of the TS as y_s , instead of y (we suppose we do not know it already). Therefore, we calculate the estimated value $y_e(k_s + 2)$ according to ARMAX-RLS, and next it is assigned again to the predicted signal y_s at this time, and so on.

We can see in the down plot of the figure the predicted y_s (the dotted plot is the real y which has not been taken into account for the prediction; it is shown only for comparison purpose). Obviously, the more predicted values, the worse prediction we have, because the predicted signal is built with the previous predicted values, instead of the previous real ones.

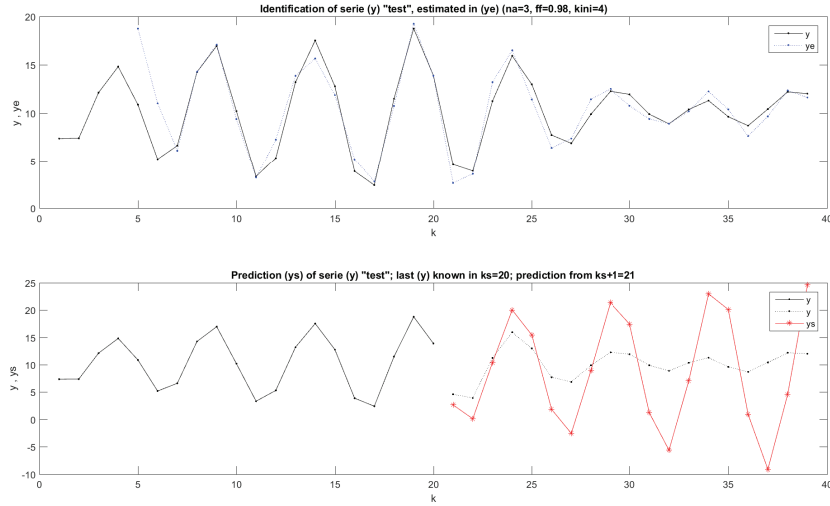


Fig. 2. Example of identification (top) and prediction from $k_s = 20$ (down).

3.2 Energy Data

We have collected data from a library building of the University of Extremadura (UEX), Spain (Figure 3). This Wi-Fi network is composed of 10 access points accessed by 2,907 users along 73 days: 10 full weeks labeled from #1 to #10, plus the first three days next to the last week; nevertheless, we only consider five of these 10 weeks, as we explain later. The energy data in each AP were collected each 60 seconds in the 12 hours daily period from 9h to 21h, in order to filter the energy data closer to the real users' behavior, since the usual activity of the network users happens in that period in the library building.

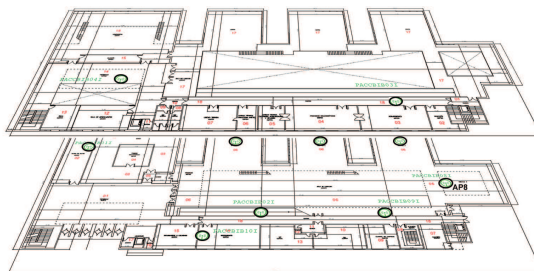


Fig. 3. Deployment of 10 APs in the Library building at the Caceres UEX campus.

From the collected data, we have built 10 time series, one for each AP, showing the total daily energy due to the users activity in the period. Our purpose is to predict the energy in the APs when we do not have more real data. The prediction is more accurate in the next day to the last known day, and it gets worse when the predicted day moves away. The three last days (71 to 73) are not part of the series, but they are left for comparison purposes with the three first predicted days.

We are interested in analyze the users' behavior through the energy patterns. For this purpose, we remove all the weekends, since the library is closed then. Moreover, we delete those weeks where one of their days does not reflect an usual behaviour (holiday, network down, etc). Taking into account these constraints, we only consider five weeks (25 days), as they are shown in Figure 4.

3.3 Access Point Technology

The model of wireless device where the data were collected is Alcatel-Lucent IAP-215. They have dual radio technology, supporting standards 802.11ac at 1.3 Gbps in 5 GHz band and 802.11n at 450 Mbps in 2.4 Ghz band. These APs are usual in high-performance Wi-Fi infrastructures, where up to 256 users can access simultaneously. The devices are configured as hive under InstantOS operating system from Aruba. At the same time, the hive is monitored by Nodowifi, a management tool from UEX [11].

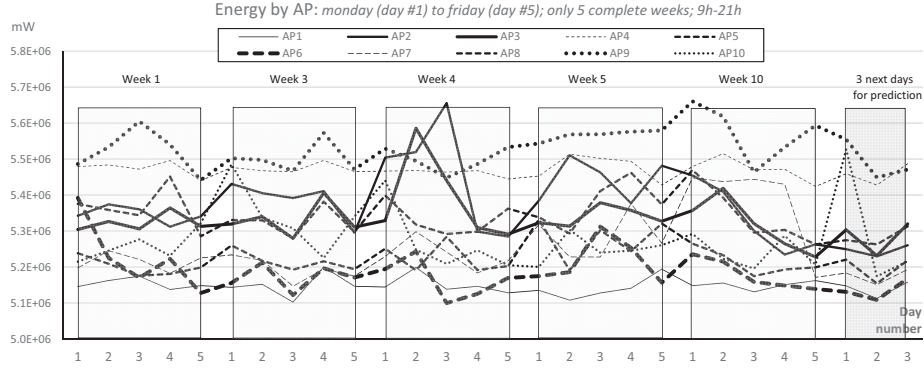


Fig. 4. Ten time series corresponding with the energy in each AP, sampled daily considering only working days: from Monday (day 1) to Friday (day 5). Each series is composed of 25 samples. The three last days considered for comparison purposes with the predicted days are also shown.

The IAP-215 device includes a CPU Freescale P1010 800 MHz, 256 MB SDRAM and 32 MB flash memories. The power and channel assignment in the installed APs is driven by Adaptive Radio Management (ARM) technology from Aruba, which analyzes the frequency spectrum in order to select the optimal power and channel, according to the regional configuration and the interferences and signals received from other devices. In any case, the power is selected in the range from 13 dBm to 18 dBm in both radios.

Last, the power of the APs is supplied by PoE (Power Over Ethernet) of 48 VDC with 802.3af compliance. The maximum operational energy consumption is 14.9 W, the minimum with radius is 4.5 W, and the power base emitting both radios at 18 dBm without connected users is 7 W. The efficiency of the PoE energy conversion to 12 VDC is around 88%.

4 Experimental Results

Figure 5 shows the prediction of the time series. Dotted and continuous lines are real (y) and predicted (ys) series respectively.

The value of na determines the initial time k_{ini} from which RLS builds the ARMAX model. The more na is, the more value for k_{ini} . Therefore, as the time series have a low number of samples ($NM=25$), we should consider a low value for na ; otherwise, the identification would start later, the recursive calculations would consider few data and, consequently, the identification would be worse.

The predicted values were compared with the real ones using the variable v (variation) defined in (9). The predicted data and their corresponding variations are shown in Table 1, and the absolute values of these variations are shown graphically in Figure 6 in order to notice better the prediction accuracy.

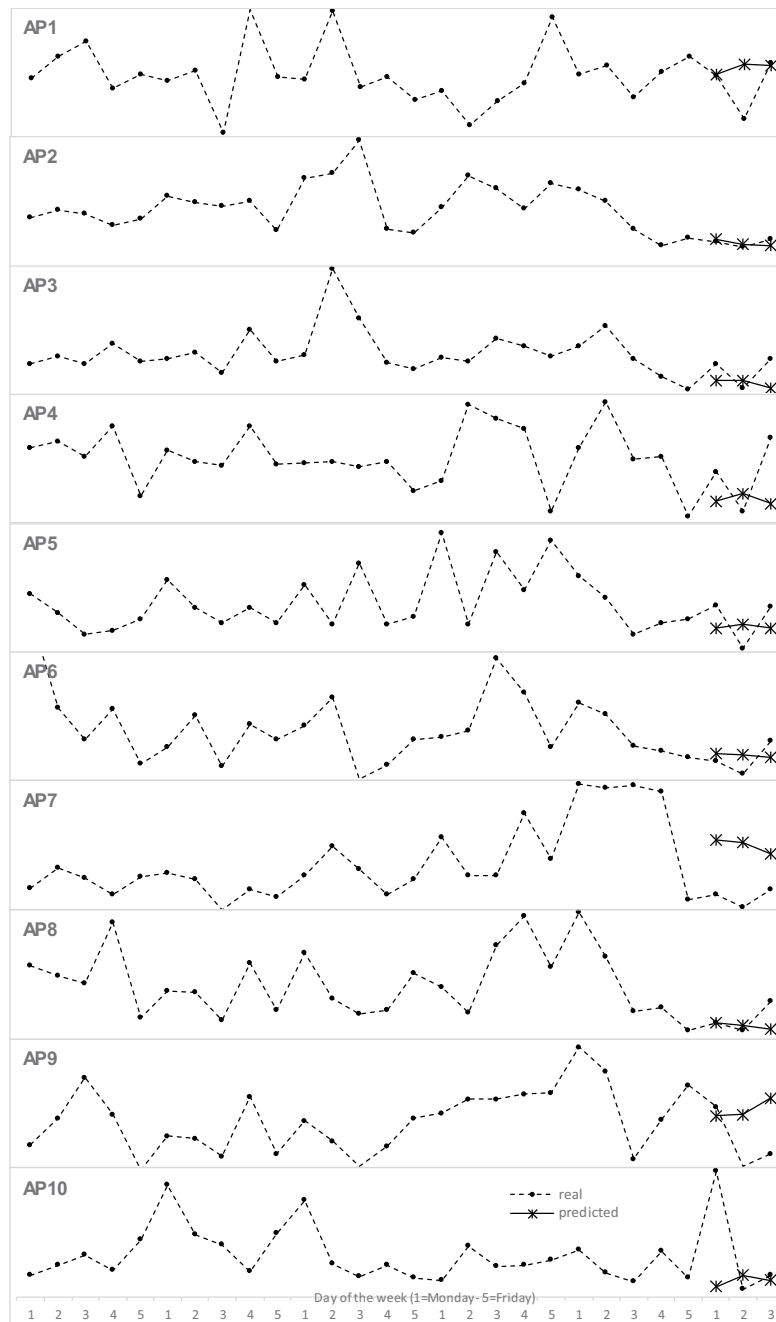


Fig. 5. Predicted energy (continuous line) in each AP in the next three days to the last considered day (sample number 25). The known energy data (dotted line) correspond with the working days of the five considered weeks, from Monday (1) to Friday (5). The real energies of the three last days are not considered in the time series, but they are useful for calculating the prediction performance. Energy is displayed in mW.

$$v(\%) = \frac{y_s - y}{y} \times 100 \quad (9)$$

		AP1	AP2	AP3	AP4	AP5
Real	$k_s + 1$	5,147,200	5,249,500	5,303,400	5,459,200	5,220,400
energy	$k_s + 2$	5,112,500	5,229,500	5,231,300	5,427,600	5,154,000
(mW)	$k_s + 3$	5,157,600	5,260,200	5,319,500	5,486,700	5,218,400
Predicted	$k_s + 1$	5,148,392	5,255,567	5,253,432	5,436,080	5,184,900
energy	$k_s + 2$	5,156,477	5,239,271	5,253,643	5,442,510	5,190,563
(mW)	$k_s + 3$	5,155,487	5,231,616	5,234,009	5,433,628	5,184,372
Variation	$k_s + 1$	0.02%	0.12%	-0.94%	-0.42%	-0.68%
	$k_s + 2$	0.86%	0.19%	0.43%	0.27%	0.71%
	$k_s + 3$	-0.04%	-0.54%	-1.61%	-0.97%	-0.65%
		AP6	AP7	AP8	AP9	AP10
Real	$k_s + 1$	5,130,600	5,183,100	5,274,600	5,554,300	5,526,900
energy	$k_s + 2$	5,108,400	5,150,600	5,262,600	5,448,900	5,173,800
(mW)	$k_s + 3$	5,167,400	5,192,800	5,312,900	5,470,800	5,216,100
Predicted	$k_s + 1$	5,144,935	5,312,812	5,274,947	5,538,894	5,180,182
energy	$k_s + 2$	5,143,113	5,305,317	5,270,491	5,540,667	5,213,637
(mW)	$k_s + 3$	5,137,779	5,279,461	5,263,770	5,569,270	5,197,231
Variation	$k_s + 1$	0.28%	2.50%	0.01%	-0.28%	-6.27%
	$k_s + 2$	0.68%	3.00%	0.15%	1.68%	0.77%
	$k_s + 3$	-0.57%	1.67%	-0.92%	1.80%	-0.36%

Table 1. Real and predicted energies, and the corresponding variations, for each access point, in the next three days to the last day considered ($k_s = 25$).

In general, the predictions are good; as regards they do not moves away too much from the real values (less than 2% in almost all the cases). Only two of the 30 predictions have variations of 3% and 4%. Besides, the variations for the first predicted day ($k_s + 1$) are under 1% in eight of the 10 access points.

The variation can be positive or negative, showing the trend to increase or decrease the own time series (according to the day in the week, the trend in the network use will be greater or smaller).

We can see how the prediction gets worse when it is far from the last known value ($k_s = 25$). This is observable in six of the 10 access points; in the other 4 APs, the prediction is different, depending on the behavior and variability of the own time series.

Finally, we remember that the size of the time series and its behavior influences on the prediction results. On the one hand, the more weeks the TS has, the more accurate the prediction will be, because we have more samples of the same day. On the other hand, the variability of the TS (a very different behavior between consecutive days or for the same day in each week) implies a

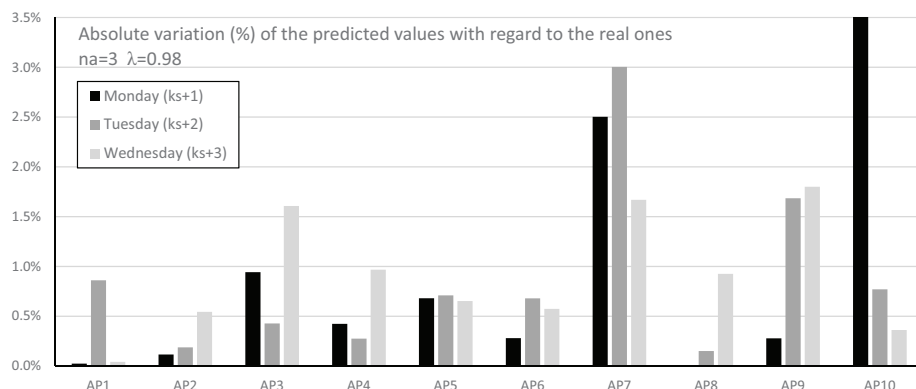


Fig. 6. Absolute variation of the predicted energy with regard to the real energy, in each access point, for the next three days to the last known day (k_s): Monday ($k_s + 1$), Tuesday ($k_s + 2$) and Wednesday ($k_s + 3$)).

worse identification. Ideally, the prediction will be better as more stable be the behavior of the time series. The worse results of our work can be blamed to the access points that show high variability in its behavior (for example, AP7).

5 Conclusions

We have applied the time series analysis to the daily energy consumption in the access points of a Wi-Fi network, in order to predict the values for the next days. This knowledge can be useful for network maintenance purposes. We have collected data from a real wireless infrastructure, in an academic environment, accessed by thousands of students during several weeks. The time series, one for each access point, were modeled following an auto-regressive formulation and a recursive estimation. The result models were applied to do the prediction, whose results were compared with real energies in order to analyze the performance.

We think the obtained results are good enough, since the differences between prediction and real values are under 2% in almost all the cases. Nevertheless, we point out some research future efforts to improve the prediction. First, we can add other parameters different than energy consumption as possible data to be predicted for maintenance issues, like number of users, sessions and data traffic, since these information is easily accessible from the AP itself. Second, the time series should be larger, in order to analyze more data of the same type that reflect the usual behavior; besides, this would allow us to build larger models; for example, we may consider a wider window of days in order to give more representation to the different parts of the year (holidays, beginning of the academic year, examination period, etc.). Third, we can consider other networks composed of APs of different types and models in order to lead to different consumption behavior depending on the energetic efficiency. Last, we can optimize

the settings of the main parameters involved in the prediction processes, such as the forgetting factor in the recursive identification algorithm; in this case, we would use the error made as cost or objective function.

Acknowledgments. This work was partially funded by the Government of Extremadura under the project IB16002, and by the AEI (State Research Agency, Spain) and the ERDF (European Regional Development Fund, EU) under the contract TIN2016-76259-P (PROTEIN project).

References

1. Hamamoto, R., Takano, C., Obata, H., Aida, M., Ishida, K., Mitton, N., Kantarci, Melike, E., and Gallais, A.: Setting Radio Transmission Range Using Target Problem to Improve Communication Reachability and Power Saving. In: *Proceedings of Ad Hoc Networks 7th International Conference*, pp. 15–28. Springer (2015)
2. Konstantinidis, A., Yang, K.: Multi-objective K-connected Deployment and Power Assignment in WSNs using a problem-specific constrained evolutionary algorithm based on decomposition. *Computer Communications* 34, 83–98 (2010)
3. Zhang, H., Chu, X., Guo, W. and Wang, S.: Coexistence of Wi-Fi and heterogeneous small cell networks sharing unlicensed spectrum. *IEEE Communications Magazine* 53, 158–164 (2015)
4. Shen, Y., Jiang, C., Quek, T., Zhang, H., Ren, Y.: Pricing equilibrium for data redistribution market in wireless networks with matching methodology. In: *2015 IEEE International Conference on Communications (ICC)*, pp. 3051–3056 (2015)
5. Scellato, S., Musolesi, M., Mascolo, C., Latora, V., Campbell, A.: NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems. In: *Proc. of the 9th International Conference on Pervasive Computing 2011*, pp. 152–169 (2011)
6. Hamamoto, R., Takano, C., Obata, H., Ishida, K.: Improvement of Throughput Prediction Method for Access Point in Multi-rate WLANs Considering Media Access Control and Frame Collision. In: *2015 Third International Symposium on Computing and Networking (CANDAR)*, pp. 227–233 (2015)
7. François, J., Leduc, G.: AP and MN-Centric Mobility Prediction: A Comparative Study Based on Wireless Traces. In: *Proceedings of the 6th International IFIP-TC6 Networking Conference*, pp. 322–332 (2007)
8. Gmach, D., Rolia, J., Cherkasova, L., Kemper, A.: Workload Analysis and Demand Prediction of Enterprise Data Center Applications. In: *Proceedings of the 2007 IEEE 10th International Symposium on Workload Characterization*, pp. 171–180 (2007)
9. Soderstrom, T., Stoica, P.: *System Identification*. Prentice-Hall Int., London (1989)
10. Ljung, L.: *System Identification. Theory for the User*. Prentice-Hall, Englewood Cliffs, N.J., 1999.
11. Nodowifi project. Wireless Labs, University of Extremadura (Spain), <http://www.nodowifi.es>

A Combination of Variational Mode Decomposition with Neural Networks on Household Electricity Consumption Forecast

Vanessa Haykal, Hubert Cardot, and Nicolas Ragot

Université François Rabelais Tours, Computer Science Lab (LI, EA 6300), France

Abstract.

Recently, there has been a significant emphasis on the forecasting of the electricity demand due to the increase in the power consumption. This paper presents the computational modeling of electricity consumption based on Neural Network (NN) training algorithms. The noise in signals, which are caused by various external factors, often corrupt demand series and influence consequently on the model performance. For accurate electricity demand forecasting, we propose a novel approach that combines a NN MLP (multilayer perceptron) with VMD (variational mode decomposition)-based signal filtering. Using the daily electricity demand series of EDF (Electricité De France) obtained from the UCI machine learning repository, this paper demonstrates that the proposed VMD-NN model greatly improves the forecasting error comparing to existing stationary stochastic process such as the autoregressive moving average (ARMA) model.

Keywords: neural network algorithms, time series, household electricity consumption forecast, variational mode decomposition, multiresolution analysis

1 Introduction

Domestic energy consumption [1] is the total amount of energy used in a house for household work. The amount of energy used per household varies widely depending on the standard of living of the country, the climate, and the age and type of residence. Energy demand forecasting is a very important task in the electric power distribution system to enable appropriate planning for future power generation. Quantitative and qualitative methods have been utilized previously for the electricity demand forecasting. These methods fail to provide effective results. With the development of the advanced tools, these methods are replaced by efficient forecasting techniques. According to common classifications [2], demand forecasting models are classified based on two different criteria: the forecasting horizon and the aim of the forecast, also we can divide them into linear and nonlinear models and a third group consists of models that use a combination of both.

This paper presents an improved method for forecasting, we use the VMD-NN model. The VMD is a fully adaptive method for the analysis of nonlinear and non-stationary properties of time series. The original series will be decomposed by the VMD method into several high and low frequency signals. These sub-series will be used in the NN model in order to make the prediction. The forecasting results of this work have revealed that the VMD-NN model outperforms the NN itself and the ARMA models.

The rest of this paper is organized as follows: Section 2 introduces the notions of the classical forecasting models namely the autoregressive moving average and the artificial neural networks. Section 3 shows in details the theory of the recently developed variational mode decomposition. Section 4 contains our experiments and results, we start this part by defining the practical error measurement, we describe different steps to get the optimal training algorithm on our dataset, and we sum up our work with some simple quantitative performance evaluations compared to the baseline models. Section 5 concludes on the effectiveness of our novel approach, and includes some future directions and expected improvements.

2 Classical Forecasting Methods

2.1 ARMA Process

The various researches [3] have used these methods with time series data for the electric power consumption. In [4] Zhu, Guo, and Feng studied the issue of household energy consumption in China from the year 1980 to 2009 with construction on of VAR model. There were two forecasting methods that used ARIMA and BVAR. The results showed that both of them can predict the sustained growth of household energy consumption (HEC) trends. Ediger and Akar [5] applied SARIMA (Seasonal ARIMA) methods to estimate the future primary fuel energy demand in Turkey from the years 2005 to 2020. The research work of Contreras et al. [6] applied ARIMA methods to predict next day electricity price in Californian markets. Conejo et al. [7] applied wavelet transform and ARIMA models to predict day-ahead electricity price of mainland Spain in year 2002.

In this paper, we use the ARMA process [8, 9] as a reference model. It has become a popular linear statistical model for stationary time series analysis and forecasting. The ARMA (p, q) generating process is given by

$$\varphi(B) v_t = \theta(B) e_t$$

where v_t and e_t are respectively the actual value and random error at time period t , B is the backshift operator. The error term e_t are assumed to be independently and identically distributed (*iid*) with a mean $E(e_t) = 0$ and a variance $V(e_t) = \sigma^2$. The polynomials $\varphi(B)$ and $\theta(B)$ are given by

$$\begin{aligned}\varphi(B) &= (1 - \varphi_1 B - \dots - \varphi_p B^p) \\ \theta(B) &= (1 - \theta_1 B - \dots - \theta_q B^q)\end{aligned}$$

where p is the number of autoregressive orders, q is the number of moving average orders, θ is the autoregressive coefficient, and φ is the moving average coefficient.

In particular, the autoregressive (AR) component is expressed by the coefficients φ that represent a linear relationship between the value predicted by the model at time t and the past values of the interest rate variation time series. Similarly, the moving average (MA) component is expressed by the coefficients θ that represent a linear relationship between the value predicted by the model at time t and the error term e .

2.2 Artificial Neural Networks

Based on some literature reviews, the non-linear models, derived from the artificial neural networks (ANNs), have gained more and more attention since the second half of the 80's. This evolution is due to the fact that certain researchers achieved great advances on ANNs.

Artificial neural networks in Fig.1 [10] are a class of statistical learning models inspired by the physiology of biological neural networks. Each neuron performs a specific kind of computation. First, a weighted sum of the input variables and the bias term b is built, with the result being then processed by an activation function $f(t)$. Once the single neuron operation is specified, one can easily calculate the network outputs given an input vector by evaluating the output of each layer by forward input propagation. The result is a function of the network configuration, i.e. its topology and the value of the connection weights. It will be the job of the training phase to learn the weights in order to induce the desired computation.

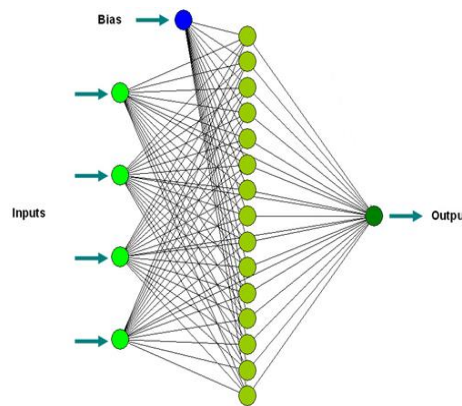


Fig. 1. Schematics of a fully connected multilayer perceptron with four inputs and a bias unit. The weighted input sum is added to the bias term and then enters as argument of the activation function f which generates the output (Neuromaster, 2015)

This has been overcome by the back-propagation [11, 12] algorithm; nowadays it is widely applied in training multilayer perceptron. Given a supervised training set $\{x_i, t_i: i = 1 \dots N\}$ with x_i input variables and t_i target variables, we denote by y_i the correspondent output computed by the network when x_i is fed forward. In general, we have $t_i \neq y_i$. A global error on the training set can be then defined as a quadratic function of the form

$$E(\mathbf{w}) = \frac{1}{2N} \sum_i ||\mathbf{t}_i - \mathbf{y}_i||^2$$

and can be seen as a function of the network weights \mathbf{w} . Other error definitions are possible, for example by choosing a different norm. The idea behind back propagation is to minimize this error by updating the weights using the gradient descend [13] method (with k as iteration index), i.e.

$$w_{ij}^{(k)} \rightarrow w_{ij}^{(k)} - \alpha \frac{\partial E(\mathbf{w})}{\partial w_{ij}^{(k)}}$$

The calculation of the partial derivatives is thus crucial for the algorithm. It is done by using directly the dependence of the error function on the training set instances. When all the instances have been used, one ‘epoch’ of training is completed. Usually many epochs of training are needed in order for the error function to converge to a local or global minimum, resulting in longer training periods.

3 Variational Mode Decomposition

More recently, a new multiresolution technique called variational mode Decomposition (VMD) was introduced by Dragomiretskiy and Zosso (2014) yields better results in signal processing domain specifically in the case of signals without prior knowledge. In [14], they propose an entirely non-recursive variational mode decomposition model. The model looks for an ensemble of modes and their respective center frequencies. We apply this technique on our dataset before using the predictive neural network model described in Section 2.2. In this part, we mentioned a few concepts and tools from signal processing that will constitute the building blocks of the VMD model.

3.1 Denoising Problem

Using a simple denoising problem, an underlying signal f_0 consist of an unknown signal f corrupted by an additive noise, namely the zero-mean Gaussian noise. The use of the Wiener filter is to estimate the unknown signal using an original signal as input. The filter is based on a statistical theory in order to minimize the mean squared error classically addressed using Tikhonov regularization [15].

$$\min_f \{ \|f - f_0\|_2^2 + \alpha \|\partial_t f\|_2^2 \},$$

The Euler-Lagrange equations are typically solved in Fourier domain:

$$\hat{f}(\omega) = \frac{\hat{f}_0}{1 + \alpha\omega^2},$$

where \hat{f} is the fourier transform of the signal f . This solution corresponds to convolution with a Wiener filter, where α represents the variance of the white noise, and the signal has a low-pass $1/\omega^2$ power spectrum prior.

3.2 Constrained Model

For a multicomposition real valued signal f , VMD assumes that f is composed of a given number of subsignals u_k (modes). Each mode is regarded as an amplitude-modulated and frequency-modulated (AM-FM) signal and has mostly compact frequency ω_k around a center pulsation [16].

To assess the bandwidth of the modes, the following scheme is proposed by Dragomiretskiy and Zosso (2014):

- (a) Compute the associated analytic signal by means of the Hilbert transform to obtain a unilateral frequency spectrum for each mode.
- (b) Shift the frequency spectrum of each mode to the baseband by mixing with an exponential tuned to the estimated center frequency.
- (c) Estimate the bandwidth through the H^1 Gaussian smoothness of the frequency translated function, that is, the squared L^2 -norm of the gradient.

The resulted constrained variational problem is the following:

$$\begin{aligned} \min_{\{u_k\}, \{\omega_k\}} \quad & \left\{ \sum_k \left\| \partial_t \left[\left(\sigma(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ \text{s.t.} \quad & \sum_k u_k(t) = f. \end{aligned}$$

where f is the signal, u is its mode, w is the frequency, σ is the Dirac distribution, t is time script, k is number of modes, and $*$ denotes convolution.

Thus, we intend to minimize the sum of the bandwidths defined as the squared L^2 -norm of the gradient of the demodulated signal components. To solve the constrained variational problem [16], the augmented Lagrangian is introduced and the non-constrained variational problem is gotten by

$$\begin{aligned} L(\{u_k\}, \{\omega_k\}, \lambda) \\ = \alpha \sum_k \left\| \partial_t \left[\left(\sigma(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 \\ + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle, \end{aligned}$$

where α denotes the balancing parameter of the fidelity constraint and λ is the lagrangian multiplier. The saddle point here is to get the optimal solutions of u_k and w_k using the alternate direction method of multipliers (ADMM).

3.3 Minimization with respect to u_k

The subproblem can be written as the following equivalent minimization problem:

$$u_k^{n+1} = \arg \min_{u_k \in \mathbb{R}} \left\{ \alpha \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| f - \sum u_i + \frac{\lambda}{2} \right\|_2^2 \right\}.$$

The solution of this quadratic optimization problem is found by using Parseval/Plancherel Fourier isometry and exploiting the Hermitian symmetry [16]. All the modes can be obtained from the below equation in the frequency domain through updating each mode and its center frequency ω_k constantly:

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \hat{\lambda}(\omega)/2}{1 + 2\alpha(\omega - \omega_k)^2}.$$

This equation is regarded as the Wiener filtering result of the current residue with signal prior $1/(\omega - \omega_k)^2$. Consequently, the mode in time domain is obtained as the real part of the inverse Fourier transform of this filtered analytic signal.

3.4 Minimization with respect to ω_k

As before, the optimization takes place in the Fourier domain. The relevant subproblem thus reads:

$$\omega_k^{n+1} = \arg \min_{\omega_k} \left\{ \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\}.$$

The new center frequency is put at the center of gravity of the corresponding mode's power spectrum, which can be updated by

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega}.$$

4 Experimentation

4.1 Performance Measure

The forecasting performance [9] is examined using the root mean of squared errors (RMSE). It measures the deviation between actual and predicted values. A small value of RMSE means that the predicted time series values are closed to the actual values. Thus, it can be used to evaluate the prediction error. The computation of this criterion is given as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (v_i - p_i)^2}{n}}$$

where v and p represent respectively the actual and predicted value and n is the total number of the sample data points.

4.2 Data Preprocessing and Training

This research used dataset [17] about the electric power consumption in one household that has a sampling rate one minute over a long period of time from the years 2006 to 2010. Following the existing study [3], we use the “Global Active Power” variable which is the household global minute-averaged active power (in kilowatt).

The raw data were not ready for constructing the forecast model because some values are missing and the recorded time frames are inappropriate. The lack of some information [3] may decrease the predictive efficiency of the forecasting model. To fill the missing data, we use the previous value, where we assume that the current data will be similar to the previous ones as shown in Fig.2.

Date	Time	Global_active_power
21/12/2006	11:19:00	0.244
21/12/2006	11:20:00	0.244
21/12/2006	11:21:00	0.242
21/12/2006	11:22:00	0.244
21/12/2006	11:23:00	0.244
21/12/2006	11:24:00	0.244
21/12/2006	11:25:00	0.246
21/12/2006	11:26:00	0.246
21/12/2006	11:27:00	0.244
21/12/2006	11:28:00	0.244

Fig. 2. Fill the missing data by the previous value “0.244” (extracted from [3])

We aggregate the minute-by-minute data into daily observations, and then we got a new sample of 1442 data points. In Fig.3, we represent the shape of the daily HEC time series. It is clearly shown that the data points are highly fluctuated and non-stationary, since their means and variances change over time.

The data is divided into two parts. The first group is the training dataset which contain data from 26/12/2006 to 31/12/2009 (1112 observations) for the construction of the predictive models. The second group is the test sample which contain data from 01/01/2010 to 26/11/2010 (330 observations).

We want to predict the future value based on n previous days, where n is considered as the window size of the time series. There is no specific rule to define the window, in our case, we choose the first thirty lags for training the artificial neural networks.

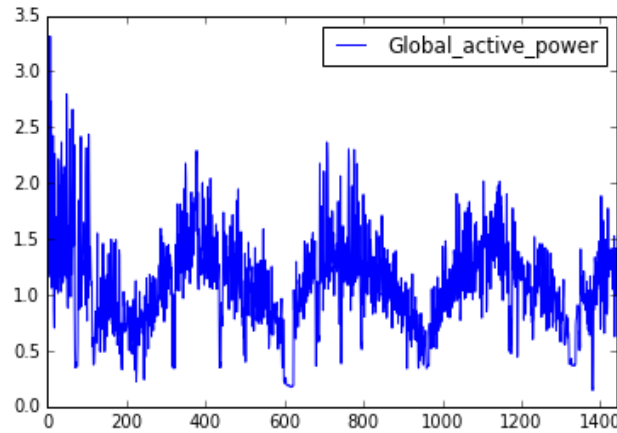


Fig. 3. The daily time series of the Global Active Power from 2006 to 2010

The main aim of this work is to determine the optimal NN for electricity demand forecasting. The model is generated by using the grid search cross validation technique. In brief, the grid search is simply an exhaustive searching algorithm with a manual specified subset for hyperparameter optimization. It must be guided by some performance metric, typically measured by the k-fold cross-validation. The cross validation algorithm does this by splitting the training dataset into k subsets and takes turns training models on all subsets except one which is held out, and evaluating model performance on the held out validation dataset. The process is repeated until all subsets are given an opportunity to be the held out validation set. The performance measure is then averaged across all models that are created.

We used this technique to identify the number of layers (length) and the number of neurons in each layer (width). First, we vary the length from 1 to 5 layers with a width of 10, 60, 110 neurons per layer. We found that the optimal model structure consists of a hidden layer with 110 neurons. Then, we tried again the algorithm for one layer with a range of neurons between 90 and 120; the record gives 100 neurons as the best fit. Thus, the optimal neural network is composed of one layer with 100 neurons. Throughout the training, we use an epoch equal to 1000. Considering that, one epoch is a multiple number of iterations for the gradient descent updates until we show all the data to the NN, and then start again.

As proposed in this paper, we work on the VMD-based signal filtering method to reduce the noise. The VMD algorithm requires predetermining the number of variational modes to be extracted. However, it is not easy to set a rule to determine an appropriate number of modes.

On the one hand, we tried our experiments on the training sample with mode = 6 in order to exemplify the theory described in Section 3. We illustrate in Fig.4 the results of the VMD algorithm on the non-stationary HEC dataset in order to assess the clarification of the proposed approach (Dragomiretskiy & Zosso, 2014).

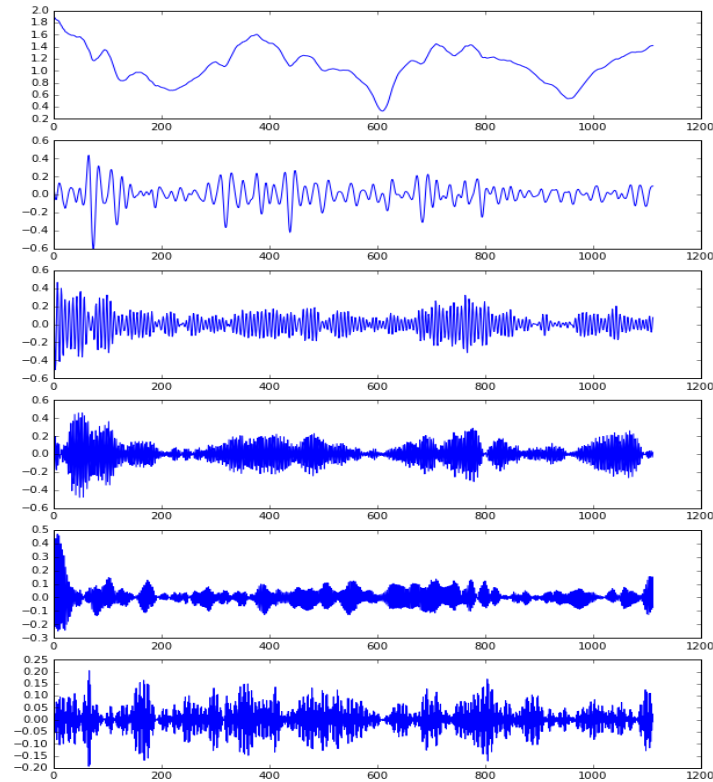


Fig. 4. Applying the VMD = 6 on the train sample

One can clearly see an oscillating low-frequency pattern. The first mode captures the low-frequency oscillation of the baseline. Then, the distinct spikes of the train sample create important higher-order harmonics in the next modes. The 6th mode is the highest frequency mode and contains the most noise with a highly non-sinusoidal spikes.

On the other hand, in order to contribute to the forecasting system, we repeatedly apply the VMD algorithm on every window of the time series to elaborate the modes. We tried our experiments on different levels of decomposition: 8, 10, 15, and 20. In each case, the modes are integrated in the optimized MLP model for 7 day-ahead forecast. Based on the RMSE values explained in the Section 4.1, we found that VMD = 15 gives the most efficient result.

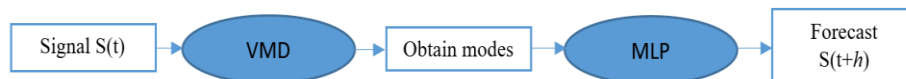


Fig. 5. The VMD-NN model components

As a partial conclusion, the following diagram in Fig.5 explains the different steps of the hybrid combination VMD-NN model in order to build this novel forecasting approach. We mean by signal $S(t)$, the series values that correspond to the window range from 1 to 30. Applying the VMD on each signal, we obtain the 15 different variational modes. Then, using the previous optimized MLP (100,) as indicated in Fig.5, we considered the variational modes as inputs of the model in order to obtain the prediction on horizon h .

4.3 Results

We define the Carbon Copy as the model that takes exactly the same value of the previous day, which means that the predicted value is equal to the actual value. For comparison purpose with ARMA models (Chujai et al., 2013), the horizon $h = 7$ was employed.

Based on the test sample and the optimal models, we make a summary of the RMSE values shown in the table 1. Using the NN model, we found a significant improvement in the error comparing to ARMA process for the proposed dataset. The RMSE decreases from 0.34 to 0.272. However, the novel approach VMD-NN model clearly shows an efficient reduction in the error among all the previous studies. Its corresponding RMSE is equal to 0.077. The VMD-NN greatly outperforms the NN itself by decreasing the RMSE from 0.272 to 0.077 for predicting the household electricity consumption dataset.

Models	RMSE $h = 7$
Carbon Copy	0.374
ARMA (Chujai et al. [3])	0.340
NN	0.272
VMD-NN	0.077

Table 1. RMSE comparative analysis

Thus, based on the VMD-NN model, the RMSE analysis shows that we divide the error by 4.4 ($\sim 0.34/0.077$) comparing to ARMA, and by 3.5 ($\sim 0.272/0.077$) comparing to NN.

The VMD technique cannot be applied to ARMA model. The variational modes cannot be implemented in its algorithm, since it only involves regressing the variable on its own lagged (i.e. past) values.

The RMSE of the VMD-NN could be also minimized by making a new optimization of the hyperparameters of the neural network. But, as long as we get a significant decrease in the RMSE, we restrict our study to the MLP (100,) to show the effectiveness of the VMD comparing to the same previous MLP.

In Fig. 6, we only plot the first 150 observations of the test sample to clearly show the difference between the curve of the predicted and original values. The curves are very close to each other, thus the VMD-NN model fits the data very well.

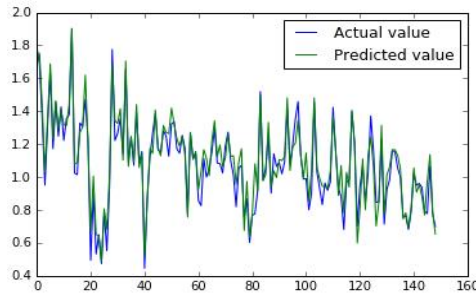


Fig. 6. The actual versus predicted values

Besides the RMSE measures, we also examined the distribution of the forecast errors in order to check the normality of the distribution. The histogram in Fig.7 shows that the errors are normally distributed between $[-0.3, 0.3]$ where the highest point on the curve represents the most probable event in the error close to zero, while all other possible occurrences are equally distributed around the center, creating a downward-sloping line on each side of the peak.

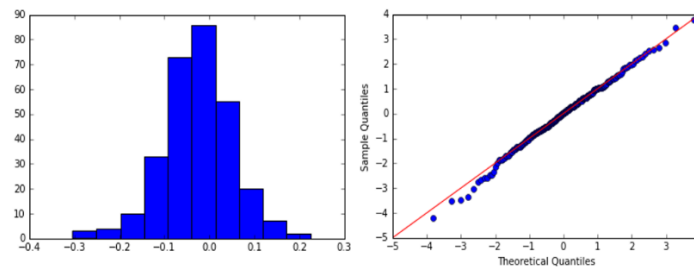


Fig. 7. The histogram (left) and the Q-Q plot (right) of the forecast errors

We use the Q-Q plot in Fig.7 as a test to verify the normality. Roughly speaking, the Q-Q plot take the sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution known as the standard normal distribution with mean 0 and standard deviation 1. If both sets of quantiles come from the same distribution, we should see the points converge to the straight line. As long as the blue points in Fig.7 are close to the red line, the normality can be assumed, and we have stability in the model error.

5 Conclusion

Due to the lack of research on this UCI dataset, our objective was to build a forecast system to make some improvement comparing to existing studies specifically on a daily level. Experiments with RMSE statistical criteria, clearly demonstrate that VMD-based Neural Network model significantly achieved the lowest forecasting error among models. This indicates that this novel approach can be used as a very promising methodology specifically for non-stationary and noisy time series. The VMD is considered as a

new adaptive multiresolution technique, and this is the main advantage of adopting this approach.

Finally, a comparative study of accuracy of the VMD combined with other machine learning models such as support vector machines could be considered for future works to also examine its effectiveness.

6 References

1. T.M. Usha and S. Appavu alias Balamurugan, "Computational modeling of electricity consumption using econometric variables based on neural network training algorithms," 2016
2. L. Hernandez, C. Baladron, J. M. Aguiar, B. Carro, A. J. Sanchez-Esguevillas, J. Lloret and Joaquim Massana, "A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings," IEEE Communications Surveys & Tutorials, pp. 14601495, Third Quarter 2014.
3. P.Chujai and N. Kerdprasop "Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models," 2013
4. Q. Zhu, Y. Guo and G. Feng, "Household energy consumption in China forecasting with BVAR model up to 2015," 2012
5. Volkan Ş. Ediger, Sertaç Akar, "ARIMA forecasting of primary energy demand by fuel in Turkey," Energy Policy, vol.35, 2007, pp.1701-1708.
6. Javier Contreras, Rosario Espinola, Francisco J. Nogales, and Antonio J. Conejo, "ARIMA models to predict next-day electricity prices. Power Systems," IEEE Transactions on 2003, vol.18, no. 3, pp. 1014-1020.
7. Antonio J. Conejo, Miguel A. Plazas, Rosa Espinola, and Ana B.Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," IEEE Trans. Power Syst., vol. 20, no. 2, 2005, pp. 1035-1042.
8. G. Box, G. Jenkins, "Time Series Analysis: Forecasting and Control," Holden-Day, San Francisco, 1970.
9. S. Lahmiri, "Interest rate next-day variation prediction based on hybrid feedforward neural network, particle swarm optimization, and multiresolution techniques," 2016
10. Q. Yi Feng, R. Vasile, M. Segond, A. Gozolchiani, Y. Wang, M. Abel, S. Havlin, A. Bunde, and H. A. Dijkstra, "ClimateLearn: A machine-learning approach for climate prediction using network measures," 2016
11. Bishop, C. M. "Pattern recognition and machine learning," Springer, New York, 2006.
12. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation. In parallel distributed processing: Exploration in the microstructure of cognition," volume I, Bradford Books, Cambridge, MA, 1986
13. Yann A LeCun, L. Bottou, Genevieve B Orr, K.-R. Müller, "Efficient backprop," 2012
14. S. Liu, G. Tang, X. Wang, and Y. He, "Time-Frequency Analysis Based on Improved Variational Mode Decomposition and Teager Energy Operator for Rotor System Fault Diagnosis," 2016
15. A. N. Tichonov, "Solution of incorrectly formulated problems and the regularization method," Soviet Mathematics, vol. 4, pp. 1035–1038, 1963.
16. K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," IEEE Transactions on Signal Processing, vol.62, no.3, pp. 531–544, 2014.
17. <https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption>

Nonparametric panel stationarity testing. An application to crude oil production

Landajo, M.¹, Presno, M.J.¹ and Fernández González, P.¹

¹ University of Oviedo, Spain

Abstract. A nonparametric panel stationarity test is proposed that offers the advantage of not requiring *a priori* specification of the trend function for each of the series in the panel. A bootstrap implementation of the test is outlined and its finite sample performance is analyzed via Monte Carlo simulations. The stochastic properties of monthly crude oil production are then analyzed for a panel of 20 -both OPEC and non-OPEC- countries in the period from January 1973 to December 2015. Our analysis detects strong evidence of non-stationarity, both globally and group-wise for both OPEC and non-OPEC countries. A case-by-case study reveals that stationarity is rejected for 8 out of the countries under study, with stationarity being relatively more frequent among OPEC members.

Keywords: Stationarity, Panel, Nonparametric, Oil Production.

1 Introduction

A potential drawback of conventional unit root and stationarity testing stems from its lack of robustness to misspecification of the trend function of the series. Conclusions may change depending on the presence of breaks (as well as their number and even the speed of change¹) and whether non-linear features are taken into account or not. In practice this is a serious limitation, as for many series it is hard to *a priori* specify a simple parametric form for their deterministic trend component. The problem is even more serious in the panel case as many panel test statistics are averages of the corresponding test statistics for the individual components of the panel, so correct model specification is required simultaneously for *all* the series in the panel, at the risk of undue rejection of the null hypothesis of the test as a consequence of model misspecification for some components of the panel.

Our goal in this work is developing an approach that properly addresses the above limitations. For this we shall rely on nonparametric panel stationarity testing. In a recent paper, Landajo and Presno (2013, LP hereafter) propose a fully nonparametric (univariate) stationarity test that offers the advantage of not requiring *a priori* specification of the trend of the series. In this paper we build on that contribution, proposing a panel, bootstrap-based extension of the LP test.

¹ See Landajo and Presno (2010).

The proposed test is applied to analyze stationarity in monthly oil production, along the period from 1973 to 2015, for some of the leading producer countries. We shall focus on a panel of 20 OPEC and non-OPEC countries encompassing more than 80% of global oil production.

The remainder of this paper is organized as follows. Section 2 outlines the methodology and includes the technical details on implementation of the proposed test and an extensive simulation analysis on its finite sample performance under several trend specifications and time series models. In Section 3, the datasets are presented, together with the empirical results and a discussion. Some concluding remarks are included in Section 4.

2 Methodology

2.1. The model and the nonparametric panel stationarity test

We consider a panel of N (*fixed*) time series $\mathbf{y}_t = (y_{1,t}, \dots, y_{N,t})$ generated by the following multivariate process:

$$\begin{aligned} y_{i,t} &= \mu_{i,t} + \theta_i^*(t/T) + \varepsilon_{i,t}, \\ \mu_{i,t} &= \mu_{i,t-1} + u_{i,t}; \quad t = 1, \dots, T; \quad T = 1, 2, \dots; \quad i = 1, 2, \dots, N \end{aligned} \quad (1)$$

with $\theta_i^*: [0, 1] \rightarrow \mathbb{R}$ being the trend function of the i -th time series in the panel. $\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \dots, \varepsilon_{N,t})$ is a zero mean random vector process (both serial dependence and cross-section correlation among the components of $\boldsymbol{\varepsilon}_t$ is allowed). In addition, for any $i = 1, \dots, N$, the processes $\{\varepsilon_{i,t}, t = 1, 2, \dots\}$ and $\{u_{i,t}, t = 1, 2, \dots\}$ are assumed to be independent of each other having zero means and respective (finite) variances $E(\varepsilon_{i,t}^2) = \sigma_{i,\varepsilon}^2 > 0$ and $E(u_{i,t}^2) = \sigma_{i,u}^2 \geq 0$; $\{\mu_{i,t}\}$ starts with $\mu_{i,0}$, which is assumed to be zero for each $i = 1, \dots, N$.

We consider the following panel stationarity testing problem:

$$H_0: q_i \equiv \frac{\sigma_{i,u}^2}{\sigma_{i,\varepsilon}^2} = 0 \text{ for } i = 1, \dots, N, \text{ versus } H_1: \sum_{i=1}^N q_i > 0 \quad (2)$$

In the above setting, under the null hypothesis (H_0), all the series of the panel are stationary around their respective deterministic trend functions, whereas at least one of the series includes a unit root under the alternative (H_1).

Stationarity can be tested separately for each component of the panel by using the nonparametric stationarity test derived by Landajo and Presno (2013). In that setting, the trend function $-\theta_i^*(t/T)$ of time series $y_{i,t}$ is first estimated nonparametrically by OLS regression of $y_{i,t}$ on the elements of a cosine basis. The resulting estimate has the form:

$$\hat{\theta}_i(t/T) = \hat{\beta}_{i,0} + \sum_{j=1}^{m_T} \hat{\beta}_{i,j} \cos(j\pi t/T) \quad (3)$$

Model complexity (m_T) in (3) grows with sample size (T) obeying a suitable deterministic rule (e.g., a rule as $m_T = [cT]^{1/5}$, with $c > 0$ and $[\cdot]$ denoting the integer part function, is often appropriate). Then the raw (KPSS-type) stationarity test statistic

for series $y_{i,t}$ is readily computed from the OLS residuals of the above regression, namely:

$$\hat{S}_{i,T} = \frac{\sum_{t=1}^T E_{i,t}^2}{\hat{\sigma}_i^2 T^2} \quad (4)$$

where, $E_{i,t} = \sum_{k=1}^t \hat{\varepsilon}_{i,k}$ with $\hat{\varepsilon}_{i,k} = y_{i,k} - \hat{\theta}_i(k/T)$, $k = 1, \dots, T$, and $\hat{\sigma}_i^2$ is a suitable estimator for the long run variance of $y_{i,t}$. Finally, the standardized test statistic for series $y_{i,t}$ is computed as follows:

$$\hat{Z}_{i,T} = \frac{\hat{S}_{i,T} - \mu_{m_T}}{s_{m_T}} \quad (5)$$

with μ_{m_T} and s_{m_T} being suitable standardization factors.² It is readily checked (Landajo and Presno, 2013) that the null distribution of $\hat{Z}_{i,T}$ approaches the standard normal as T increases, whereas under H_1 the nonparametric panel test statistic diverges in probability to $+\infty$, so a consistent test statistic is readily obtained.

In the panel setting (2), we can test for the null of joint stationarity by using the following nonparametric panel stationarity (NPS, hereafter) test statistic:

$$\bar{Z}_T = \frac{\sum_{i=1}^N \hat{Z}_{i,T}}{N} \quad (6)$$

which is a simple average of the standardized nonparametric stationarity test statistics for each element of the panel.³

\bar{Z}_T is easily calculated once the scalar test statistics have been obtained for each component of the panel and, by construction, it is assured to have limiting power approaching 1 as T grows (for every fixed N). Unfortunately, the limiting null distribution of \bar{Z}_T is unknown excepting some especial cases⁴, though it can be readily bootstrapped, which renders a feasible test. In Section 2.2 below the details for the bootstrap implementation are included⁵. Section 2.3 summarizes the results of a Monte Carlo simulation study on the finite sample performance of the proposed test, showing that it performs suitably in realistic settings.

2.2. Bootstrap implementation of the NPS test

We shall assume that, for any $i = 1, \dots, N$ and some known $p_i < \infty$, the weakly stationary process $\{\varepsilon_{i,t}, t = 1, 2, \dots\}$ has the $AR(p_i)$ representation $\varepsilon_{i,t} = \sum_{k=1}^{p_i} \varphi_{ik} \varepsilon_{i,t-k} + v_{it}$. The following sequence is applied to implement the bootstrapped test:

Algorithm

1. Select the number of bootstrap resamples (B) and the complexity order (m_T). Set $b=1$.
2. For each $i = 1, \dots, N$, fit by OLS the model $y_{i,t} = \hat{\beta}_{i0} +$

² Namely, $\mu_{m_T} = \sum_{j=m_T+1}^{\infty} (j\pi)^{-2}$, $s_{m_T}^2 = 2 \sum_{j=m_T+1}^{\infty} (j\pi)^{-4}$, and $s_{m_T} = \sqrt{s_{m_T}^2}$.

³ A great many panel extensions (e.g., Carrion-i-Silvestre *et al.*, 2005) of classical stationarity and unit root tests are derived in this simple average fashion.

⁴ For instance, if the panel is composed of time series having independent random error processes, $\sqrt{N}\bar{Z}_T$ is approximately standard normal as T increases, for any fixed N .

⁵ Matlab codes are available from the authors upon request.

$$\sum_{j=1}^{m_T} \hat{\beta}_{ij} \cos\left(\frac{j\pi t}{T}\right) + e_{i,t}.$$

3. For each $i = 1, \dots, N$, fit by Yule-Walker (or OLS) the model $e_{i,t} = \sum_{k=1}^{p_i} \hat{\phi}_{ik} e_{i,t-k} + \hat{v}_{i,t}$ with p_i obtained by minimization of the Schwarz information Criterion (SIC). Compute $\hat{\sigma}_i^2$ and the observed test statistic \hat{Z}_{iT} . Compute \bar{Z}_T .

4. For each $i = 1, \dots, N$, generate centered residuals for the AR models:

$$\hat{v}_{i,t} = \hat{v}_{i,t} - (T - p_i)^{-1} \sum_{t=1+p_i}^T \hat{v}_{i,t}, \quad t = 1 + p_i, \dots, T.$$

5. Set bootstrap starting values (e.g., $\varepsilon_{i,0}^* = \dots = \varepsilon_{i,1-p_i}^* = 0$, or random values drawn from the sample distribution of $e_{i,t}$, $i = 1, \dots, N$).

6. Draw with replacement a random sample of observations $\mathbf{v}_t^* = (v_{1,t}^*, \dots, v_{N,t}^*)$, $t = 1, \dots, T$, from the sample distribution of vector $\hat{\mathbf{v}}_t = (\hat{v}_{1,t}, \dots, \hat{v}_{N,t})$.

7. For each $i = 1, \dots, N$, generate the pseudo series $\varepsilon_{i,t}^* = \sum_{k=1}^{p_i} \hat{\phi}_{i,k} \varepsilon_{i,t-k}^* + v_{i,t}^*$ and $y_{t,i}^* = \hat{\beta}_{i0} + \sum_{j=1}^{m_T} \hat{\beta}_{ij} \cos\left(\frac{j\pi t}{T}\right) + \varepsilon_{i,t}^*$, $t = 1, \dots, T$. Compute $\hat{\sigma}_i^{2*}$ (the bootstrap analogue for $\hat{\sigma}_i^2$). Compute \hat{Z}_{iT}^* and \bar{Z}_T^* (these are, respectively, the bootstrap analogues of \hat{Z}_{iT} and \bar{Z}_T).

8. Set $b = b + 1$ and repeat steps 5 to 7 while $b \leq B$.

9. Compute the bootstrap approximation for the critical value, namely:

$$p^B = B^{-1} \text{card}\{\bar{Z}_T^* > \bar{Z}_T\}$$

10. Reject H_0 if p^B is below the specified significance level (α).

□

Model complexity (m_T) is determined through the same kind of deterministic rules proposed in Landajo and Presno (2013), namely, $m_T = \lceil cT^{1/5} \rceil$, with c being some reasonable constant (further details are provided in Section 2.3).

In this paper we rely on a (roughly) parametric estimator for σ_i^2 , as we assume that an $\text{AR}(p_i)$ model provides an accurate approximation to the underlying data generating processes, and thus we estimate σ_i^2 parametrically. The maximum lag order in the above $\text{AR}(p_i)$ models is limited to $\max p = \lceil dT \rceil^{1/5}$, with $d = 1$.⁶

2.3. Monte Carlo analysis

In this section we analyze the finite sample performance (size and local power) of the proposed NPS test, first under i.i.d. errors and then the research will be extended to time series. Our computer-based experiment considers several trend models, panel sizes, sample sizes, and signal-to-noise ratios ($q = 0, 0.01, 0.1$).

For the case of i.i.d. errors, we considered data sets generated under model (1) above, with the following trend specifications:

(M1) $\theta^*(x) = \beta_0$,

(M2) $\theta^*(x) = \beta_0 + \beta_1 x$,

⁶ Other possibilities would involve use of fully nonparametric (spectral window) estimators for the long run variance of the error processes (e.g., the class considered in Pötscher and Prucha, 1991), although in our simulations the parametric approach provided more accurate results.

$$(M3) \theta^*(x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

$$(M4) \theta^*(x) = \beta_0 + \beta_1 x + \beta_3 [1 + \exp\{-\gamma_1(x - \omega_1)\}]^{-1},$$

with (independent uniformly distributed) randomly selected parameters $-2 < \beta_k < 2$, $k = 0, \dots, 3$; $0 < \gamma_1 < 100$, and $0.05 < \omega_1 < 0.95$; $0 < x < 1$. Models M1 and M2 are classical linear trend specifications. Model M4 incorporates a smooth transition in the level of the series (with ω_1 and γ_1 being, respectively, the relative position of the timing of the transition midpoint and the speed of transition –gradual for small values of γ_1 and approaching a break as that parameter increases). Finally, M3 is a quadratic model.

Throughout the simulation process, for each fixed model specification, each component of the panel and each Monte Carlo replication, the parameters of the trend model are randomly generated from independent uniform distributions with support on the above mentioned intervals. Hence, the trend parameters randomly change with each Monte Carlo replication and each component of the panel. We consider panel and sample sizes $N = 1, 5, 10, 20$, and $T = 100, 200, 300, 400$ (for the i.i.d. processes) and $T = 200, 400, 600$ (in the time series case), and signal-to-noise ratios $q = 0, 0.01$, and 0.1 . Cross-correlation is also allowed for. In model (1), the error term $\varepsilon_{i,t}$ allows for the presence of a common factor z_t , under the following form:

$$\varepsilon_{i,t} = \sqrt{1-\rho} \varepsilon'_{i,t} + \sqrt{\rho} z_t \quad (7)$$

where $\varepsilon'_{i,t}$ is the idiosyncratic random component (to be detailed below) and $\{z_t, t=1, \dots, T\}$ is an i.i.d. $N(0,1)$ process independent of $\{\varepsilon'_{i,t}, t=1, \dots, T\}$. Coefficient ρ allows us to incorporate cross-correlation. We consider the case cases $\rho=0, 0.5$ and 1 . The case $\rho=0$ corresponds to absence of cross-correlation, whereas under $\rho=1$ the random error processes of all the components of the panel coincide under the null hypothesis. The case $\rho=0.5$ is an intermediate, more realistic setting.

The test is conducted at 5% significance level. 1,000 Monte Carlo replications were generated for each case. In order to reduce computational complexity, the null distribution of the tests (and corresponding critical values) are approximated by using $B = 200$ bootstrap replications. The simulation analysis was implemented in Matlab.

2.3.1. Simulation results for i.i.d. processes.

In this case the idiosyncratic component $\{\varepsilon'_{i,t}, t=1, \dots, T\}$ is an i.i.d. $N(0,1)$ process. A summary of results is provided in Table 1 below.⁷

Overall, results indicate that the empirical size of the test (at rows $q = 0$) is close to the nominal (5%) level. Seemingly, test size is roughly unaffected by changes in model

⁷ Model complexity in the i.i.d. setting is determined through the rule $m_T = \lceil 5T^{1/5} \rceil$.

specification, series length (T), the number of series in the panel (N), and cross-correlation intensity (ρ). As for the power of the test, as expected it increases with q (the signal-to-noise ratio), N , and T . For fixed q , T , and N , we do not observe significant differences in the power of the test for the various trend specifications considered. As to the effect of ρ , power clearly decreases as cross-correlation intensity rises.

Table 1. Empirical size ($q = 0$) and local power ($q > 0$) of the NPS test under i.i.d. processes. 1,000 Monte Carlo replications.

N	T	q	$\rho=0$				$\rho=0.5$				$\rho=1$			
			M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
1	100	0					4.8	3.7	5.9	6.1				
1	100	0.01					6.1	5.4	6	6.5				
1	100	0.1					17.5	17.5	14.8	18.1				
1	200	0					4.7	3.4	4.6	5.9				
1	200	0.01					9.4	8.5	7.9	8.6				
1	200	0.1					54.9	56.3	54.4	54.9				
1	300	0					4.9	4.1	5.8	6.1				
1	300	0.01					14.3	15.8	15	15.6				
1	300	0.1					87.5	89.2	86.6	89.1				
1	400	0					5.3	5.3	5	6.6				
1	400	0.01					23.2	21.2	20.1	22.1				
1	400	0.1					97.8	98.1	98	98.5				
5	100	0	4.4	4.2	3.7	5.1	5.9	4.4	4.1	4.8	6.2	4.3	5.1	5.9
5	100	0.01	6.7	6.2	6	7.1	5.6	6.6	6.6	7	5.9	5.7	4.8	5.6
5	100	0.1	36.3	38.2	40	37.6	26.1	23.5	22.6	26.3	14.2	15.8	14.5	15.4
5	200	0	4.4	4.2	4.6	5	4.8	4	5.5	4.8	5.1	4.9	3.5	5.5
5	200	0.01	12.8	12.5	15.6	17	9.7	9.3	10.2	12.2	7.4	5.9	7.2	7
5	200	0.1	98.5	98.2	97.9	97.7	94.1	89.5	92.3	91.5	66	70.4	70.1	69.4
5	300	0	4.7	5.5	5.4	7.2	4.7	6	5.5	4.9	5.7	5.5	5.8	5.9
5	300	0.01	32.9	34.1	31.4	33.5	20	18	19	20.6	12.5	12	11.7	12
5	300	0.1	100	100	100	100	99.9	100	99.9	100	99.1	99.5	98.9	99.8
5	400	0	5.6	5.6	5.3	6.3	4.4	5.4	6.2	7	3.9	3.6	4.9	4.3
5	400	0.01	57.1	58	54.3	56.3	34.2	33.5	34.5	39.8	19.5	19.2	17.1	20.1
5	400	0.1	100	100	100	100	100	100	100	100	100	100	100	100
10	100	0	3.2	4.1	4.1	5.6	4.7	5	5.6	5.6	4.4	5.4	4.5	5.6
10	100	0.01	5.4	6.6	6.8	9.6	6.3	5.1	6.8	8.6	5.1	5.4	5	6.6
10	100	0.1	57.3	56.8	54.5	63.5	28.1	28.9	30.8	30.1	15.4	15.9	11.8	14.9
10	200	0	4.2	4.3	4.7	5.8	5.2	6	5.8	5.4	3.7	4.6	4.2	5.5
10	200	0.01	18.9	17	20.3	22.7	9.5	9.5	9.6	11.4	7.7	8.1	5.9	7.8
10	200	0.1	99.9	99.9	99.9	100	97.5	97.1	98.1	98.4	70.8	69.6	71.6	71.7
10	300	0	3.6	5.1	6.1	6.1	5	3.6	4.5	6.6	5	4.8	4	7
10	300	0.01	49.6	47.6	46.9	50.6	21.1	21	21.6	23.2	11.8	10	9.7	12.7
10	300	0.1	100	100	100	100	100	100	100	100	99.9	100	100	100
10	400	0	5.8	5.2	4.6	7.6	4.3	4.2	4.1	6.3	4.1	4.9	4.6	5.6
10	400	0.01	79.6	79.6	77.1	80.9	39.5	38.1	39.5	40.3	19.4	18.6	19.2	18.3
10	400	0.1	100	100	100	100	100	100	100	100	100	100	100	100
20	100	0	4.1	4.3	3.7	6.1	4.3	5.5	5.8	5.5	4.1	4.5	2.1	5.9
20	100	0.01	8.7	5.6	7	9.5	5.1	5.3	6.2	6.5	4.8	6.5	4.6	5.2
20	100	0.1	80	81.1	80.8	85.8	31.3	29.9	30.3	37.3	14.1	16.2	14.4	12.8
20	200	0	3.8	3.7	4.4	5.8	3.6	4.5	4.3	5.8	4.6	4.3	5.4	5.6
20	200	0.01	28.4	24.6	27.9	32.8	8.7	9.5	11.4	11.4	7.4	8.3	7.5	8.1
20	200	0.1	100	100	100	100	99.5	99.5	99.1	99.8	73.7	72.3	74.1	74.3
20	300	0	5.3	4.8	3.2	9.1	5	5.4	6.4	6.5	5.8	6	4.3	4.9
20	300	0.01	67.3	69.4	68.9	74.8	22.1	21.6	22.7	24.4	12.5	9.5	11.6	14.1
20	300	0.1	100	100	100	100	100	100	100	100	99.9	100	100	99.9
20	400	0	5.3	4.9	4.3	7.9	5.2	5.3	3.8	4.9	3.9	5.8	5.1	5.4
20	400	0.01	97	96.6	97.2	96.2	44	41.7	44.4	44.5	18.1	17.2	18.9	18.6
20	400	0.1	100	100	100	100	100	100	100	100	100	100	100	100

2.3.2. Time series simulation results

Then the above study was extended in order to allow for serial dependence. We consider the idiosyncratic component of the series generated under the following time series models, for any $i = 1, \dots, N$:

(I) AR: $\varepsilon'_{i,t} = \varphi_{i1}\varepsilon'_{i,t-1} + v_{i,t}$, with $-0.8 \leq \varphi_{i1} \leq 0.8$,

(II) MA: $\varepsilon'_{i,t} = \delta_{i1}v_{i,t-1} + v_{i,t}$, with $-0.8 \leq \delta_{i1} \leq 0.8$,

(III) ARHET: $\varepsilon'_{i,t} = \varphi_{i1}\varepsilon'_{i,t-1} + \sqrt{\pi_i}v_{i,t}$, with $-0.5 \leq \varphi_{i1} \leq 0.5$,

$\pi_i = \delta_{i0} + \delta_{i1}(\varepsilon'_{i,t-1})^2 + \delta_{i2}(\varepsilon'_{i,t-1})^2$, $0 < \delta_{ij} < 0.4$, $j = 0, 1, 2$,

(IV) BIL: $\varepsilon'_{i,t} = \varphi_{i1}\varepsilon'_{i,t-1} + \varphi_{i2}\varepsilon'_{i,t-2}v_{i,t-1} + v_{i,t}$, with $-0.4 \leq \varphi_{ij} \leq 0.4$, $j = 1, 2$,

(V) NLMA: $\varepsilon'_{i,t} = \delta_{i1}\varepsilon'_{i,t-1} + \delta_{i2}v_{i,t-1}v_{i,t-2} + v_{i,t}$, with $-0.4 \leq \delta_{ij} \leq 0.4$, $j = 1, 2$,

with the components of the basis process $\{v_{i,t}, i = 1, \dots, N; t = 1, \dots, T\}$ being a sequence of independent $N(0,1)$ random variables. In the simulations, for each Monte Carlo replication and each $i = 1, \dots, N$, coefficients φ_{ij} and δ_{ij} are drawn at random from independent uniform distributions with support on the above intervals.

Models I and II above are examples of classical (respectively, AR and MA) linear time series models, whereas the remaining specifications will allow us to analyse the performance of the NPS test in nonlinear settings (Model III is an AR specification with heteroskedastic errors; Model IV a bilinear time series; Model V a nonlinear MA time series).

In order to calculate the long run variance estimators $\hat{\sigma}_i^2$, AR processes were fitted separately to each residual series in the panel, with AR complexity determined by the SIC, with maximum lag order⁸ set at $maxp = \lceil T^{\frac{1}{5}} \rceil$.

Results (under the smooth transition trend specification, M4, with cross-correlation intensity fixed at $\rho=0.5$) are reported in Table 2.⁹ As in the i.i.d. case, the power of the test increases with q , N , and T although -as expected- serial dependence reduces the power of the test with respect to the i.i.d. case.¹⁰

⁸ For the sake of simplicity, the highest lag order selected by the SIC is used simultaneously to fit all the residual series.

⁹ Results for the other trend specifications and cross-correlation levels -omitted for brevity- are similar to those reported here.

¹⁰ The deterministic rule $m_T = \lceil 4T^{1/5} \rceil$ is used for model complexity in the time series setting.

Table 2. Empirical size ($q = 0$) and local power ($q > 0$) of the NPS test under several time series models. 1,000 Monte Carlo replications. Model M4. $\rho=0.5$.

N	T	q	AR model	MA model	ARHET model	BIL model	NLMA model
1	200	0	12.8	13.7	6	0.9	2.4
1	200	0.01	14.4	15.5	6.1	0.2	0.8
1	200	0.1	23.9	25.5	5.4	1	0.5
1	400	0	5.3	4.4	3.1	0.2	0.4
1	400	0.01	11.3	7.2	3.1	1.7	1.6
1	400	0.1	22.1	20.4	11	5.4	4.7
1	600	0	9.4	2.5	1.7	0.2	1
1	600	0.01	19	13	11.9	3.2	4.5
1	600	0.1	40.1	38.2	27.9	15.7	16.9
5	200	0	5.1	6.5	2.4	3	3.1
5	200	0.01	7.4	8.6	6	2	3.6
5	200	0.1	31.9	51.8	20.4	18.3	18.5
5	400	0	5.6	5.6	4.9	2.8	2.8
5	400	0.01	19.5	22.6	18.7	12.3	11.7
5	400	0.1	67.6	73	59.6	67.2	67.5
5	600	0	8.6	7.3	6.2	4.5	3.8
5	600	0.01	55.5	62.9	63.3	46.9	48.3
5	600	0.1	96.2	98.8	90.9	98.1	97.7
10	200	0	2.5	4	3.2	2.1	2.6
10	200	0.01	5.7	7	3.9	4.3	3.5
10	200	0.1	33	59.3	30	30	30.1
10	400	0	5.8	3.2	5.3	3.7	4.1
10	400	0.01	20	26.2	22.1	15.4	17.3
10	400	0.1	84.5	91.4	65.7	86.1	85.6
10	600	0	8.2	7.1	6.2	6.3	7
10	600	0.01	69.3	79.5	76.9	70.3	68.2
10	600	0.1	99.9	100	98.8	100	100
20	200	0	2.2	2.2	3.7	3	2
20	200	0.01	3.1	5.7	5.1	2.6	3.2
20	200	0.1	31.6	64	30.9	39.6	40.6
20	400	0	3.9	3.7	6.1	3.7	4.1
20	400	0.01	20.7	24.5	23.4	20.2	20
20	400	0.1	95.4	98.7	76.4	96.1	97
20	600	0	7.8	6.9	7.5	6.4	4.2
20	600	0.01	80.5	88.1	81.6	77	76.4
20	600	0.1	100	100	99.9	100	99.7

3 Empirical analysis

The time series to be analyzed are the logged monthly values of the production of crude oil including lease condensate (in thousand barrels of oil per day), for the period between January 1973 and December 2015 (so the total number of observations is 516 for each country). The source of the data is the Energy Information Administration (EIA) of the U.S. Department of Energy. We considered 12 OPEC members (Algeria, Angola, Ecuador, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates -UAE-, and Venezuela) and 8 non-OPEC countries (Canada, China, Egypt, Mexico, Norway, Russia, the United Kingdom -UK-, and the United States -US-). This list of countries – amounting to 82.7% of world oil production in 2015- is similar to that considered by Maslyuk and Smyth (2009). As in that paper, we do not adjust for seasonality in the time series under study, since the seasonal pattern was not as strong as that observed in the oil consumption series and the effects of seasonal filters on the test have not been researched up to date.

In this Section we test for trend stationarity of the oil production series. Three cases will be considered: (i) first we test for stationarity of aggregate oil production (considering successively three aggregates: global, OPEC, and non-OPEC). Then, (ii) we apply panel stationarity testing to the panel of the 20 countries, considering also subpanels of OPEC and non-OPEC states. Finally, (iii) we undertake a detailed research on stationarity of oil production in each country separately.

Step (ii) in the analysis will be carried out by resorting to the NPS test proposed in Section 2 above, whereas steps (i) and (iii) (which only involve separate analysis of individual series) will be implemented by relying on a bootstrapped version of the nonparametric LP stationarity test. Technically, as the latter is an adaption of the general NPS test to panels including a single series, steps (i) and (iii) do not increase the conceptual/computational complexity of the study (yet, according to the power analysis in Section 2.3 below, the panel test in (ii) may be expected to have higher power to detect departures from stationarity than the single-series version of the test as applied in steps (i) and (iii)). In addition, simultaneous stationarity for all the individual series in (iii) would imply panel stationarity in (ii) and stationarity of the aggregates in (i), so the conclusions of those three analyses should tend to be mutually consistent.

It is remarked in literature both the non-linear character of the series (tested by Maslyuk and Smyth, 2009) and the presence of breaks and outliers (many of them detected by Barros *et al.*, 2011 for the OPEC states). As commented above, the model-free nature of the analysis liberates researchers from the need of prior, correct specification of functional forms for the trend function in each component of the panel.

3.1. Aggregate oil production and panel analysis

Focusing on stationarity analysis for the aggregates (total oil production of the 20 countries, as well as separate totals for the 12 OPEC states and 8 non-OPEC countries), the nonparametric LP stationarity test leads us (see Table 3) to reject the null of stationarity for both the aggregate production of all the countries and that of the non-OPEC group but not for the OPEC aggregate, implying that shocks affecting the latter aggregate would be transitory in nature and tend to vanish in the long run. According to these results, in the event of any exogenous shock, stronger policy measures must be applied

to non-OPEC countries than to the OPEC ones in order to return their respective production series to their original trends.

However, as pointed out by Yang (2000), aggregate data do not fully capture the variability in the grade that countries depend on energy. In addition, it is well known that single-series stationarity tests may exhibit low power in short series. This led us to rely on panel stationarity testing in order to obtain more powerful results. Table 3 reports the observed test statistics and the conclusions of panel stationarity testing. Beginning with the global panel of 20 countries, the null of stationarity is rejected at 1% significance, suggesting that shocks would have permanent effects on oil production. Separate analysis for the OPEC and non-OPEC subpanels let us reject again the null of stationarity at 5% significance for both subpanels, but not at the 1% level for the group of OPEC countries.

Table 3. Analysis for aggregate oil production (LP test) and panel (NPS test)

Aggregate	Obs. LP test statistic (p-value)	Panel	Obs. NPS test statistic (p-value)
Total oil production	3.205 ^b (p=0.000)	All the countries	6.566 ^b (p=0.000)
Total OPEC production	-0.535 (p=0.614)	OPEC countries	3.409 ^a (p=0.032)
Total Non- OPEC production	3.205 ^b (p=0.000)	Non-OPEC countries	6.206 ^b (p=0.000)

^{a, b} denote significance at 5% and 1% level, respectively.

3.2. Country-level stationarity analysis

The above results suggest that it would also be convenient to test for stationarity individually for each of the 20 oil production series, in order to detect specific countries that may be responsible for aggregate and panel non-stationarity. The results of the individual (country) tests are reported in Table 4 below. The null of stationarity is rejected at 1% significance for Canada, China, Mexico, and the US (among the non-OPEC countries), as well as for Algeria, Iran, Nigeria, and Qatar (from the OPEC group). Therefore, the percentage of series for which the null of trend stationarity is rejected seems to be higher in the group of non-OPEC states (50%) than among OPEC countries (33.33%).

Comparing the above results with those of previous studies that applied different techniques, we have on one hand Narayan *et al.* (2008), who -for a panel of 60 countries- report that an LM linear panel-unit-root test with a single structural break provides strong evidence that crude oil production does not contain unit roots. On the other end, Maslyuk and Smyth (2009) -for a group of countries similar to that examined in this paper, and using unit root tests in a non-linear framework- find that all the countries have a unit root in at least one of the regimes examined. They justify differences in results on the basis of the assumption of linearity in Narayan *et al.* (2008), which might be somewhat restrictive. Barros *et al.* (2011) - in their analysis for OPEC countries, using fractional integration modelling and incorporating breaks and outliers- find that shocks affecting OPEC oil production have a high degree of persistence in the long run,

but a unit root is only present in some cases. Therefore, depending on the trend specification considered, different conclusions have been reached in literature. The main advantage of nonparametric testing as implemented in this paper resides in its flexibility as prior, correct specification of the trend functions is not required, so more robust results can be expected.

Table 4. Nonparametric stationarity analysis for country oil production. LP test.

Countries	Obs. test statistic (p-value)
Algeria	3.300 ^b (p=0.002)
Angola	-1.787 (p=0.992)
Ecuador	-0.725 (p=0.862)
Iran	3.332 ^b (p=0.002)
Iraq	0.936 (p=0.082)
Kuwait	-1.859 (p=0.986)
Libya	-1.408 (p=0.996)
Nigeria	4.087 ^b (p=0.000)
Qatar	3.027 ^b (p=0.004)
Saudi Arabia	0.413 (p=0.500)
UAE	-0.633 (p=0.896)
Venezuela	0.611 (p=0.478)
Canada	6.482 ^b (p=0.000)
China	5.409 ^b (p=0.000)
Egypt	-0.130 (p=0.804)
Mexico	5.636 ^b (p=0.000)
Norway	0.086 (p=0.598)
Russia	0.422 (p=0.228)
UK	0.076 (p=0.402)
US	5.352 ^b (p=0.000)

^{a, b} denote significance at 5% and 1%, respectively.

4 Concluding remarks

Traditional unit root and stationarity tests are not robust to misspecification of the trend components of the series, potentially leading to spurious results. In this paper we have proposed a nonparametric approach that bypasses that limitation. The advantage of the proposed tests resides in their remarkable flexibility as they free researchers from the need of correct specification of the trend function for each component of the panel, so that far more robust results -not depending on trend specification- can be expected.

Extensive Monte Carlo evidence reported in this paper (considering several trend specifications and various stochastic structures -including both i.i.d. and time series data-) indicates that the empirical size of the test is close to the nominal level and is roughly unaffected by such factors as changes in model specification, series length, the number of series in the panel, and cross-correlation intensity. As for the power of the

test, as expected, it increases with the signal-to-noise ratio, series length, and the number of series in the panel, with no significant differences observed for the various trend specifications considered.

We have applied the tests to analyze stationarity of monthly crude oil production in the period between years 1973 and 2015 for a panel of 20 OPEC and non-OPEC countries. Our analysis indicates that the null of stationarity is strongly rejected (at 1% significance) for both the panel of 20 countries and the subpanel of 8 non-OPEC nations. For the subpanel of 12 OPEC members the evidence is slightly weaker, with stationarity being rejected at 5% but not at 1% significance. These results suggest that disruptions in crude oil production may have permanent effects, with other macroeconomic variables also inheriting the characteristic that might thus spill over throughout the economy.

We have completed panel analysis with a case-by-case stationarity study for the individual oil production series, with a view to obtain a deeper understanding of specific factors that may have motivated rejection at the panel level. The null of stationarity is rejected for Canada, China, Mexico, and the US (among non-OPEC members), and for Algeria, Iran, Nigeria, and Qatar (in the OPEC group), with a higher percentage of non-stationary countries in the block of non-OPEC nations. Therefore, in order to explain these results, *OPEC membership* would also be a factor to be considered, mainly because of the greater coordination capacity of that organization in order to influence the oil market.

References

1. Barros, C.P., Gil-Alana, L.A., Payne, J., 2011. An analysis of oil production by OPEC countries: persistence, breaks and outliers. *Energy Policy* 39, 442-453.
2. Carrion-i-Silvestre, J.Ll., del Barrio, T., López-Bazo, E., 2005. Breaking the Panels: An Application to GDP per capita. *Econometrics Journal* 8, 159-175.
3. Energy Information Administration.
4. Landajo, M., Presno, M.J., 2010. Stationarity testing under nonlinear models. Some asymptotic results. *Journal of Time Series Analysis* 31, 392-405.
5. Landajo, M., Presno, M.J., 2013. Nonparametric pseudo-Lagrange multiplier stationarity testing. *Annals of the Institute of Statistical Mathematics* 65, 125-147.
6. Maslyuk, S., Smyth, R., 2009. Non-linear unit root properties of crude oil production. *Energy Economics* 31, 109-118.
7. Narayan, P.K., Narayan, S., Smyth, R., 2008. Are oil shocks permanent or temporary? Panel data evidence from crude oil and NGL production in 60 countries. *Energy Economics* 30, 919-936.
8. Pötscher, B.M., Prucha, I.R., 1991. Basic structure of the asymptotic theory in dynamic nonlinear econometric models. Part II: Asymptotic normality. *Econometric Reviews* 10, 253-325.
9. Yang, H.Y., 2000. A note on the causal relationship between energy and GDP in Taiwan. *Energy Economics* 22, 309-317.

Detection of temperature break point for gas storage

Andrzej Szczurek, Andrzej Kielbik, and Monika Maciejewska

Wroclaw University of Science and Technology, Faculty of Environmental Engineering,

Wbrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

{andrzej.szczurek@pwr.edu.pl

Gas Engineering Office Gazoprojekt SA,

55a Strzegomska Street , 53-611 Wroclaw Poland

akielbik@gazoprojekt.pl

Wroclaw University of Science and Technology, Faculty of Environmental Engineering,

Wbrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland

monika.maciejewska@pwr.edu.pl

Abstract. Natural gas is one of the most important fuels of the future. However, its supplies and demand are typically imbalanced. It is due to the seasonality of consumption. The lack of match between the supply and demand provides justification for gas storage. In this work, there was proposed a method of detecting the temperature break point. In the period of year when ambient temperature is higher than the break point, there exist favorable circumstances for gas storage. The temperature break point was defined as the daily ambient temperature, t when the probability, $p_1(t)$ that daily gas consumption exceeds the supply equals the probability, $p_2(t)$ that daily gas consumption is smaller than the supply. The detection method is based on the analysis of the dependence structure between ambient temperature and gas consumption. The two probabilities $p_1(t)$ and $p_2(t)$ were derived from the model of the bivariate dependence. Copula is an effective tool of studying and modeling the multivariate dependence and this concept was applied in our work. The temperature break point method was tested using real data set, which referred to an exemplary year. For this data, the temperature break point for gas storage was found equal 8 ± 0.5 °C. The obtained estimate was highly realistic. The presented method may have important policy implications.

Keywords: natural gas, copula, change-point detection

1 Introduction

Natural gas is one of the cleanest (environmentally-friendly), safest, and the most diversified fuels. It primarily serves thermal energy generation for heating buildings and homes, but also for electric power generation and cooling. Therefore, the demand for natural gas is highly seasonal and it is heavily dependent

on how the temperature fluctuates with reference to the time of the year and the time of the day. Traditionally, gas consumption increases greatly in winter. However, recently there is observed a growing summer peak demand. It is due to the electricity consumption for powering air conditioners and the like.

In winter, the demand can be as much as 6 times higher than in summer. Therefore, the accumulated surplus from the summer must be stored in order to meet the higher demand in winter. In other words, shippers want to inject natural gas into storage when the demand is low - historically in the summer - and withdraw it during times of high demand - generally, to meet the peak heating demands in winter. Natural gas from storage accounts for about 20 % of the natural gas consumed in winter. Shippers now sometimes use gas from storage in the summer as well, to meet gas-fired electric generation needs. Natural gas storage enables supply to match the demand on any given day throughout the year by adjusting to daily and seasonal fluctuations in demand while natural gas production remains relatively constant year-round [1]. Natural gas in storage also serves as insurance against any unforeseen accidents, market speculation (reducing price volatility), natural disasters such as e.g. hurricanes, or other occurrences that may affect the production or delivery of the natural gas. Storage facilities are ones of the tools needed to increase the energy security. They ensure, to some extent, the reliability of gas supply to the consumer at the lowest cost, as required by the regulatory body.

Natural gas can be stored for an indefinite period of time in natural gas storage facilities for later consumption. The most important type of gas storage is in underground reservoirs [2]. There are three main types of underground storage: depleted gas reservoirs, aquifer reservoirs and salt cavern reservoirs. Each of these types has distinct physical and economic characteristics which govern the suitability of a particular type of storage for a given application. The most prominent and common form of underground storage consists in re-filling depleted gas reservoirs. They are generally the cheapest and easiest to develop, operate and maintain, compared with two other types of underground storage. Depleted gas reservoirs are used to meet base load requirements (seasonal demand increases). They are capable of holding enough natural gas to satisfy long term seasonal demand requirements. Typically, these facilities are operated on a single annual cycle; gas is injected into storage during periods of low demand (non-heating season, which usually runs from April through October) and withdrawn from storage during periods of peak demand (heating season, usually from November to March). It means, that depleted gas reservoirs have long term injection and withdrawal seasons.

For economic and technical reasons (like pipeline capacity, physical limits on existing storage capacity), it is important to predict moments when gas should be injected during the off-peak summer months and withdrawn during the winter months of peak demand. This type of information can be extracted from time series of appropriate variables using stochastic methods of data analysis.

Time series reflect the stochastic nature of most variables over time. There are two main goals of time series analysis: (1) identifying the nature of the

phenomenon represented by the sequence of observations, and (2) forecasting, i.e. predicting future values of the variable in time. In both cases, it is required to identify the pattern in the time series data and to describe it, more or less formally.

Usually, time series analysis refers to trend, seasonal and cyclic patterns. Regardless of the depth of understanding and the validity of interpretation of the processes influencing natural gas storage possibilities, we can extrapolate the identified pattern to estimate the conditions when particular events occur and to predict future events.

In this work, there was explored the pattern of the dependence structure between ambient temperature and natural gas consumption. The analysis was aimed at estimating the temperature break point. At temperatures lower than the break point, gas consumption dominates over supply. Otherwise, the situation is the opposite. Thus, temperature break point indicates conditions which are favorable for gas storage.

Copula offers a way of describing the dependence between random variables. Due to abundance of multivariate problems, which require this kind of analysis copulas are more and more willingly used in various areas of science and technique, in particular in finances [3] and environmental science [4, 5]. The approach is also under constant development from theoretical point of view [6]. Copula is the principal element of the presented method of temperature break point detection.

2 Methods

In this work there is presented a method for detecting the temperature break point with respect to gas usage. It allows to determine the period of time, which is most appropriate for gas injection to the underground storage.

The method is based on the analysis of the bi-variate time series. The considered variables are ambient air temperature and gas consumption.

Let $U(t)$ be the daily natural gas usage, expressed as a function of temperature. Let S be the daily gas supply. This variable is independent of temperature. We additionally assume it takes constant value in time.

There are considered two functions of temperature $p_1(t)$ and $p_2(t)$, where $p_1(t)$ represents the probability that real daily gas consumption at particular temperature t is greater than the supply

$$p_1(t) = p(U(t) > S) \quad (1)$$

and $p_2(t)$ represents the probability that real daily gas consumption at particular temperature t is smaller than the supply

$$p_2(t) = p(U(t) < S). \quad (2)$$

The temperature break point is defined as the temperature for which the values of two functions are equal, $p_1(t) = p_2(t)$. In terms of graphical interpre-

tation the temperature break point is the value of argument where the plots of two functions $p_1(t)$ and $p_2(t)$ intersect.

The basis for determining $p_1(t)$ and $p_2(t)$ is the bi-variate distribution of daily temperature and gas consumption.

It was proposed to evaluate the dependence structure of the two variables using copulas [7]. Copulas are functions that describe dependencies among variables and provide a way to create distributions which model correlated multivariate data [8]. This approach has the capability of capturing the complex nonlinear multivariate relationship between parameters.

The copula approach for dependence modeling is based on the Sklar's theorem [7]. It states that when C is a d -variate copula and F_1, \dots, F_d are univariate cumulative distribution functions (cdf-s), then the function

$$F(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d)) \quad (3)$$

is a d -variate cdf with margins F_1, \dots, F_d . In other words, $F(\mathbf{x})$ is the joint distribution with marginal distributions F_1, \dots, F_d .

The principal value of copula concept consists in the fact that it allows for separation of two major characteristics of multivariate data. The first is the individual behavior of variables x_1, \dots, x_d . It is represented by marginal distributions F_1, \dots, F_d . The second is the dependence structure of these variables. It is captured by the copula function, C . One should notice that F_1, \dots, F_d may be viewed as uniform random variables u, \dots, v . Because of fundamental properties of cumulative distribution function, the range of values for these variables is always $[0, 1]$. In the direct manner, copula operates on the variables u, \dots, v and it is insensitive to the underlying distributions F_1, \dots, F_d , which link them with the variables x_1, \dots, x_d .

There are available a number of parametric copula functions. Here belong for example: Gauss copula, t-Student copula and a family of Archimedean copulas, including Clayton, Gumbel and Frank ones. The essential step in applying these copulas for modeling the dependence structure of the experimental data is to determine the degree of association between variables u, \dots, v . Whole range of correlation coefficients may be used for this purpose, but rank correlation estimates (Kendall, Spearman) are favored.

In this work there was considered copula-based joint probability model for the bi-variate problem

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)) \quad (4)$$

where: x_1 is the daily ambient air temperature and x_2 is the daily average natural gas consumption, $F(x_1)$ and $F(x_2)$ the respective cumulative distribution functions. The joint distribution may also be expressed as

$$F(u, v) = C(u, v) \quad (5)$$

where variables u and v are defined as follows $u := F_1(x_1)$ and $v := F_2(x_2)$.

Frank copula was found most appropriate for modeling the examined bi-variate data. Frank copula has the following analytical function

$$C_{\alpha}(u, v) = -\frac{1}{\alpha} \ln \left[1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{(e^{-\alpha} - 1)} \right] \quad (6)$$

where α is the Frank copula parameter. It is a measure of association between variables u and v , based on Kendall's rank correlation.

A two-step procedure was implemented to evaluate the joint distribution of x_1 and x_2 . It consisted in marginal modeling and copula parameter estimation.

The criterion for the quality of copula fitting was the root mean square error, which represented the distance between the empirical distributions of the measurement data and the data generated when using the particular copula function.

Joint cumulative distribution $F(x_1, x_2)$, obtained from copula modeling, was the basis for estimation of joint probability density function $f(x_1, x_2)$. For this purpose we applied the grid with the predefined resolution Δx_1 in the dimension x_1 and Δx_2 in the dimension x_2 . The following formula was applied

$$f\left(x_1 + \frac{\Delta x_1}{2}, x_2 + \frac{\Delta x_2}{2}\right) = F(x_1 + \Delta x_1, x_2 + \Delta x_2) - F(x_1, x_2 + \Delta x_2) - F(x_1 + \Delta x_1, x_2) + F(x_1, x_2) \quad (7)$$

where: $x_1 = x_{1min} : \Delta x_1 : x_{1max}$ and $x_2 = x_{2min} : \Delta x_2 : x_{2max}$.

Via domain quantization the problem was moved from the continuous to discrete domain. The following was assumed about probability that temperature and gas consumption take particular discrete values:

$$p\left(X_1 = x_1 + \frac{\Delta x_1}{2}, X_2 = x_2 + \frac{\Delta x_2}{2}\right) = f\left(x_1 + \frac{\Delta x_1}{2}, x_2 + \frac{\Delta x_2}{2}\right) \Delta x_1 \Delta x_2. \quad (8)$$

Therefore, functions $p_1(t)$ and $p_2(t)$ are directly expressed as follows

$$p_1(t) = p(X_1 = t, X_2 > S) \quad (9)$$

and

$$p_2(t) = p(X_1 = t, X_2 < S) \quad (10)$$

and they may be easily computed.

All calculations were carried out using Matlab [8].

3 Data

Two major assumptions were made regarding analyzed data. They refer to the temporal resolution and the time span.

1. In most countries, the gas consumption is publicly reported in terms of daily average consumption, while the specialized agencies have access to higher resolution data. We demonstrated the performance of the method using the bi-variate time series of daily ambient air temperature and daily gas consumption. Hence, in the presented analysis, the break point is the average daily temperature. In principal, the method may be applied to data featured by higher temporal resolution. Mind, the status of break point would change in such case e.g. this could be one hour average. Using lower resolution data is not recommended, because the sufficient size of data set is necessary to study the dependence structure.
2. It is required that data describes the time period of at least one year. This requirement is justified by the cyclic character of ambient temperature variation and gas consumption variation. In both cases the period of the basic cycle is one year.

Further, we present the results of the analysis performed, with the proposed method, for the exemplary data set. It consists of daily ambient air temperatures and daily natural gas consumption in Poland, throughout the year 2006. The temperature was quantized with the resolution of 0.5°C and the gas consumption was quantized with resolution of $0.5 \text{ mln m}^3/\text{h}$. Average daily gas supply was computed equal $34.74 \text{ mln m}^3/\text{h}$.

4 Results

The time series of daily ambient air temperature and natural gas consumption in Poland, in 2006, are shown in Fig. 1. It displays a typical behavior of the two variables in time. Namely, the temperature is low in winter and high in summer. Contrarily, gas consumption is high during winter and small in summer. The transition between large and small gas consumption takes place in spring, while at the same time of year ambient temperature climbs from its lowest towards highest values. The opposite behavior of both variables is typical for autumn. As shown in Fig. 2 the two variables are negatively correlated.

The dependence structure between the temperature and gas consumption displayed in Fig. 2 was evaluated using copula concept. Several parametric copulas were considered for empirical data fitting, namely Gauss, t-Student and all members Archimedes family. In Table 1 there is shown the root mean square error of fitting. It represents the difference between the empirical joint distribution (relative frequency histogram) and the joint distribution obtained from the data simulated using copulas functions.

As shown in Table 1, Frank copula best represented the correlation structure between the examined variables.

The bi-variate cumulative distribution of temperature and gas consumption was computed using the analytical form of Frank copula. The joint distribution is shown in Fig. 3. The Frank copula parameter that controls the strength of dependence was estimated as $\alpha = -20.84$.

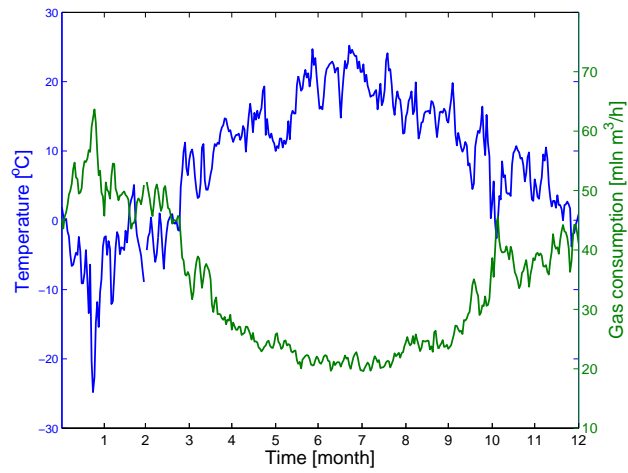


Fig. 1. Time series of daily ambient air temperature and daily natural gas consumption in Poland, in 2006. Time axis starts at 1.01.2006. 00:00

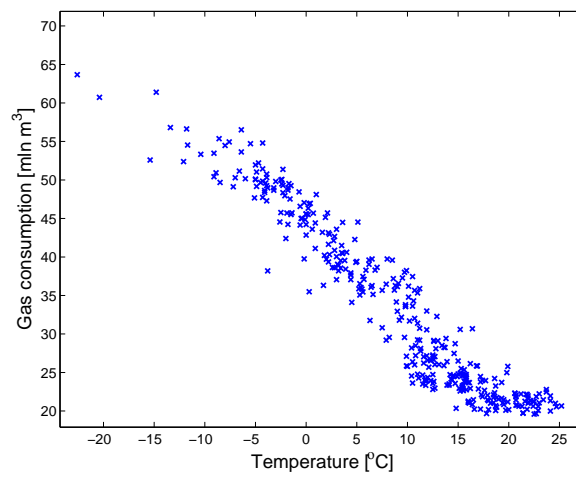


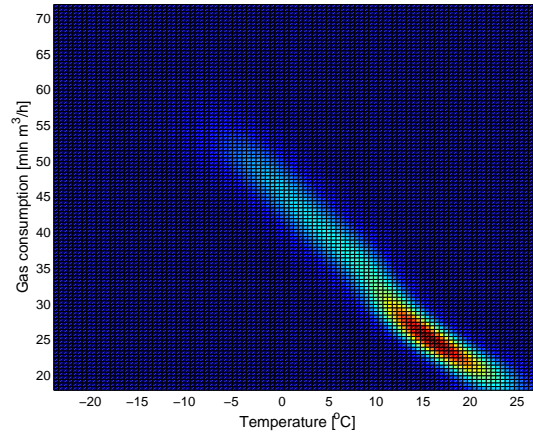
Fig. 2. Relationship between daily ambient temperature and daily natural gas consumption in Poland, in 2006.

Table 1. RMSE of fitting empirical bivariate distribution with a selection of parametric copula functions.

Copula function	RMSE
Gauss	0.017
t-Student	0.016
Clayton	0.020
Gumbel	0.045
Frank	0.015

From the comparison of Fig. 2 and Fig. 3 it may be concluded that the bivariate distribution which results from copula modeling well accounts for the dependencies in the empirical data. As a result, it successfully retrieves the true underlying distribution. The copula function parameter confirms negative correlation of ambient temperature and gas consumption. Additionally, the function itself also reveals the nonlinear character of the relationship. As shown in Fig. 2 and Fig. 3, in general, gas consumption decreases upon ambient temperature increase. However, at high temperatures, over approximately 10°C the decrease rate is smaller as compared to lower temperature range, under 10°C. The analysis discloses that in warm period of year, gas consumption is less temperature-dependent.

Fig. 4 presents the temperature dependency of two probabilities $p_1(t)$ and $p_2(t)$. The first is the probability that real daily gas consumption is higher than the supply, and the second is the probability that real daily gas consumption is lower than the supply. They were calculated based on the bivariate distribution, which shown in Fig. 3, as explained in Section 2.

**Fig. 3.** Bivariate distribution of daily ambient air temperature and daily natural gas consumption in Poland in 2006, estimated by Frank copula, $\alpha=-20.84$

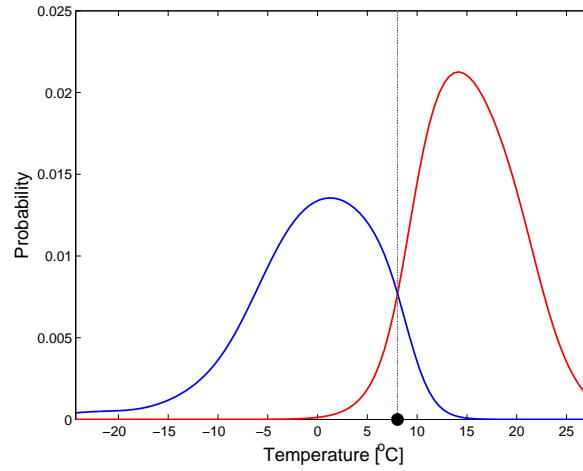


Fig. 4. Break point temperature. The coordinate of the the coordinate of intersection of two curves $p_1(t)$, the probability that real daily gas consumption exceeds the supply (blue line) and $p_2(t)$ probability that real daily gas consumption is smaller than the supply (red line).

It is shown in Fig. 4, that plots of two functions meet in a single pint. This meeting point indicates temperature conditions when it is equally likely that gas consumption exceeds the supply and stays behind it. The coordinate of this point is the temperature break point. We found it equal 8 ± 0.5 °C. As shown in shown in Fig. 4, at temperatures lower than the break point, the probability $p_1(t)$ is higher than $p_2(t)$. This points at the higher likelihood of the supply deficiency. At temperatures higher than the break point, probability $p_2(t)$ is higher than $p_1(t)$. This indicates excessive supply i.e. conditions which are favorable for gas storage.

Using break point temperature there may be estimated the recommended period of gas storage filling. It was defined as the longest continuous time period when the temperature remained over the break point. From our analysis, in 2006 it lasted between 107th and 286th day of the year. This corresponds to nearly 6 months, which is half a year.

5 Conclusions

In this work, there was proposed a method of temperature break point detection with respect to gas storage. It is the ambient temperature associated with the highest probability that the gas supply and demand are balanced. Temperature break point detection is based on the analysis of the dependence structure between daily temperature and daily natural gas consumption. In order to achieve this goal, copulas were applied. In particular, it was found that Frank copula allowed for the most accurate evaluation of the bi-variate distribution of the two

variables. For the studied exemplary data set, the temperature break point was 8 ± 0.5 °C and the conditions favorable for gas storage occurred between the 107 and 286 day of the year. The presented method may have important policy implications.

References

1. Jong C.: Gas storage valuation and optimization. J Nat Gas Sci Eng. 24, 365-378 (2015)
2. Confort M.J.F., Mothe Ch.G.: Estimating the required underground natural storage capacity in Brazil from the gas industry characteristics in countries with gas storage capacities. J Nat Gas Sci Eng. 18, 120-130 (2014)
3. Cherubini U., Luciano E., Vecchiato W.: Copula Methods in Finance, J. Wiley (2006)
4. Masina M., Lamberti A., Archetti R.: Coastal flooding: A copula based approach for estimating the joint probability of water levels and waves. Coast. Eng. 97, 37-52 (2015)
5. Torres R., De Michele C., Laniado H., Lillo R. E.: Directional multivariate extremes in environmental phenomena, Environmetrics 28:e2428, 1-15, (2017)
6. Khan N. M.: GQL Estimation in Bivariate Non-Stationary Poisson Time Series Model Based on Copula Approach, In: International work-conference on Time Series ITISE 2016, pp. 482-493, Granada (2016)
7. Nelsen, R.B.: An Introduction to Copula, Springer (2006)
8. MathWorks - Makers of MATLAB and Simulink, <https://www.mathworks.com/>

An econometric analysis of the merit order effect in electricity spot price: the Germany case

François Benhmad ¹ * Jacques Percebois ²

June 2017

Abstract

In this paper, we carry out an econometric analysis for Germany, in order to investigate impact of wind energy and Photovoltaic feed-in on electricity spot price level, the so-called merit-order effect.

We have used an ARMA-X- GARCH-X modeling in order to assess the joint impact of RES on the electricity spot price level as well as on spot price volatility in Germany.

Our main empirical findings suggest that wind power and Photovoltaic feed-in decreases electricity spot price. However, their impact on electricity spot prices volatility are quite different. Indeed, the solar Photovoltaic power has a lowering on impact electricity price volatility whereas the wind feed-in exacerbates it.

Keywords: RES, Electricity spot prices, merit order effect, volatility.

JEL classification: Q41, Q42, Q48

* Corresponding author

¹ Montpellier University, Site Richter, Avenue Raymond Dugrand , CS79606, 34960 Montpellier Cedex2, France, Phone: +33.4.34.43.25.02, E-mail address : francois.benhmad@umontpellier.fr

² Art-Dev, Montpellier University, Site Richter, Avenue Raymond Dugrand, CS79606, 34960 Montpellier Cedex2, France, Phone: +33.4.34.43.25.04, E-mail address : jacques.percebois@univ-montpl.fr

adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

1. Introduction

Renewable energy is a key component of the EU energy strategy. It started with the adoption of the 1997 White paper and has been driven by the need to decarbonise the energy sector and address growing dependency on fossil fuel imports from politically unstable regions outside the EU.

Various RES supporting schemes are operating in Europe, mainly feed-in tariffs, fixed premiums, green certificate systems. The German Renewable Energy Act, "Erneuerbare-Energien-Gesetz" (EEG), a well known support scheme, has provided a favorable feed-in tariff (FIT) for a variety of renewable energy sources (RES) since the year 2000. It also gives priority to electric power in-feed from RES over power in-feed from conventional power plants, i.e., fossil- and nuclear-fuel thermal and already existing hydro-based power plants. Thus, all renewable sources combined made up 25 per cent of gross electricity production in 2016 and are Germany's second most important source of electricity generation after lignite (BDEW, 2016).

The goal of this paper is to carry out an econometric analysis to investigate the well-known merit-order effect consisting on a downward pressure of RES on the spot electricity price, by using a data sample of daily electricity spot prices in Germany for the 2012-2016 period.

There are two main contributions of this study to the literature. Firstly, in contrast to the previous studies, we take into account the joint impact of wind feed-in and solar photovoltaic on electricity price and volatility with a more recent dataset allowing us to assess the learning curve of new technologies integration at the energy-mix of Germany.

Secondly, an ARMA-X-GARCH-X modeling is used with wind and photovoltaic power generation as exogenous variables included in the mean and the variance equation. The goal is to assess the joint impact of intermittent renewable electricity generation on the electricity spot price level as well as on spot price volatility in Germany.

Our main findings suggest that intermittent wind feed-in and solar photovoltaic power generation not only decrease the spot electricity price in Germany but also have an impact on its price volatility. However,

photovoltaic has a downward impact whereas the wind feed-in has an opposite impact-upward- on electricity spot price volatility.

Several papers have carried out empirical analysis on the impact of RES in electricity markets, finding evidence of the merit-order effect. It is worth noting that several authors have explored this topic. For Germany, Bode and Groscurth (2006) find that renewable power generation lowers the electricity price. Neubarth et al. (2006) show that the daily average value of the market spot price decreases by 1 €/MWh per additional 1,000 MW wind capacity. Sensfuss et al. (2008) show that in 2006, renewables reduced the average market price by 7.83 €/MWh. Weigt (2008) concludes that the price was on average 10 €/MWh lower. Nicolosi and Fürsch (2009) confirm that in the short run, wind power feed-in reduces prices whereas in the long run, wind power affects conventional capacity, which could eventually be substituted. For Denmark, Munksgaard and Morthorst (2008) conclude that if there is little or no wind (<400MW), prices can increase up to around 80 €/MWh (600 DKK/MWh), whilst with strong wind (>1500MW) spot prices can be brought down to around 34 €/MWh (250 DKK/MWh). Jonsson et al. (2010) show that the average spot price is considerably lower at times where wind power production has been predicted to be large. Sáenz de Miera et al. (2008) found that wind power generation in Spain would have led to a drop in the wholesale price amounting to 7.08 €/MWh in 2005, 4.75 €/MWh in 2006, and 12.44 €/MWh during the first half of 2007.

Gelabert et al. (2011) find that an increase of renewable electricity production by 1 GWh reduces the daily average of the Spanish electricity price by 2 €/MWh. Wurzburg et al. (2013) find that additional RES generation by 1 GWh reduces the daily average price by roughly 1 €/MWh in German and Austrian integrated markets. Woo et al. (2011) carry out an empirical analysis for the Texas electricity price market and showed a strong negative effect of wind power generation on Texas balancing electricity prices. Huisman et al. (2013) obtained equivalent results for the Nord Pool market by modeling energy supply and demand. Ketterer (2014) also examined wind power in German electricity markets and found that an additional RES generation by 1GWh led to a reduction of daily spot price by approximately 1€/MWh.

Benhmad and Percebois (2016) also explored German electricity markets for a more recent dataset and found similar results consisting on a reduction of daily spot price by approximatively 1€/MWh for each an additional GWh of wind feed-in.

The paper is organized as follows. Section 2 provides an overview of the merit order effect. In section 3, we carry out an empirical analysis and discuss the main findings. Section 4 provides some concluding remarks.

2. The merit order effect

The electricity market operates according to day-ahead bidding. Indeed, the transmission system operators basically receives the bids from all power producers for the quantity and cost for each hour of the next day and then assigns the dispatch based on the lowest cost producer until demand is met. All producers who dispatch get the marginal price of the last producer that dispatched. This conventional approach consists in ranking the power plants of the system in ascending order of their marginal cost of generation. This approach is called the merit order.

Traditionally, the hydroelectric power plants are the first to be dispatched on the grid, followed respectively by nuclear plants, coal-fired and/or combined-cycle gas turbines (CCGT), and then open cycle gas turbine (OCGT) plants and oil-fired units with the highest fuel costs.

Gas plants are usually the marginal producers. But, due to EU ETS price weaknesses, carbon prices have plunged to record low prices making it more expensive to burn gas than coal. Moreover, The U.S. coal surpluses export due to shale gas revolution has lowered coal prices in Europe. Therefore, the price competitiveness of more polluting coal-fired plants, allow them to be dispatched before the gas turbine and to be the key of electricity price setting.

However, a pricing based on marginal costs could never allow RES to recover their fixed costs. Indeed, the photovoltaic (PV) and wind power plants have a high average cost and their load factor is too low due to intermittency. Therefore, subsidising renewable energy sources by feed-in tariff (FIT) scheme allowing their average costs to be recovered

corresponds to a support mechanism outside the market. By granting an economic return above the market price, these supporting schemes have promoted RES development in several European electricity markets.

As the renewable energy sources (RES) have priority access to grid at zero marginal cost, electricity from RES induces a shifting of the supply curve to the right. Without RES feed-in, during full and peak times, the marginal power plant is logically a combined-cycle gas-fired plant. However, RES make the coal-fired plant becoming the marginal plant. Therefore, RES have a downward impact on average equilibrium price called *merit order effect* (Sioshansi, 2013).

3. Empirical evidence

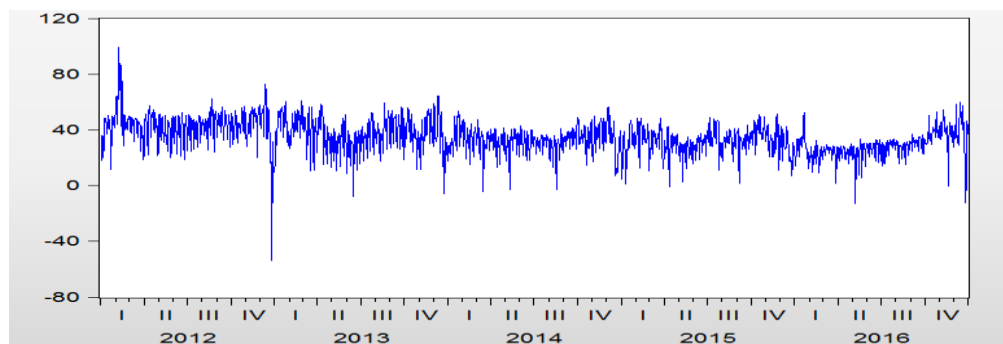
3.1 Data

The sample data covers the period going from the 1st January 2012 to the 31st December 2016, summing up to 1827 observations.

The day-ahead electricity German market consists on hourly contracts with physical delivery on the next day. The daily prices are calculated as the average weighted price over these hourly contracts.

Figure 3 provides a plot of the data which exhibits seasonality, periods of extreme volatility, price spikes and a mean-reverting behavior.

Figure 3. Daily EEX day-ahead spot prices (€/MWh)



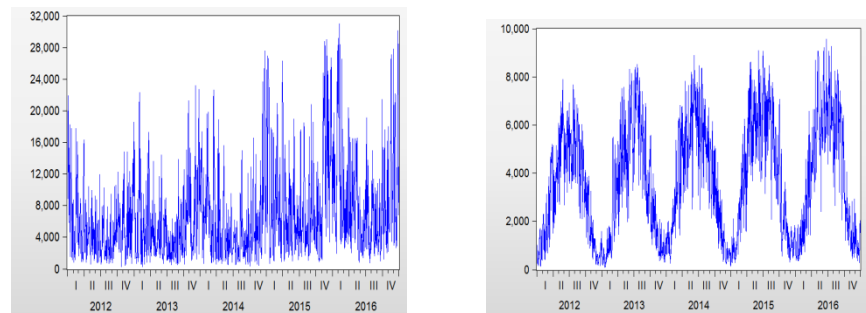
The descriptive statistics of German electricity spot prices summarized in Table 1 show that values of sample mean are close to 34.97 and a standard deviation of 11.74. The sample kurtosis (6.86) indicates that price distribution exhibit fat tails and negative skewness a greater probability of large falls in electricity price than large increases.

Table 1 Descriptive statistics of German electricity spot prices

Mean	34.97
Std.Dev.	11.74
Skewness	-0.33
Kurtosis	6.86
Jarque-Bera	1171.37
Prob.	0.0000

The RES generation data collected by the four German transmission system operators (TSO)¹, for the full period is illustrated in Figure 4.

Figure 4 .Wind power and photovoltaic feed-in (2012-2016)



¹ The data are available in 15-minute format. For this study, 15-minute MW data are averaged for each hour and again averaged to MWh per day. There is four transmission system operators (TSO) in Germany and one TSO in Austria: *Amprion GmbH*, *TenneT TSO GmbH*, *50hertz Transmission GmbH*, *EnBW Transportnetze*, and *APG-Austrian Power Grid AG*.

The descriptive statistics (Table 2) show that the wind power and photovoltaic forecasts fed into the grid has a respectively a daily mean of 6817 and 3651 MWh per day but a high variability.

Table 2. Descriptive statistics of wind feed-in and photovoltaic

	Wind	PV
Observations	1827	1827
Mean	6817.63	3651
Std.Dev.	5652.75	2380
Skewness	1.54	0.28
Kurtosis	5.32	1.91
Jarque-Bera	1135.58	113.14
Prob.	0.0000	0.0000

3.2 Empirical methodology: ARMA-X-GARCH-X model

In order to explore the link between daily electricity spot price and RES wind in-feed (wind and photovoltaic), we should carry out a linear regression using least squares method. As electricity spot prices deviates from the normal distribution due to more frequent large outliers, outliers should first be removed before conducting the regression analysis. Values that exceed 3 times the prices standard deviation are removed and replaced with the value of 3 times the standard deviation. Furthermore, electricity demand has a typical seasonal pattern as it varies throughout the day, during the week, as well as across the year. Therefore, models of electricity prices should incorporate seasonality by using dummy variables. After outliers removal and seasonal adjustment, we carry out an augmented Dickey-Fuller (ADF) test (Dickey and Fuller, 1981) to test for stationarity.

Table 3. ADF unit root test on adjusted electricity spot prices

	t-statistic	Prob.
ADF teststatistic	-8.588311	0.0000
Critical Value (5%)	-2.862924	

The spot electricity prices are then stationary. As electricity is not storable, the price tends to spike and then revert as soon as the divergence of supply and demand is resolved (Escribano et al., 2011).

For the Wind power and photovoltaic generation, after adjusting their seasonal dynamics, the ADF test on (Wind_sa) and (PV_sa) reveals their stationary behavior (the ADF t-statistic is respectively -19.17 and -23.18 whereas the 5% critical value is -2.86).

Even after removing out seasonality and outliers, electricity spot prices still present high order serial correlation in its structure which could be filtered out by an autoregressive moving average (ARMA) filter (Box and Jenkins, 1976). Therefore, the impact of wind-in feed and photovoltaic on electricity prices is explored according to the following ARMA-X model where the wind feed-in and photovoltaic power considered as exogenous variables X:

$$(spot_sa)_t = \alpha_0 + \sum_{i=1}^p \alpha_i (spot_sa)_{t-i} + \sum_{j=1}^q \beta_j \varepsilon_{t-j} + \delta wind_sa_t + \lambda pv_sa_t + v_t$$

The selection of autoregressive lag p could depend on AIC minimization, and q is assumed to be 0. According to the Akaike information criterion, the best choice was lag p = 7 which corresponds to a weekly seasonality.²

The estimation results reported in Table 4 (Column A) reveal a negative impact of wind power on the electricity price in Germany. Indeed, for each additional GWh of wind feed-in, the electricity price decreases by 1€/MWh at the spot market. Therefore, and given the average wind electricity generation during 2012-2016, the merit-order effect roughly corresponds to an average price decrease, in absolute terms, of approximately 7€/MWh.

² The results, not reported here, are available upon request.

Table 4. Wind and Photovoltaic feed-in impact on electricity prices and volatility

Dependant variable : electricity spot prices

Sample : 1.1.2012 31.12.2016

	(A)	(B)
Mean equation		
Constant	-0.46 (0.65)	0.00 (0.99)
Wind	-0.00099(0.00)	-0.0010 (0.00)
PV		-0.00098(0.00)
Variance equation		
Constant	3.59 (0.00)	3.44(0.00)
Alpha	0.31 (0.00)	0.31(0.00)
Beta	0.56 (0.00)	0.56 (0.00)
Wind	0.00016 (0.00)	0.00016 (0.00)
PV		-0.0004 (0.00)
Adj.R squared	0.7467	0.7584
AIC	5.7713	5.6973
BIC	5.8116	5.7426

Note: AIC and BIC stand respectively for Akaike and Bayesian information criterion, p-values are in parentheses.

An ARCH test following Engle (1982), carried out on residuals data shows that residuals are heteroskedastic. Thus, GARCH(1,1) specification (Bollerslev,1986) is used to explore the joint impact of wind in-feed and photovoltaic generation on spot electricity price level and also on price volatility dynamics. Therefore, our empirical analysis is based on ARMA (p,q)-X-GARCH(1,1)-X modeling where wind feed-in and photovoltaic generation are taken as exogenous variables in the mean equation as well as in the variance equation.

The empirical results reported in Table 4 (Column A) show that wind electricity has not only reduced the electricity spot prices (-0.001), in absolute terms approximately 7 €/MWh, but also induced an increase of their volatility (positive sign +0.00016 at the conditional variance equation).

The estimation results (Column B) also reveal a negative impact of solar photovoltaic generation on electricity prices of the same magnitude as wind feed-in. Indeed, for each additional GWh of photovoltaic feed-in, the electricity price decreases approximatively by 1€/MWh at the spot market, in absolute terms, of approximately 3.65€/MWh. However, photovoltaic generation, in contrast to wind generation, also induced a downward pressure of their volatility (negative sign - 0.0004 at the conditional variance equation).

Indeed, the upward effect on electricity prices volatility induced by highly intermittent wind feed-in is largely offset by the photovoltaic downward effect. Thus, the mixture of installed electricity generation capacities consisting on wind and solar photovoltaic allows German electricity market volatility to be less higher than it would be if Germany had only installed wind generation capacities.

4. Conclusion:

The feed-in tariffs support scheme, consisting in buying intermittent electricity at a fixed price off-market has clearly induced a huge market penetration of RES in Germany. The fact that this intermittent electricity has statutory priority on the grid and participates in spot market auctions at a zero marginal cost, leads to a downward trend in the equilibrium price: the so-called merit-order effect.

The purpose of the paper consists in quantifying the merit order effect of wind feed-in and photovoltaics in Germany during the 2012-2016 period. One of the major findings is that the day-ahead electricity spot price fell by 1€/MWh for each additional GWh respectively for the two renewable energy sources. Moreover, the wind electricity generation has an increasing effect on the spot prices volatility which is largely offset by photovoltaics with their downward impact on volatility.

However, although the volatility is controlled by a mixture of installed capacities of RES, the merit order effect remains a big challenge for Germany. This negative effect of RES could significantly be limited by the interconnections with Germany neighbouring countries especially France, allowing it to export its surplus wind power. Therefore, the development of the renewable energy sources should be accompanied by a market coupling.

References

- [1] Benhmad, F., Percebois, J., (2016), “Wind power feed-in impact on electricity prices in Germany 2009-2013”, July, Volume 13-Issue 1, *European Journal of comparative economics*.
- [2] BMU (2016), “Renewable Energies Driving Germany’s Energiewende”, Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, www.bmu.de/english · www.erneuerbare-energien.de (October 2016)
- [3] Bode S., Groscurth H.M. (2006), ‘The Effect of the German Renewable Energy Act (EEG) on the electricity price’, *HWWA Discussion Paper* (348).
- [4] Bollerslev T. (1986), ‘Generalized autoregressive conditional heteroskedasticity’, *Journal of Econometrics* (31), 307–327.
- [5] Box G.E.P., Jenkins G.M. (1970), ‘Time Series Analysis Forecasting and Control’, Holden-Day, San Francisco.
- [6] Dickey, D.A., Fuller, W.A., (1981), ‘Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root’, *Econometrica* (49), 1057-1072.
- [7] Engle, R. (1982), ‘Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation’, *Econometrica* (50), 987-1007.
- [8] ENTSO-E (2012), ‘Load and consumption data: Specificities of member countries’, *Report*, European Network of Transmission System Operators for Electricity, Brussels.
- [9] Escribano A., Ignacio Peña J., Villaplana P., (2011), ‘Modeling electricity prices: International evidence’ *Oxford Bulletin of Economics and Statistics* (73), 622-650.
- [10] Gelabert L., Labandeira X., Linares, P., (2011), ‘An ex-post analysis of the effect of renewable and cogeneration on Spanish electricity prices’ *Energy Economics* (33), S59-S65.
- [11] Jonsson T., Pinson P., Madsen H., (2010), ‘On the market impacts of wind energy forecasts’, *Energy Economics* (32), 313–320.

- [12]Keles D., Genoese M., Most D., Ortlieb S., and Fichtner W., (2013), ‘A combined modeling approach for wind power feed-in and electricity spot prices’ *Energy Policy* (**59**), 213-225.
- [13]Knittel C.R., Roberts M.R., (2005) ‘An empirical examination of restructured electricity prices’, *Energy Economics* (**27**), 791-817.
- [14]Ketterer J.C., (2014), ‘The impact of wind power generation on the electricity price in Germany’, *Energy Economics* (**44**), 270-280
- [15]Mugele C., Rachev S.T., Trück S., (2005), ‘Stable modeling of different European power markets’, *Investment Management and Financial Innovations* (**2**), 65–85.
- [16]Munksgaard J., Morthorst P.E., (2008), ‘Wind power in the Danish liberalised power market Policy measures, price impact and investor incentives’, *Energy Policy* (**36**), 3940–3947.
- [17]Neubarth J., Woll O., et Weber C., Gerecht M., (2006), ‘Influence of Wind Electricity Generation on Spot Prices’, *Energiewirtschaftliche* (**56**) , 42–45.
- [18]Nicolosi M., Fürsch M., (2009), ‘The impact of an increasing share of RES-E on the conventional power market - The example of Germany’, *Zeitschrift für Energiewirtschaft* (**33**), 246–254.
- [19]Sáenz de Miera G., del Rio Gonzalez P., Vizcaino I., (2008), ‘Analysing the impact of renewable electricity support schemes on power prices: The case of wind electricity in Spain’, *Energy Policy* (**36**), 3345–3359.
- [20]Sensfuß F., Ragwitz M., and Genoese M., (2008), ‘The merit-order effect: A detailed analysis of the price effect of renewable electricity generation on spot market prices in Germany’ *Energy Policy*, (**36**):3086-3094.
- [1]Sioshansi F., (2013), ‘Evolution of global Electricity markets’. Ed.Elsevier, June 2013.
- [21]Woo C.-K., Horowitz I., Moore J., and Pacheco A., (2011), ‘The impact of wind generation on the electricity spot-market price level and variance: The Texas experience’ *Energy Policy*, (**39**):3939-3944.
- [22]Wurzburg K., Labandeira X., and Linares P., (2013), ‘Renewable generation and electricity prices: Taking stock and new evidence for Germany and Austria’ *Energy Economics*(**40**), 159-171.

Pattern sequence similarity based techniques for wind speed forecasting

N. Bokde¹, A. Troncoso², G. Asencio-Cortés²,
K. Kulat¹, and F. Martínez-Álvarez²

¹Department of Electronics and Communication, Visvesvaraya National Institute of Technology,
Nagpur, India

neerajdhanraj@gmail.com, kdkulat@ece.vnit.ac.in

²Division of Computer Science, Universidad Pablo de Olavide, ES-41013 Seville, Spain
guaasecor@upo.es, ali@upo.es, fmaralv@upo.es

Abstract. The accurate prediction of wind speed has turned into a very hot topic in recent years. The wind is a clean source of energy that is increasingly being used. In this work, the suitability of applying pattern similarity based algorithms to forecast wind speed is explored. In particular, the PSF algorithm is selected for this problem and data of India are predicted for the first time. Additionally, two kind of predictions are made: single instance prediction and one step ahead predictions. Comparisons to well-known ARIMA and Bayesian methods are reported, with clear superiority of PSF, in terms of standard RMSE and MAE error metrics. These results are very promising and suggest extensive efforts in this line, in order to develop adapted versions and modifications of PSF.

Keywords: Wind speed, time series, pattern sequence, forecasting.

1 Introduction

Wind power is a renewable energy source, that can be relied on for the long-term future. Since wind is non-polluting, its generated energy does not produce gases or radioactivity and is currently being used across the world.

The power grid is affected by the uncertain nature of wind energy. Hence, the ability of predicting wind speed is a task of utmost relevance for both energy managers and electricity traders, in order to minimize the aforementioned uncertainty when this kind of renewable energy is used. Precise wind speed forecasts can be used in many contexts, typically in the evaluation of wind energy potential, in wind power planning or in the design of wind farms. For all the mentioned, to accurately predict wind speed has become a critical task with deep impact and huge benefits for the human kind.

For the first time, the concept of Pattern Sequence-based Forecasting (PSF) method was proposed in [17] and presented in an updated form with a detailed study in [18]. This method found various applications in time series analysis, including electricity price forecasting [18, 11], electric vehicle charging energy

consumption [16], electricity demand forecast [23, 13] or photovoltaic energy demand [9]. These articles highlight the better performance of PSF method over other state-of-the-art methods.

To assess the performance of this methodology in wind speed, data from India have been analyzed for the first time. In particular, years 2012 to 2015 have been analyzed with horizon of prediction of one hour, one day, and one week. Reported results outperform those of other well-established methods, such as ARIMA or Bayesian methods. The successful application of PSF in this context opens new opportunities in this emerging research field.

The rest of the paper is structured as follows. Section 2 reviews and discusses relevant works in the field of wind speed forecasting. Section 3 describes the methodology applied. Results are reported and discussed in Section 4. Finally, the conclusions drawn in this work are summarized in Section 5.

2 Related work

A revision of the different approaches recently published in the literature related to the wind speed forecasting is reported in this section.

Several previous works have dealt with wind speed prediction problems in wind farms using data from measuring towers. The majority of the approaches used modern regression techniques, many of them based on neural networks: multi-layer perceptrons [24, 14, 22], fuzzy-based neural approaches [15], two-hidden layer neural networks [10], fast training neural approaches [20] and abductive networks [1].

In the last years, support vector machines (SVM) have been also proposed to obtain wind speed predictions. In [21], a SVM was applied to predict wind speed from Mexico. The SVM was optimized by a genetic algorithm, showing a good performance when comparing with time series models such as autoregressive integrated moving average (ARIMA).

Regression trees, well-known as Classification and Regression Trees (CART) [4], have been widely applied to predict both wind speed and wind power or to obtain rules between wind and other variables. CART has been used to discover relationships between wind speed and several weather conditions such as atmospheric pressure, temperature, solar radiation, and humidity [19]. The simulation results reported the station pressure and sea-level pressure as the most important variables explaining the wind speed. It can be noticed that the input variables for forecasting the wind speed are meteorological variables while in this work the input variables are the wind measurements of neighbors towers. Moreover, regression trees have been also proposed for predicting the wind energy production. For example, CART has been applied by Clifton et al. [5] to obtain the power output of a wind turbine from wind speed, turbulence intensity and wind shear. The results provided by CART were two or three times better than those of a standard power curve method typically used in the industry.

In Fugon et al. [8] a comparison of different data mining techniques for short-term wind power forecasting in three real wind farms can be found. Namely,

hourly power production and weather forecasts comprising a period of 18 months were considered.

3 Methodology

This section describes the PSF algorithm, an algorithm that has already been successfully applied in other research fields. One of the main features of PSF lies in its ability of discover patterns in the historical data and to make use of them to generate accurate forecasts.

The PSF method can be broadly divided into two steps: (i) labeling of time series using a clustering technique and (ii) sequence based forecast. In the first step, labels are given to time series on the basis of the well-known k-means clustering algorithm, and the time series is partitioned into various cluster centers. The optimal number of clusters is determined by means of a wise combination of three indices: Silhouette [12], Dunn [7] and Davies-Bouldin [6].

By contrast, in the second step, the sequence-based prediction is done following three sub-steps: (a) optimum window size selection, (b) matching pattern sequences and (c) estimation.

In the labeled series, the last sequence of the label with a window of size W is searched for and very next value of each repetition of the window is kept in a new vector. The arithmetic mean of this vector is assigned as a label for next value to be predicted. The selection of W is one of the most important steps on which prediction performance exhibits high dependency. Once the labels for predicted values are obtained, the process of de-normalization is performed on label sequence to achieve actual prediction values.

To continue with further predictions, the predicted label is appended to the sequence of labels in the series prior to the procedure of de-normalization. The block diagram of PSF method and its associated pseudocode are as shown in Figs. 1, 2 and 3, respectively. These figures are adopted from [3]. W is searched for within the label series such that the sequence of labels in the window should repeat at least once within the series. The detailed information of these steps involved in PSF is available in [18, 3]. The efforts are taken in [3] to present an R package for **PSF** [2].

4 Results

This section reports the results achieved by PSF, as well as a comparative study with ARIMA and Bayesian methods. But, first, Section 4.1 describes the dataset used and, second, Section 4.2 the quality parameters used to assess the performance of the considered methods. The results themselves can be found in Section 4.3.

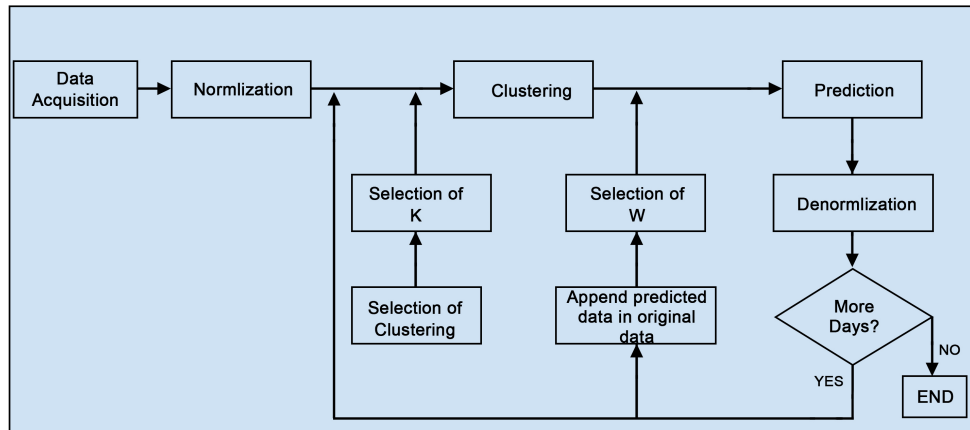


Fig. 1. Block diagram of the PSF algorithm.

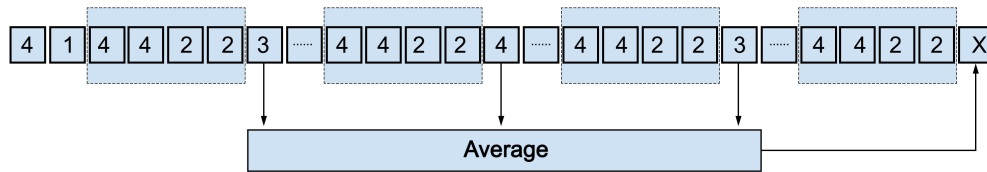


Fig. 2. Prediction with the PSF algorithm.

4.1 Data description

This section describes the data used in this work. The wind speed time series used in the present study has been collected by State load dispatch centre located in the Maharashtra state region of India. The wind speed mean values are collected on an hourly basis, in m/s, for four consecutive years, from January 2012 to December 2015. The total length of the time series is 34,320 values and its statistical parameters including mean and median are shown in Table 1.

Table 1. Summary of wind speed time series.

	Minimum	Maximum	Median	Mean
Wind speed data	101	543	346.5	344.8

Input: Dataset D , number of clusters K , labeled dataset $[L_1, L_2, \dots, L_{d-2}, L_{d-1}]$, length of the window W and Test Set T

Output: Forecasts $\hat{X}(d)$ for all days of T

```

PSF()
   $ES_d \leftarrow \{\}$ 
   $\hat{X}(d) \leftarrow 0$ 
  for each day  $d \in T$ 
     $S_W^{d-1} \leftarrow [L_{d-W}, L_{d-W+1}, \dots, L_{d-2}, L_{d-1}]$ 
    for each  $j$  such as  $X(j) \in D$ 
       $S_W^j \leftarrow [L_{j-W+1}, L_{j-W+2}, \dots, L_{j-1}, L_j]$ 
      if ( $S_W^j = S_W^{d-1}$ )
         $ES_d \leftarrow ES_d \cup j$ 
    for each  $j \in ES_d$ 
       $\hat{X}(d) \leftarrow \hat{X}(d) + X(j + 1)$ 
     $\hat{X}(d) \leftarrow \hat{X}(d) / \text{size}(ES_d)$ 
     $D \leftarrow D \supset \hat{X}(d)$ 
     $[L_1, L_2, \dots, L_{d-1}, L_d] \leftarrow \text{clustering}(D, K)$ 
     $d \leftarrow d + 1$ 
  return  $\hat{X}(d)$  for all days of  $T$ 

```

Fig. 3. Pseudocode for the PSF algorithm.

4.2 Quality parameters

The prediction with PSF method is compared to ARIMA and Bayesian methods. Prediction of these methods are compared for various durations (varies from 1 hour to few days) and the comparison is assessed by means of Root Mean Square Value (RMSE) and Mean Absolute Error (MAE), as performance measures. Their formulas are shown in Eq. 1 and 2.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |X_i - \hat{X}_i|^2} \quad (1)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |X_i - \hat{X}_i| \quad (2)$$

where X_i is the original data, \hat{X}_i is the corresponding predicted data, and N is the number of samples in X .

4.3 Wind speed forecasting

In this section, the prediction of wind speed time series is performed with PSF and compared to few well-established state-of-the-art time series predic-

tion methods. The performance of PSF is examined with two types of prediction techniques: single instance prediction and one step ahead predictions.

In single instance prediction, predictions for various time intervals are performed in a single step, whereas in one step ahead forecast technique, the prediction is performed one step by another. In this technique, one step is predicted at a time and the predicted value is appended on the time series and the forecasting continues until desired time instances are predicted.

The observations describing the comparison of various prediction methods are shown in Table 2. Among all three methods under study, the RMSE values for PSF method are found to be lesser than that of both ARIMA and Bayesian methods for all time durations.

Note that for the application of these methods, PSF parametrization consisted in setting $K=2$ (number of clusters) and $W=10$ (size of the window). These values are automatically generated if using the R package [3]. Alternatively, the *forecast* R package was used to apply ARIMA, which suggested that $ARIMA(4,1,2)$ was the most suitable method for this particular time series. As for Bayesian methods, lower and upper quantiles for the forecast interval estimate were set to 0.025 and 0.975, respectively and the seasonal state model was set to 24 hours.

Table 2. Error comparison for prediction at single instance.

Methods	PSF		Bayesian		ARIMA	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
1 Hour	5.03	4.63	7.31	6.02	6.98	5.90
2 Hours	8.48	7.96	12.03	10.98	11.70	10.05
6 Hours	38.75	31.11	45.11	42.66	47.54	43.83
12 Hours	48.56	42.96	53.69	48.34	59.41	54.35
1 Day	51.55	45.03	67.07	62.12	72.01	66.41
2 Days	63.43	57.35	80.35	71.72	87.90	78.77

Similarly, in the case of one step ahead forecast technique, the performance of the PSF method is compared to ARIMA and Bayesian methods. The time series are predicted for next 20 and 40 hours. Table 3 shows the RMSE errors comparison. It analysis reveals the superiority of the PSF method for wind speed time series prediction. The prediction with PSF method is shown in Fig 4, which shows its accuracy while performing one-step ahead forecasting. Apart from this, the summary of the 40 hours predicted values are shown in Table 4. It shows the mean and median of the time series used.

5 Conclusions

Wind speed for India data have been successfully forecasted in this work. Four years, from 2012 to 2015, have been predicted with the PSF algorithm in two

Table 3. Error comparison for prediction at one step-ahead forecast.

Methods	PSF		Bayesian		ARIMA	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
20 Hours	29.88	27.90	37.12	35.63	46.09	49.22
40 Hours	40.43	35.61	47.32	41.76	59.83	54.03

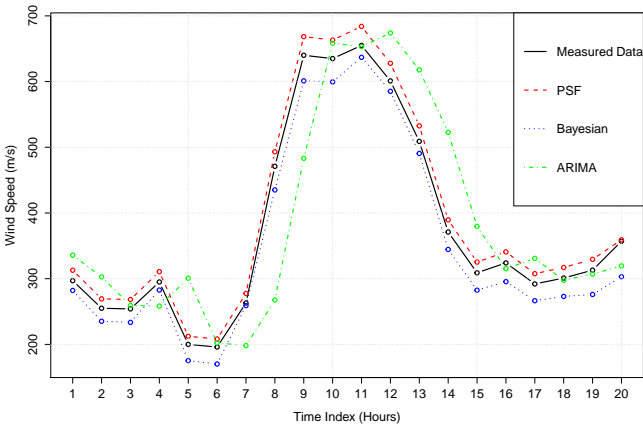


Fig. 4. One-step ahead forecast comparison.

Table 4. Summary of predicted values for duration of 40 hours.

	Minimum	Maximum	Median	Mean
Validating time series	196	655	311	376.9
PSF	198.2	674	317.4	384.2
Bayesian	170.1	636.9	282.6	351.4
ARIMA	206.8	681	324.5	382.6

different ways –one single sample and one step ahead– reaching quite promising results. Additionally, other well-known algorithms, ARIMA and Bayesian methods, have been used for comparative purposes, confirming the robust values obtained by PSF. Given the good performance exhibited by PSF in wind speed forecasting, future works are directed towards the development of adapted versions of PSF to this particular time series.

Acknowledgments.

The authors would like to thank the Spanish Ministry of Economy and Competitiveness and Junta de Andalucía for the support under projects TIN2014-55894-C2-R and P12-TIC-1728, respectively.

References

1. Abdel-Aal, R.E., Elhadidy, M.A., Shaahid, S.: Modeling and forecasting the mean hourly wind speed time series using gmdh-based abductive networks. *Renewable Energy* 34(7), 1686–1699 (2009)
2. Bokde, N., Asencio-Cortés, G., Martínez-Álvarez, F.: PSF: Forecasting of Univariate Time Series Using the Pattern Sequence-Based Forecasting (PSF) Algorithm (2017), <http://www.neerajbokde.com/cran/psf>, r package version 0.4
3. Bokde, N., Asencio-Cortés, G., Martínez-Álvarez, F., Kulat, K.: PSF: Introduction to R Package for Pattern Sequence Based Forecasting Algorithm. *The R Journal* 9(1), 324–333 (2017)
4. Breiman, L., Friedman, J., Ohlsen, R., Stone, C.: Classification and regression trees (1984)
5. Clifton, A., Kilcher, L., Lundquist, J., Fleming, P.: Using machine learning to predict wind turbine power output. *Environmental Research Letters* 8(2), 024009 (2013)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2), 224–227 (1979)
7. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics* 4(1), 95–104 (1974)
8. Fugon, L., Juban, J., Kariniotakis, G.: Data mining for wind power forecasting. In: *Proceedings of the European Wind Energy Conference & Exhibition*. pp. 1–6 (2008)
9. Fujimoto, Y., Hayashi, Y.: Pattern sequence-based energy demand forecast using photovoltaic energy records. In: *Proceedings of the IEEE International Conference on Renewable Energy Research and Applications*. pp. 1–6 (2012)

10. Grassi, G., Vecchio, P.: Wind energy prediction using a two-hidden layer neural network. *Communications in Nonlinear Science and Numerical Simulation* 15(9), 2262–2266 (2010)
11. Jin, C.H., Pok, G., Park, H.W., Ryu, K.H.: Improved pattern sequence-based forecasting method for electricity load. *IEEJ Transactions on Electrical and Electronic Engineering* 9(6), 670–674 (2014)
12. Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons (2009)
13. Koprinska, I., Rana, M., Troncoso, A., Martínez-Álvarez, F.: Combining pattern sequence similarity with neural networks for forecasting electricity demand time series. In: *Proceedings of the IEEE International Joint Conference on Neural Networks*. pp. 940–947 (2013)
14. Kramer, O., Gieseke, F., Satzger, B.: Wind energy prediction and monitoring with neural computation. *Neurocomputing* 109, 84–93 (2013)
15. Ma, X., Jin, Y., Dong, Q.: A generalized dynamic fuzzy neural network based on singular spectrum analysis optimized by brain storm optimization for short-term wind speed forecasting. *Applied Soft Computing* 54, 296–312 (2017)
16. Majidpour, M., Qiu, C., Chu, P., Gadh, R., Pota, H.R.: Modified pattern sequence-based forecasting for electric vehicle charging stations. In: *Proceedings of the IEEE International Conference on Smart Grid Communications*. pp. 710–715 (2014)
17. Martínez-Álvarez, F., Troncoso, A., Riquelme, J.C., Aguilar-Ruiz, J.S.: LBF: A labeled-based forecasting algorithm and its application to electricity price time series. In: *Proceedings of the IEEE International Conference on Data Mining*. pp. 453–461 (2008)
18. Martínez Álvarez, F., Troncoso, A., Riquelme, J.C., Aguilar-Ruiz, J.S.: Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering* 23(8), 1230–1243 (2011)
19. Mori, H., Awata, A.: Feature extraction of meteorological data using regression tree for wind power generation. In: *Proceedings of the IEEE Sustainable Energy Technologies*. pp. 1104–1107 (2008)
20. Saavedra-Moreno, B., Salcedo-Sanz, S., Carro-Calvo, L., Gascón-Moreno, J., Jiménez-Fernández, S., Prieto, L.: Very fast training neural-computation techniques for real measure-correlate-predict wind operations in wind farms. *Journal of Wind Engineering and Industrial Aerodynamics* 116, 49–60 (2013)
21. Santamaría-Bonfil, G., Reyes-Ballesteros, A., Gershenson, C.: Wind speed forecasting for wind farms: A method based on support vector regression. *Renewable Energy* 85, 790–809 (2016)
22. Sheela, K.G., Deepa, S.: Neural network based hybrid computing model for wind speed prediction. *Neurocomputing* 122, 425–429 (2013)
23. Shen, W., Babushkin, V., Aung, Z., Woon, W.L.: An ensemble model for day-ahead electricity demand time series forecasting. In: *Proceedings of the International Conference on Future Energy Systems*. pp. 51–62 (2013)
24. Velázquez, S., Carta, J.A., Matías, J.: Comparison between ANNs and linear MCP algorithms in the long-term estimation of the cost per kWh produced by a wind turbine at a candidate site: a case study in the Canary Islands. *Applied Energy* 88(11), 3869–3881 (2011)

Improving the performance of machine learning models by integrating partly physical control response models in short-term forecasting of aggregated power system loads

Pekka Koponen¹, Harri Niska², Reino Huusko³

¹ Energy Systems, VTT Technical Research Centre of Finland, Espoo Finland,
pekka.koponen@vtt.fi

² Environmental Informatics University of Eastern Finland, Kuopio, Finland,
harri.niska@uef.fi

³ Loiste Electricity Network Kajaani, Finland, reino.huusko@loiste.fi

Abstract. Combining the strengths of different modelling approaches and various information sources is studied in short-term forecasting of aggregated electrical loads that are controllable and include e.g. thermal storage capacity. Measurement data driven models tend to fail in forecasting power during rare situations such as dynamic control actions and extreme weather conditions. The thermal dynamics of the loads, large outdoor temperature variations, and changes in the technologies contribute to this challenge. Here we study a model integration approach using field trial data covering about 7000 houses and 27 months. Control responses and load saturation are forecast using a physically based structure. The residual is forecast with a machine learning model designed and tuned to learn also system dynamics. The load forecast is the sum of these component forecasts. The forecasting accuracy of this hybrid method is compared with using the machine learning alone. The results show improvement in the accuracy.

Keywords: forecasting, machine learning, physically based models, smart grid.

1 Introduction

Accurate forecasts of the power flows in the distribution system are a critical enabler for high penetrations of distributed power generation and demand response. Ignoring the explicit presence of active demand in the model of the load leads to unsatisfactory forecasts according to [1] and [2].

This contribution belongs to a project Response funded by the Academy of Finland, which studies the following research hypotheses. 1) Hybrid models can combine the benefits of different load modelling approaches, thus providing models that (a) forecast relatively accurately in different situations including also those that have not been experienced before, (b) adapt to expected and unexpected changes in the load behavior, and (c) are easy and fast to maintain. 2) Models that combine all relevant available information forecast dynamically controlled aggregated load more accurately than black box models (purely data driven models) or purely physically based models.

There are several ways to improve forecasting accuracy by combining forecasting methods. An approach is to run several forecasting algorithms in parallel and use a weighted average of the forecasts while adjusting the weights according to the situation as learned in the identification [3]. A hybrid ARIMA-ANN model for time series prediction is proposed and studied by [4]; there a multilayer perceptron forecasts the residual of the ARIMA system. We found forecasting the control responses using ARIMA unreliable and inaccurate. The obvious reasons include nonlinearities, nonstationary behavior and limited amount of test responses. We use a model with a physically based structure to forecast the control responses and the saturation of the load, and the machine learning models forecast the residual. Then the load forecast is the sum of these two component forecasts. We successfully applied this approach for electricity spot price based direct control of the aggregate loads of full storage heating houses [2].

In the present contribution, we explain the methods of [2] and give a new summary of the results. Then we apply and modify the approach of [2] to very seldom activated emergency load control of partial storage heating houses located in a climate with large temperature variations. A further difference is that the control responses are modelled from aggregated 3 minute interval measurements from the primary substations in addition to the hourly interval measurements from the smart billing meters. That enables forecasting the emergency control responses with 3 minute time resolution, which is necessary. We also apply and compare two machine learning methods: support vector machine (SVM) and multilayer perceptron (MLP). According to the literature, such as [5], SVM has many methodological benefits and produces smaller forecasting errors.

2 The forecasting problem

The problem studied is to forecast aggregated powers of customer groups that include active demand (AD). The focus is on short-term forecasting: each day at 9 a.m. the power during the next day is forecast with one hour or 3-minute time resolution. Hourly interval consumption measurements from the previous day are available from each customer. The behavior of individual customers is very stochastic but their aggregated behavior is rather well predictable. The outdoor temperature in the region has large variations and the AD responses and loads have highly nonlinear behavior due to saturation of cooling and especially heating. Accurate forecasts during high load situations, such as very cold temperatures, are very important, because then the balancing errors are exceptionally costly and the operational margins in the distribution grid are small.

Two AD forecasting cases are studied using load control field test data. These are

- 1) forecasting about 700 full storage electrically heated houses subject to electricity spot price based direct load control in Helsinki, and
- 2) forecasting partial storage electrically heated houses subject to both emergency load control and Time-of-Use (ToU) load control. (The reported results represent 5188 customers. Slightly over 7500 customers were controlled in the verification tests, but we removed from this study all those sites that had gaps in the data or clearly different load behavior.)

In the first case the identification period was 12 months and the verification period was 7 months. In the latter case, the identification data covered 13 months and the verification data 14 months. For forecasting the emergency load control responses, time resolution of the forecasts must be better than 3 minutes. With the physically based response model structures, this is easy to achieve.

Hourly interval consumption history of each customer is available thanks to ubiquitous smart metering. In addition to them, we used outdoor temperature measurements and forecasts, and power measurements from the primary substations to identify and verify the emergency response models.

3 Background research for the emergency load control case

Paper [6] developed and studied physical model based short term daily energy forecasting using the identification data of the emergency load control case. Fig. 1 shows how the daily mean power per house and outdoor temperature varied in the identification period. The developed forecast is shown denoted as simulation. The figure shows aggregated sliding 24 mean powers and sliding 24 h mean outdoor temperature. Here they demonstrate how large the temperature variation is, how much the loads depend on it and how short the extreme temperatures are. It was found out that most of the temperature dependence comes from heating loads, but there is significant cooling load in summer time, and it has somewhat different dynamics than the heating.

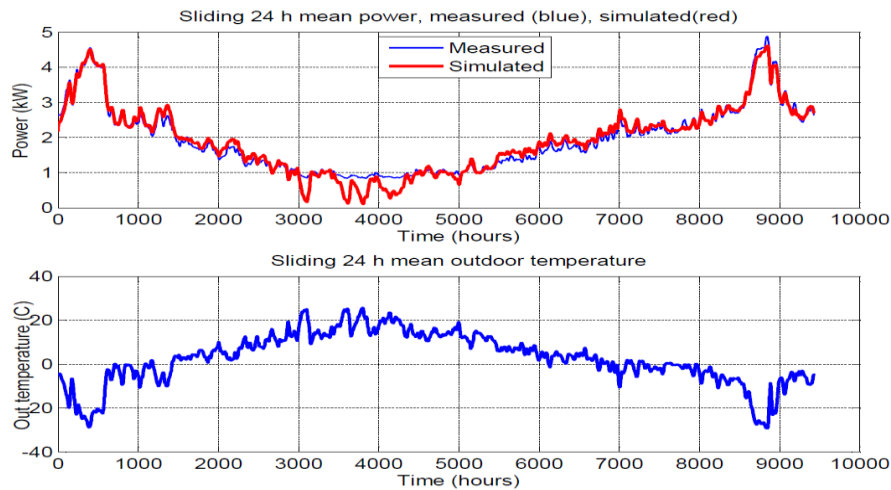


Fig. 1. Temperature dependence of the load in 2011 [6].

4 The hybrid forecasting approach

Machine learning methods alone tend to have challenges in forecasting the dynamics of power during temperature dependent active demand responses and during the load

saturation. Here we study a potential solution. We forecast the control responses and load saturation using model structures based on the thermal dynamics of the houses. We identified the parameter values from the identification field tests also taking into account the feasible parameter ranges estimated from the building codes. Then the machine learning methods were taught to forecast the residual of the physically based model. The residual is also a dynamic process so the machine learning models applied need to include capabilities to model the dynamics.

Fig. 2 shows the resulting main structure of the forecasting model.

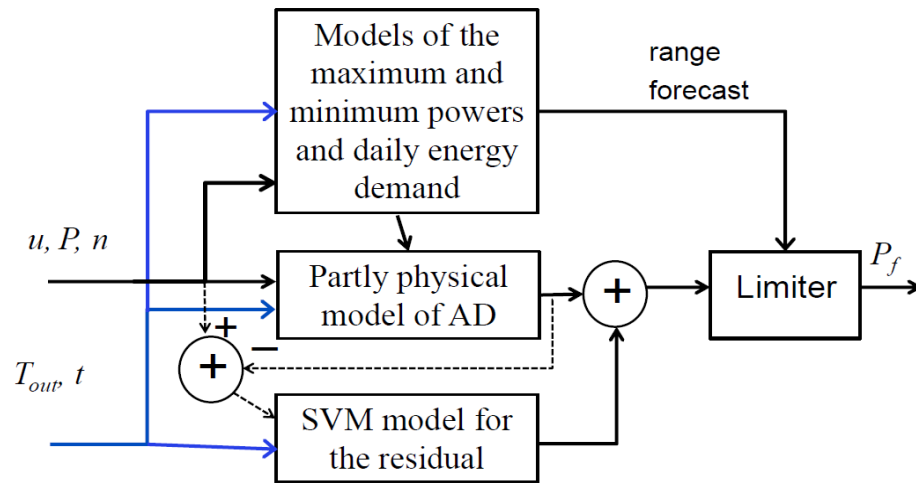


Fig. 2. The machine learning model forecasts the residual of the partly physically based response model.

P is the measured power, P_f the power forecast, u is the control signal and n is the number of houses. T_{out} is a combination of the measured and forecast temperatures as available at the time of forecasting the power. Time t is also an input signal. Each controlled group has its own control signal u and model. The residual model may be common to all groups or each group may have its own residual model. Which one is better depends on the performance and complexity in the particular case.

5 The machine learning methods

In this study, two standard machine learning methods: multi-layer perceptron (MLP) and support vector machine (SVM) are evaluated. Both the methods are largely adopted in the domain of electric load forecasting, e.g. [7].

5.1 Support vector machine (SVM)

Basically SVM is a machine learning technique for data classification and non-linear regression. For more technical details the reader is referred e.g. to [8].

In this study, epsilon(ϵ)-SVM (or SVR) with the radial basis kernel function based on the LIBSVM package was used to execute the model runs. We have adopted the direct prediction scheme for both the machine learning models by using delayed power and temperature values as regressors. Alternatively, this could be considered as a recursive forecast problem by performing one step ahead prediction. The hyper parameters of SVM were defined as follows. The gamma of kernel function was defined using the LIBSVM default, i.e. $1/\text{number of features}$ whereas C (value 20) was defined based on experimental testing. The value of epsilon ϵ (0.1) was defined based on the defaults.

The input variables of SVM model were selected based on the previous study [9], which showed that timing variables: day of year, day of week, hour of day, and day length and few delayed outdoor temperature values (-0 hour, -9 hour, and some longer delays, here we used -19 hour) are required to produce sufficient prediction accuracy within the direct prediction scheme. Additionally, we have used delayed power value (here -48 hour) as the SVM model input.

The input data were normalized between -1 and 1. The variance scaling was also tested to prevent influence of potential outliers, but it was not observed to enhance the accuracy.

5.2 Multilayer Perceptron (MLP)

Following, the basic outlines given in the SVM model definition, the standard MLP model was trained to forecast hourly mean powers using timing variables, outdoor temperature and power measurements. The MLP network with one hidden layer (25 nodes) was trained using Levenberg-Marquadt algorithm. In total 3000 training epochs were utilized. A subset of the identification (training) data (5%) were used to control potential over-fitting and to ensure external prediction power. Discontinuous input variables (such as hour of day, day of week) were divided into continuous form by using sine and cosine transformations. This transformation was adopted in case of SVM model, as well.

5.3 Modelling the system dynamics with the machine learning methods

A set of delays is introduced to the forecasting model and during the identification those delays are selected that best improve the fit.

6 Partly physically based control response models

The responses of active demand (AD) to control signals are modelled using models of the thermal dynamics for the buildings and their heat storages. In the houses, the temperature controls are often on-off type. The heating is either on full power or zero power. Such a model is very inaccurate in forecasting the aggregated behavior if a large number of models with stochastic disturbances is not run in parallel. Thus we use a

continuous controller in the house model. We fit it to the observed aggregated responses. It turned out that it forecasts accurately the aggregated responses also when the heating in the individual houses is controlled on-off.

For many model parameters a feasible range was estimated from the building codes that set the minimum requirements for the dimensioning and operation of heating, ventilation and insulation. Then the model parameters were estimated by fitting them to the measured test responses in the identification data. In the partial storage heating case the parameters were identified using nonlinear constrained optimization (such as sequential quadratic programming, SQP, or nonlinear conjugate gradient methods). Several initial guesses were used, because multiple local optima were sometimes detected.

The tests of the emergency load control in the identification data set did not include load control actions in cold enough temperatures. Thus it is not possible to model the saturation of heating powers from them. The number of tests was also very small and the information on the rough geographical location of each controlled customers was unknown. Similar emergency load control field tests using power measurements from 13 substations in different cold temperatures had been implemented in 1996-1997 and summarized in [10] in chapter 6 and Appendix B. Some of the response models identified from them were applied to the new identification test data. Good forecasting performance was observed. Thus for the emergency load control we use the dynamics and saturation from the old models as such. Only static gain of the model is identified on-line from the past measurements. A figure of the structure of the emergency load control response model is given in the Appendix B of [10]. It comprises four internal temperatures, the corresponding heat storage capacities, the connecting heat conductivities and ventilation heat loss. The internal state of the temperature controller is in the model, too. The input variables are the following three temperatures: outside air, ground and the set point of the inside temperature.

The model for the responses of full storage heating (for space heating and hot domestic water) includes only the heat storage and its heating element. The thermal dynamics of the building are taken into account only via the forecasting of the heat demand. This response model is given in [2].

The modelling of partial storage heating response still needs some research. For it the model of full storage heating tends to be too simple alone. The very simple response model may nevertheless be adequate in combination with a suitable machine learning model that possibly compensates the shortcomings.

7 Results

Both in the spot price based control case and in the emergency control case the applied machine learning models did not alone forecast the load control responses accurately enough while the hybrid model accurately forecast also the responses. A further initial observation is that also when loads were not dynamically controlled the hybrid models consistently had a slightly better forecasting accuracy than the machine learning models alone. The forecasting performance in exceptional weather situations, and near summer time winter time clock changes, improved.

7.1 Results in the full storage heating case

We studied the dynamically controlled full storage heating case in [2]. There the identification data include 365 days and the verification period was 208 days long. The test included about 700 houses divided in two separately controlled groups.

Table 1 summarizes the results added with a new row. The performance criterion is root mean square error (RMSE) of the forecast normalized to the annual mean power. The first two rows represent models that have a physically based structure. Only the latter one includes a model of the control responses. The method on the last row models the responses using SVM based machine learning. All the other rows are different versions of the approach shown in the Fig. 2, where partly physical models forecast the control responses and saturation, and SVM forecasts the residual. The residual model is common to both groups, because it gave a slightly better performance than forecasting the group residuals separately.

Table 1. Forecasting the residual using SVM improves the forecasting performance

RMSE (normalized)	Identification	Verification
partly physical without response model	0.99105	1.14260
partly physical with response model	0.33606	0.52645
response model and SVM	0.22893	0.36391
response model, SVM and minimum	0.22841	0.34487
response model, SVM and range limit	0.22827	0.34400
SVM	0.17224	0.75300

The SVM alone forecast very well the identification data but not the verification data. This suggests that the 365 dynamically controlled days in the identification data set were not enough for the SVM to generalize the control responses correctly. The hybrid method clearly outperformed its component methods. Using a physically based model for range limitation gives a small further improvement in forecasting performance. Fig. 3 shows a sample of the best forecast of Table 1 compared to the measured power in verification.

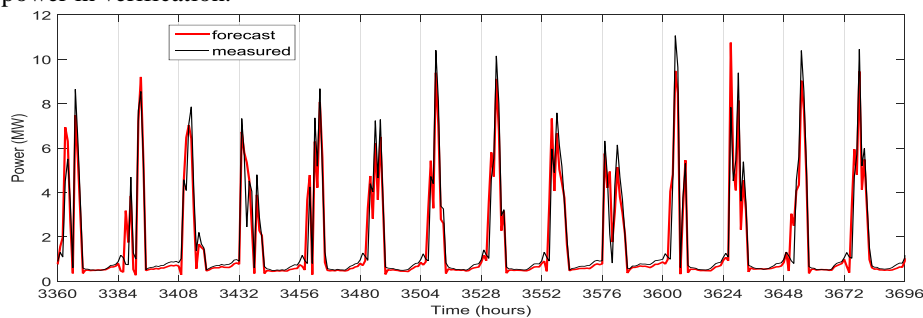


Fig. 3. Forecast and measured aggregated power in verification of the full storage heating case, an example.

The control signals were dynamically scheduled according to the electricity market energy prices and forecast heating needs. When the control signal allows the heating and the heat storage is not full, a thermostat turns heating on and high power peaks occur. The physically based response model models the aggregated behavior of such heat storage system.

7.2 Results in the emergency load control case

The control response model.

The identification period was 13 months long and included some emergency load control tests in early 2013. Then about 8600 electricity customers were subject to the tests and we selected 7062 of them for the response modelling. The test comprised two main groups controlled at different times thus enabling the response identification by reference group comparison. The main groups were split further to subgroups. An identified response of hourly interval powers in outdoor temperature -5°C is in Fig. 4.

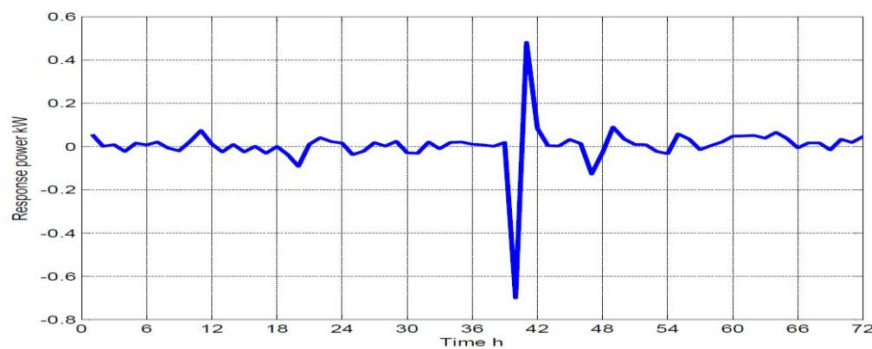


Fig. 4. Emergency load control response identified by reference group comparison, when outdoor temperature was about -5°C ; the control command was applied on hour 40.

The tests in the identification data set were not alone adequate for modelling the emergency load control responses in the temperature range of interest. It was only possible to model the control response in one temperature. Similar emergency load control response identification tests had been implemented in three adjacent power distribution areas in winter 1996-1997 using power measurements from 13 substations [6]. Then the hourly interval measurements of each customer were not available, but the temperature range covered by the tests was $-7\ldots-29^{\circ}\text{C}$, which was wide enough to see also the load saturation in some tests. The dimensioning temperature in the area was -32°C and based on the test results the actual dimensioning of heating and insulation was somewhat better. In the 1996-1997 tests six separately controlled groups of houses were applied based on the thermal dynamics and usage of the buildings. A simple thermal dynamics model was developed, its feasible parameter ranges defined based on building codes and the parameters identified using a nonlinear constrained optimization method, Sequential Quadratic Programming. Now we chose one of those six models

for comparison with the 2013 control tests and the results are shown in Tables 2 and 3 below. In the Table 2 the measured responses are uncertain, because the power measurement includes all the loads in the whole distribution area. Especially the reduction of the load in the test 2 has large uncertainty due to apparent simultaneous changes in the other loads. In Table 3 the measured responses are much more accurate.

Table 2. Comparison of the average house responses of power measurements in 2013 with the model identified in 1997 (3 minute time resolution).

Test nr.	Source	Temperature during test °C	Previous mean 24 h temperature °C	Number of houses controlled	Load step down kW	Load step up kW
1	measurement	-5.5	-6.9	4757	1.1	2.5
1	old model	-5.5	-6.9	4757	0.95	2.2
2	measurement	-4.5	-9.0	2305	1.2-1.6	2.3
2	old model	-4.5	-9.0	2305	0.98	2.2

Table 3. Comparison of the average house responses of hourly interval powers in 2013 with the model identified in 1997.

Test nr.	Source	Houses controlled	reduction in load kWh/h	next hour pay-back kWh/h
1+2	measurement	7062	0.7	0.5
1+2	old model	7062	0.93	0.41

The old model forecast reasonably well the responses in the emergency load control in the identification data. The time constants are slightly too short and can be adjusted accordingly. For clarity in the following, we use the dynamics of the old model as such. Only the scaling of the response model is identified on-line from the latest power measurements available during the forecasting.

Integration with machine learning models.

Tables 4 and 5 compare the forecasting performance of the hybrid approach with the machine learning methods. Hourly interval powers are forecast. In the verification, the controlled houses were in six groups and the four biggest groups are shown here.

Table 4. Comparison of machine learning with the hybrid methods over the verification period.

Method	RMSE (normalized)			
	Group 1	Group 2	Group 3	Group 4
SVM alone	0.1457	0.1756	0.6850	0.6866
MLP alone	0.1283	0.1651	0.8904	0.9622
SVM with response model	0.1161	0.1290	0.3801	0.3767
MLP with response model	0.1108	0.1361	0.4758	0.4622

Table 5. Comparison of machine learning with the hybrid approach when emergency load control was applied in verification; RMSE is evaluated over two 48 h periods, one for each of the two control actions.

Method	RMSE (normalized)			
	Group 1	Group 2	Group 3	Group 4
SVM alone	0.2180	0.2784	1.1122	1.1380
MLP alone	0.2037	0.2880	1.3519	1.4536
SVM with response model	0.1162	0.1350	0.5681	0.5820
MLP with response model	0.1126	0.1750	0.6186	0.6493

SVM and MLP produced roughly equal accuracy and they could not predict emergency control load situations. By combining the methods with the physically based response model, also the dynamic control situations were predicted with good accuracy. We prefer the use of SVM models, because MLP has many well-known challenges, such as a risk of over-fitting.

In the verification, the groups were different from the identification. In the identification, the average customer size was the same in all the four main groups. In the verification, the average customer size in the groups 3 and 4 was clearly smaller. The change in the group size from identification to verification resulted in large errors in the forecast. We compared two solutions: 1) each identification test group was split to two subgroups based on the average annual power thus enabling the machine learning to learn the dependence on the average group, and 2) the hybrid forecast was scaled using feedback from the measurement history available when making the short-term forecasts. The same feedback from the measured average power of customer scaled the partly physically based control response model in all alternatives.

Often the on-line feedback scaling of the response model turned out to be the most accurate although the feedback scaling took about two first weeks of data history to converge to a suitable feedback gain. Table 6 shows the results of the comparison.

Table 6. Modelling the dependence on average site power of the group.

Normalized RMSE	Without scaling		Identification from data split based on customer size		Feedback scaling to group mean size	
	Machine learning with response model					
	MLP	SVM	MLP	SVM	MLP	SVM
Group1	0.1108	0.1161	0.1712	0.1587	0.1047	0.1122
Group2	0.1795	0.1290	0.1803	0.1717	0.1431	0.1143
Group3	0.4758	0.3801	0.1365	0.1333	0.1640	0.1718
Group4	0.4622	0.3767	0.1657	0.1330	0.1820	0.1639

Figures 5 and 6 show the machine learning forecasts, hybrid forecasts and the measured responses. The hybrid forecast is the sum of the physically based response forecast and the machine learning forecast of the residual. Alone the machine learning models are not able to forecast the emergency load control responses, see also Table 5.

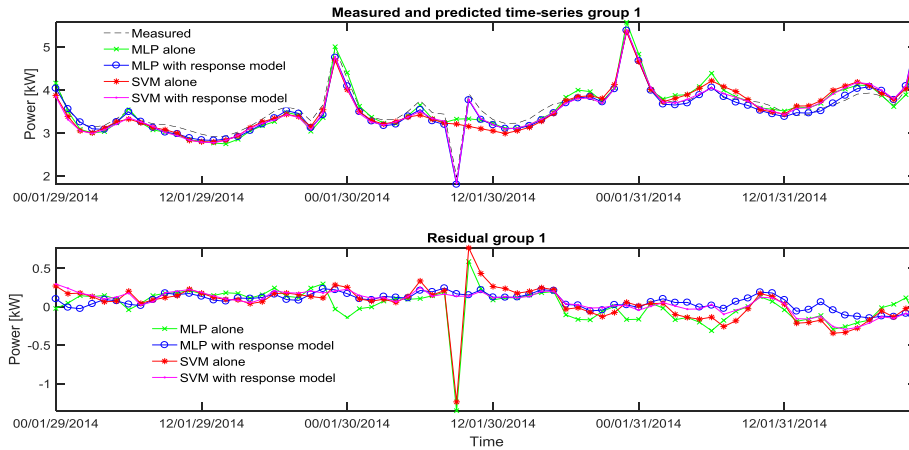


Fig. 5. The responses of the forecasting methods during emergency load control 30 January 2014 in the verification.

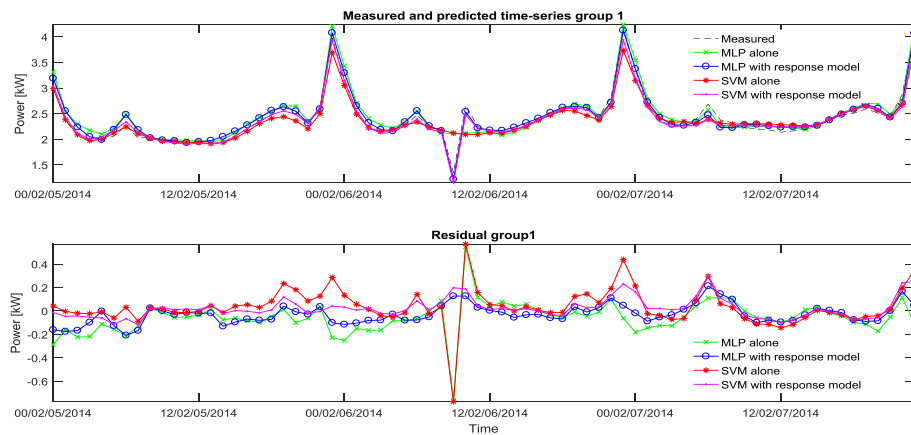


Fig. 6. The responses of the forecasting methods during emergency load control 6 February 2014 in the verification.

8 Discussion

Planned future studies include:

- forecasting the total power of the distribution area with 3 minute time resolution when dynamic load control is applied,
- other hybrid methods in AD forecasting,
- on-line implementation and field testing of the response forecasting,
- tests in cold temperatures,
- analysis and development of criteria for the performance of load forecasting, and
- estimating confidence intervals for the forecasts.

Commonly applied performance criteria reflect poorly or very poorly the costs of forecasting errors. Selection and development of performance criteria should be considered. Splitting the analysis to four groups enables getting some information on the confidence intervals of the forecasts, but further studies are needed.

Common claims are that 1) real time measurements of individual AD customers are necessary and 2) determination of a fair base case for reference and thus the actual response is ambiguous. Individual customer real time measurements improve the performance of forecasting aggregated loads so little that they may be difficult to justify. Our models always forecast the base case or reference case in addition to the responses to the planned control actions. Further studies could clarify these issues.

9 Conclusion

The results show that the hybrid model developed forecasts more accurately than the machine learning models as such. In the hybrid model, the control responses and load saturation are forecast using a physically based structure and the residual is forecast with a machine learning model designed and tuned to learn also system dynamics. The hybrid load forecast is the sum of these component forecasts.

References

1. Garulli, A., Paoletti, S., Vicino, A.: Models and techniques for electric load forecasting in the presence of demand response, *IEEE Transactions on Control Systems Technology*, Vol. 23, No. 3, May 2015, pp. 1087-1097.
2. Koponen, P., Niska, H.: Hybrid model for short-term forecasting of loads and load control responses, *IEEE PES ISGT Europe 2016*, 9-12 October 2016, 6 p.
3. Csáji, B. Cs., Kovács, A., Váncza, J.: Online learning for aggregating forecasts in renewable energy systems, *ERCIM NEWS 107*, October 2016, pp. 40-41.
4. Valenzuela, O., Rojas, I., Rojas, F., Pomares, H., Herrera, L.J., Guillen, A., Marquez, L., Pasadas, M.: Hybridization of intelligent techniques and ARIMA models for time series prediction. *Fuzzy Sets and Systems*, Vol. 159, Elsevier 2008, pp. 821-845.
5. Selakov, A., Ilic, S., Vukmirovic, S., Kulic, F., Erdeljan, A.: A comparative analysis of SVM and ANN based hybrid model for short term load forecasting, *Transmission and Distribution Conference and Exposition, 2012 IEEE PES*, 7-10 May 2012, Orlando, FL. 5 p.
6. Koponen, P.: Short-term load forecasting model based on smart metering data, daily energy prediction using physically based component model structure, *IEEE SG-TEP 2012*, Nuremberg, 3-4 December 2012, 4 p.
7. Hahn, H., Meyer-Nieberg, S., Pickl, S.: Electric load forecasting methods: tools for decision making. *European Journal of Operational Research*, Vol. 199, Elsevier 2009, pp. 902-907.
8. Vapnik, V.N.: *The nature of statistical learning theory*. New York: Springer, 1995, 188 p.
9. Niska, H., Koponen, P., Mutanen, A.: Evolving smart meter data driven model for short-term forecasting of electric loads. *IEEE ISNNIP 2015*, Singapore, 7-9 April 2015, 6p.
10. Koponen, P.: *Optimisation of load control, Final report*, VTT Energy. Espoo, 20 November 1997, 26 p. + app. 14 p. Research report ENE6/12/97, <http://www.vtt.fi/inf/julkaisut/muut/1997/R-ENE6-12-97.pdf>

A new approach to nowcast economic time series using ensembles of hidden Markov and Arima models^{*}

Álvaro Gómez-Losada^[1] and Panayotis Christidis

European Commission, Joint Research Center (JRC), Economics of Climate Change, Energy and Transport Unit. Edificio Expo, c/ Inca Garcilaso 3, 41092 Seville, Spain

¹alvaro.gomez-losada@ec.europa.eu

Nowcasting is "forecasting" the recent or current state of a phenomenon. It has also been defined as the prediction of the present, the very near future and the very recent past [1]. The term is a contraction for now and forecasting and has been used for a long time in meteorology and recently in economics [2]. Nowcasting gains an important role in central banks and policymakers since key economic values are released with a considerable delay. This is mainly due to time-consuming information collection and aggregation processes. The underlying principle of nowcasting is to exploit the information which is published early in order to obtain an early estimate before the official figure becomes available. The different definitions of nowcasting suggest it is used in different ways and with different purposes, depending on the needs and availability of data. In this work, nowcasting was used to obtain an estimate of recently-released information. Even if the latter is known, the quantitative deviation between the estimated and real values can elucidate a change of pattern in the time series under consideration. Moreover, if the confidence interval associated with the estimated value does not include the real value, the time series could be experiencing a structural break, or at least, a certain degree of shift which is difficult to detect without a formal approach. This work aims to (i) propose a new method to estimate a confidence interval for the latest values of time series to detect structural changes in them, and (ii) introduce a new method to accomplish nowcasting using hidden Markov models.

By using past observations, two ensembles of models were implemented using R software [3] to obtain a confidence interval for the five most recent values of univariate economic time series.. The first ensemble used hidden Markov and Arima models, and the second included classic forecasting models for comparative purposes. A total of 15 heterogeneous economic time series from Eurostat [4] were used to test the nowcasting performance of both ensembles. The first ensemble obtained a much narrower confidence interval than the second one, and its higher performance to detect possible shift of trends in time series became evident.

^{*} The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

The performance of the first ensemble is robust through all the time series studied, revealing it to be a promising methodology for regular use in univariate time series nowcasting. However, in order to get a clearer picture, broadening the case studies to markedly irregular economic time series or with shorter lengths would be recommendable. In addition, the performance using different forecasting horizons should be evaluated. The classification as a structural break or shift pattern in the time series according to the nowcasted values obtained remains open for further research.

References

1. Gianone, D., Reichlin, L., Small, D. Nowcasting: the real-time information content of macroeconomic data. *Journal of Monetary Economics* 55(4), 665-676 (2008).
2. Bábura, M., Giannone, D., Modugno, M, Reichlin, L. Now-casting and the real-time data flow. Working paper series 1564. European Central Bank (2013).
3. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/> (2015), last accessed 29/03/2017.
4. Eurostat Homepage, <http://ec.europa.eu/eurostat>, last accessed 2017/03/28.

Ensemble Learning Framework for Predicting Project Cost Overrun Levels in Construction Procurement Auctions

Hyosoo Moon¹, Trefor P. Williams², and Moonseo Park¹

¹ Seoul National University, Seoul 08826, South Korea

² Rutgers University, Piscataway NJ 08854-8014, USA

axis1106@gmail.com

tpw@soe.rutgers.edu

mspark@snu.ac.kr

Abstract. From the project owner's point of view, it is important to screen for the projects expected to have excessive cost overruns in the bidding stage before contracts are finalized. Cost overrun during construction stems from various reasons such as macro-economic situations, construction site conditions, design errors, or owner's change orders. Bidding stage data related to project type and delivery methods also influence cost overruns. Previous studies have made efforts to improve both explanatory and predictive models of project cost overruns. Potentially, the knowledge of construction project characteristics combined with procurement auction factors can be used to enhance predictions of project performance. The objective of this study is to develop an ensemble classification model framework that predicts the expected level of cost overrun for public sector projects using a dataset consisting of 234 projects completed between 1998 and 2013 in Korea where project characteristics such as delivery method, project type, cost data and schedule information combined with bidding characteristics are available.

Keywords: Predicting Cost Overrun Levels, Ensemble Learning, Construction Procurement Auctions, Bidding Characteristics, Project Characteristics, Project Delivery Methods, Project Type.

1 Introduction

Project cost and schedule performance are influenced by the type of project delivery method and the type of construction. These factors also impact bidding characteristics in construction procurement auctions. Williams [1] found specific bidding patterns that would produce a cost increase during the construction phase with a hybrid model using a regression model prediction and a neural network. Moon [2] has studied cost increases and change orders¹ on Design-Build (DB) and Design-Bid-Build (DBB)

¹ Change order: a written order to the contractor signed by the owner and architect, issued after execution of the contract, authorizing a change in the work or an adjustment in the con-

delivery method projects using a statistical path analysis of projects in Korea. It was found that cost increases in both DBB and DB projects were significantly influenced by the ratio of the winning bid price to the owner's estimate. Projects, where there were significant differences between the low bid and the owner's pre-bid estimate, had higher cost overruns regardless of the delivery method. Moon also found that there were differences in cost performance between different types of constructions, such as civil and building construction. These findings suggest that factors other than the delivery method may have significant impacts on project costs and schedule as well. Potentially, the knowledge of project characteristics, such as the delivery method and type of construction project, combined with other factors such as the ratio of low bid to engineer's estimate can be used to enhance predictions of project performance. It can be inferred that a combination of factors affects the ultimate project cost outcome. The type of construction, the type of procurement method, the scope of the project, the accuracy of the bids, and the level of competition can all influence the level of overruns on a project. These factors can be categorized as project and bidding characteristics that impact change orders and cost overrun during construction phase (Fig. 1).

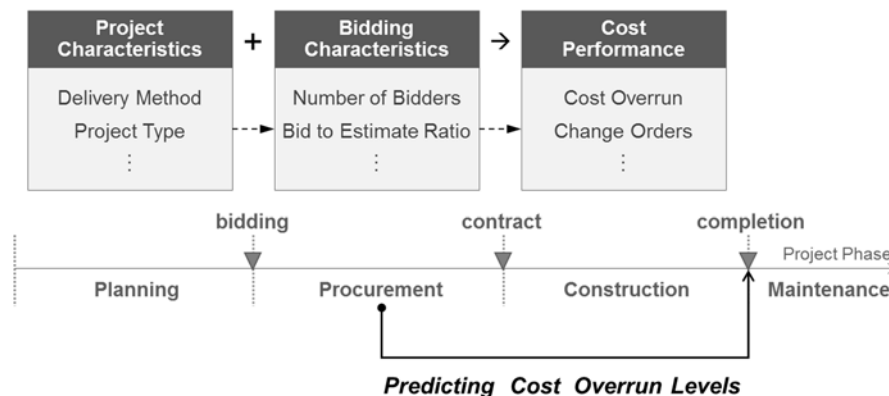


Fig. 1. Basic concept of influential factors on predicting project cost overrun levels

The objective of this paper is to use a database of 234 projects completed between 1998 and 2013 in Korea where data related to project and bidding characteristics are available, to develop an ensemble-classification model framework that predicts the expected level of cost overrun for public sector projects. The model framework developed uses two different classifiers whose outputs are combined to improve prediction results. The Korean data set consists of both DB and DBB projects, and additional information on the type of project being constructed.

tract sum of the contract time (article 12.1.1 of AIA A201). Moreover, the same terms are used in other countries; “variation order” was coined in UK, and FIDIC adopted the term “variation”.

2 Trends in Predicting Construction Project Cost

Various methods have been used to predict construction cost increases. Research that has been conducted in this area includes the application of artificial intelligence (AI) techniques and statistical methods to predict costs. Statistical methods have been used in a deductive way through hypothesis testing, while AI techniques have been used to induce the level of cost increase by using training data and model validation.

The best performances of both regression and neural network models were found using bid data to predict highway project costs [3]. For the regression model, only the low bid as input performed the best. The low bid and second lowest bid as inputs were used for the best performing neural network. Other approaches that use hybrid statistical methods and AI techniques were studied in cost overrun prediction areas. Emsley et al. [4] have developed linear regression techniques as a benchmark for the evaluation of the neural network models. Their models show that compared to the regression model, the neural network approach has the ability to model the nonlinearity in the data. Attalla and Hegazy [5] compared artificial neural networks (ANN) and regression for predicting cost deviation in reconstruction projects. They found that both models performed with similar accuracy. On the other hand, the ANN model is more sensitive to a larger number of variables. Some studies have investigated the development of prediction models using ANN and fuzzy neural networks [6], regression analysis compared traditional and weighted least-squares techniques [7], and Frequentist and Bayesian approaches [8]. Recent studies have employed advanced data-mining techniques to achieve a better prediction performance. Principal component analysis and support vector regression have been used to predict project costs [9]. The ensemble-learning method combines several methods to obtain a better predictive performance than any single method. Williams and Gong [10] used a text mining method where text from projects of the California Department of Transportation was processed and then transformed using support vector decomposition into a numeric value that was combined with other data and submitted to a stacking ensemble classifier. It was found that adding the text data improved the prediction result.

The trends survey shows that there has been an active interest in applying various modeling techniques to construction cost increases. However, it also implies that not only modeling techniques but also domain knowledge, such as construction project and bidding properties, should be combined to enhance predictions of project performance. This study will augment existing studies by considering how the project type and delivery method can improve cost predictions. Several factors are involved in determining project performance including the delivery method, the number of bidders, the type of project, the complexity of the project, accuracy of bids received, and the cost of the project. Therefore, this suggests the use of AI to classify combinations of project characteristics that produce overruns in project cost.

3 Feature Descriptions

The comprehensive understanding of project characteristics can improve the predictive models of cost overrun. This section will address data collection, descriptive statistics, project delivery method and project type as project characteristics.

This study collected a data set of 234 large construction projects costing more than 5 million dollars for each project that was awarded by the city of Seoul and completed between the dates of Jan 1, 1998 and April 30, 2013. The four types of project delivery methods acquired in the data set were categorized by Turn-key and Alternative method as DB, and Lowest bidding and Qualification as DBB method [11]. The market share analysis according to the contract year of 1992 - 2011 in Seoul is in Fig. 2. In addition, announcements of governmental policy on delivery methods were searched and matched up with the market share in the figure. It shows that the number of DB method projects increased due to governmental policy since 1992, as well as how both DB and DBB method market shares have similar rise and fall trends in the number of contracts.

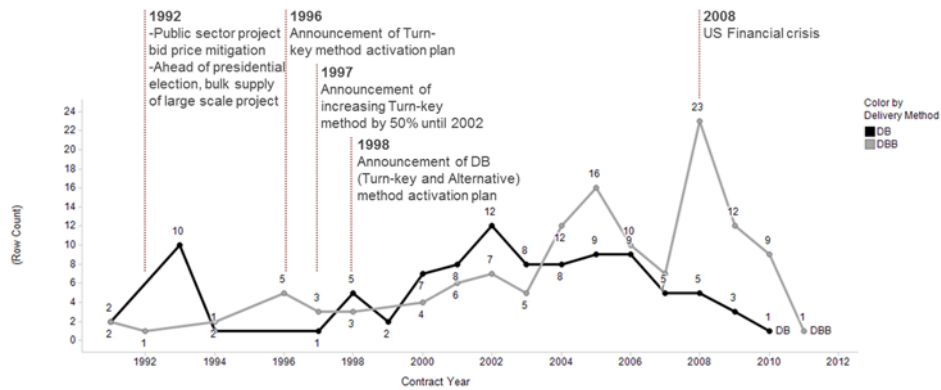


Fig. 2. Public sector project delivery method market share analysis in Seoul, South Korea (augmented from Moon 2015 [2])

The data available for the projects included two variables that describe the type of project being constructed. Table 1 shows the variable. Project Variable 1 is the general project type and Project Variable 2 is the sub-category of the general project type.

Table 1. Project variables

Project Variable 1	Project Variable 2
Civil	Sewerage
	Subway
	Road
	River
	Water Supply
Building	General Building

	Apartment
Facility	Facility
Landscaping	Landscaping

Table 2 shows that for the general Civil, Building and Facility project types, the DB delivery method had, on average, a lower level of cost growth due to change orders than DBB projects. Landscaping projects were only let using DBB.

Table 2. Average percentage cost growth due to change orders

Project Type 1	Mean Value(Std. Deviation)		Mean difference	Sample Size of DB/DBB
	DB	DBB		
Civil	13.43(23.39)	20.11(33.22)	6.68	49/61
Building	9.72(13.50)	14.49(11.28)	4.77	35/56
Facility	1.54(4.21)	20.73(22.97)	19.19	13/4
Landscape	-	36.67(24.41)	NA	0/16
Total	10.50(18.86)	19.76(25.71)	9.26	97/137

Table 3 shows that the superiority of the DB method is not clear for some specific types of construction. For road and water supply projects, the average cost increase was less than using DBB. The table also shows that subway projects in this data were only conducted as design build projects, while sewerage and landscape projects were only conducted as DBB.

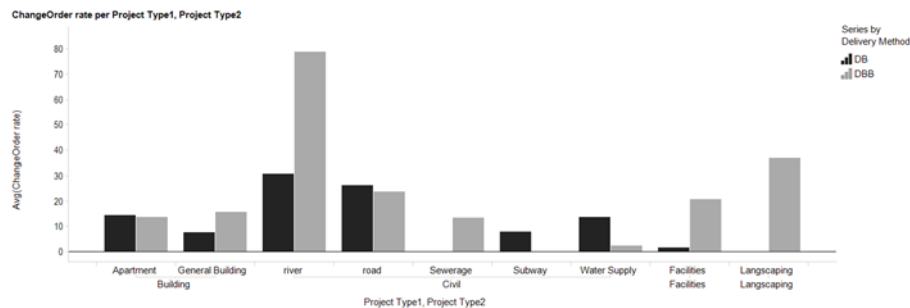
Table 3. Average percentage cost growth due to change orders for specific project types.

Project Type 2	Mean Value(Std. Deviation)		Mean difference	Sample Size of DB/DBB
	DB	DBB		
Sewerage	-	13.30(15.18)	NA	0/8
Subway	7.78(20.65)	-	NA	32/0
Road	26.26(30.09)	23.62(32.89)	-2.64	10/36
River	30.68(19.67)	78.84(70.29)	48.16	3/3
Water Supply	13.61(13.81)	2.38(9.95)	-11.22	4/14
General Building	7.63(14.24)	15.53(13.81)	7.90	24/26
Apartment	14.29(10.91)	13.58(8.67)	-0.71	11/30
Facility	1.54(4.21)	20.73(22.97)	19.19	13/4
Landscape	-	36.67(24.41)	NA	0/16
Total	14.54(16.23)	25.58(24.77)	11.04	97/137

To be more specific regarding the problem that the superiority of the DB method is not clear for some project types, we consider the project type as a confounding factor. Fig. 3(b) shows the visualization of Table 3 as a bar chart. It simplifies the comparison between DB/DBB itself (Fig. 3(a)) and DB/DBB with the project type (Fig. 3(b)).



(a) Average percentage cost growth due to change orders for only delivery method



(b) Average percentage cost growth due to change orders for delivery method with specific project types (bar chart visualization of data in Table 3)

Fig. 3. Project type as a confounding factor

4 Deriving Input Variables

To derive input variables affecting the project cost overrun, the analysis of field project work data (real-world data) and an extensive literature survey were conducted. They can be categorized as project and bidding characteristics. This section combines the former section results as the analysis of data set with an extensive literature survey to deal with the selection of input data.

Some of the existing research has identified factors that contribute to construction cost increases. Williams et al. [12] showed that there was a strong linear relationship between the natural log of the low bid and the natural log of the completed cost for highway projects in Great Britain and the United States. Skitmore and Ng [13] have developed different forms of regression models to forecast the actual construction cost and time; they found that client sector, contractor selection method, contractual arrangement, and project type could affect the final cost and time. Gkritza and Labi [14] have applied econometric models to the analysis of highway project cost overruns. They found that for a given project type and project duration, contracts of larger size or longer duration were generally more likely to incur cost overruns. These input data

used by previous researchers including section 2 literatures are chronicled in Table 4, and the possible variables that are matched with the data set are underlined.

Table 4. Input data used by previous researchers

Researcher	Project type(sample size)	Input data
Jahren and Ashe 1990 [15]	Naval facilities(U.S. 1576)	Project size, <u>the difference between the low bid and government estimate</u> , <u>the type of construction</u> , the level of competition
Williams et al.1999 [12]	Highway(UK 28, the U.S. 90)	<u>Low bid</u>
Williams 2002[1]; 2005[3]	Highway(NJ, 302); Highway(TX, 1260)	Low bid, median bid, <u>expected project duration</u> , <u>number of bids</u>
Attalla and Hegazy 2003 [5]	Reconstruction project (Canada, 50)	36 variables(scope definition and planning, tendering stage, <u>schedule</u> , cost, quality, communication, safety)
Skidmore and Ng 2003 [13]	Australian construction projects(various,93)	<u>Client sector</u> , <u>contractor selection method</u> , contractual arrangement, <u>project type</u>
Ling et al. 2004 [16]	Residential(Singapore,87)	59 variables, <u>delivery methods(DB/DBB)</u>
George et al. 2005 [6]	Industrial construction projects (U.S. 50)	25 variables(project size, contract type, relative level of complexity, site conditions, <u>design schedule</u>)
Gkriska and Labi 2008 [14]	Highway(Indiana, 1957)	<u>Project type</u> , <u>project duration</u> , contract size
Son et al. 2012 [9]	Commercial buildings(84)	64 variables (pre-project planning stage: <u>project type</u> , project size, <u>project duration</u>)
Williams and Gong 2014 [10]	Highway(California,1221)	Low bid, the completed project cost, <u>the number of bidders</u>
Sousa et al. 2014 [7]	Sanitation(Chicago,180)	<u>Delivery method(DBB)</u> , <u>project type</u> (water/ sewer)

From the combination analysis of the data set and extensive literature survey, inputs to the model were categorized into two characteristics and selected (Table 5).

Table 5. Input data to the model

Characteristic	Input data
Project characteristics	Delivery method (DB, DBB) Project type as defined by two variables (described in Table 1) Initial schedule in days
Bidding characteristics	Ratio of the bid price to the owner's pre-bid estimate Selected bid amount Award method (qualification based award or award to the lowest bidder) Number of bidders Number of companies forming a joint venture to construct the project

5 Methodology

Figure 4 shows the structure of the ensemble-learning model. The model is supposed to produce as output a prediction of the level of cost overruns that will occur during construction executions.

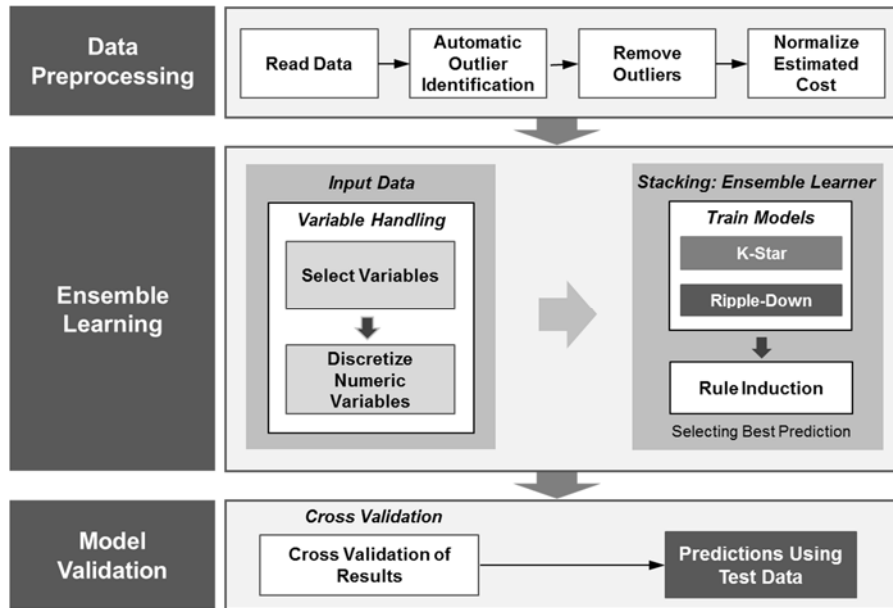


Fig. 4. The structure of the ensemble-learning model

5.1 Ensemble Classifiers and Stacking

Ensemble classifiers can provide improved prediction results by using several different classifiers and combining their results using an algorithm to select the best prediction. This model uses the stacking technique where an ensemble of classifiers is first created, whose outputs are used as inputs to a second level meta-classifier to learn the mapping between the ensemble outputs and the actual correct classes [17]. In this model, an if-then rule is induced to select which of the two model predictions to use.

5.2 Classification Algorithms

Two different classification algorithms were used to classify the project data—the Ripple-Down and the K-Star classification algorithms. The Ripple-Down algorithm automatically generates a set of classification rules from the input data [18]. It learns rules with exceptions by generating a default rule. The default or top-level rule is the class of the output that occurs most frequently. The algorithm then uses incremental

reduced-error pruning to find exceptions with the smallest error rate, finding the best exceptions for each exception and iteration [19].

K-Star is a type of algorithm called a “lazy learner.” The K-Star algorithm is an instance-based learning scheme developed by Cleary et al. [19]. Witten and Frank [20] describe the K-Star algorithm as a lazy classifier where the training instances are stored and are not employed until the classification time. They also state that K-Star uses a nearest neighbor method with a generalized distance function.

5.3 Normalization of the Original Estimate Data

The original estimate values vary widely in magnitude. It was found that normalizing the original bid amount improved the accuracy of predictions. The purpose of statistical normalization is to convert data into a normal distribution with mean = 0 and variance = 1. The formula for statistical normalization is

$$Z = (X - u) / s \quad (1)$$

Vector ‘X’ denotes the attribute values, ‘u’ is the mean of the attribute values, and ‘s’ is the standard deviation. Using this formula, we can get another vector ‘Z’ that has a normal distribution with zero mean and unit variance. It is also called the standard normal distribution, $N(0, 1)$.

5.4 Discretization of Numeric Data

This discretization is performed by simple binning. The range of numerical values is partitioned into segments of equal size. Each segment represents a bin. Numerical values are assigned to the bin representing the segment covering the numerical value. Four bins were used for each numerical variable.

5.5 Cross Validation

The data set of 234 projects is relatively small. Therefore, a cross validation scheme is needed to test and validate the model, and to prevent overtraining on training data. The cross validation operator is a nested operator. It has two sub-processes: a training sub-process and a testing sub-process. The training sub-process is used for training the cost prediction model. The trained model is then applied in the testing sub-process to make predictions. The performance of the model is also measured during the testing phase.

The data from the projects needs to be partitioned into k subsets of equal size. Of the k subsets, a single subset is used as the testing data set and the remaining $k - 1$ subsets are used as the training data set. The cross validation process is then repeated k times, with each of the k subsets used exactly once as the testing data. The k results from the k iterations can then be averaged (or otherwise combined) to produce a single estimation.

6 Conclusions and Future Work

This study provides the ensemble-learning model framework to predict project cost overrun levels in construction procurement auctions. Although the data is from the city of Seoul, this model could adapt the risk levels of other cities/countries, because the data includes data from both DB and DBB projects, which are most prevalent world-wide, and also includes data about the type of construction. Analysis of this data is supposed to compare the overrun level between DB and DBB projects according to project types. With the results of the developed model, project owners and project managers can check the projects expected to have excessive cost increase and prepare for loss of budget before contract completion during bidding phase.

However, since this model is in the beginning stage, the following process needs to be conducted for the near further studies: the ensemble-learning model should show the precision and recall for the prediction of the levels of cost overrun. Finally, the important contributors to the better-accuracy of the predictions would be shown in the further studies.

References

1. Williams, T. P.: Predicting completed project cost using bidding data. *Construction management and economics*, 20, 225-235 (2002).
2. Moon, H.: Cost performance comparison of design-build and design-bid-build focusing on mediator effect. M.S. thesis, Seoul National University, Seoul, South Korea (2015).
3. Williams, T. P.: Bidding ratios to predict highway project costs. *Engineering, construction and architectural management*, 12(1), 38-51 (2005).
4. Emsley, M. W., Lowe, D. J., Duff, A. R., Harding, A., Hickson, A.: Data modeling and the application of a neural network approach to the prediction of total construction costs." *Construction management and economics*, 20, 465-472 (2002).
5. Attalla, M., Hehazy, T.: Predicting cost deviation in reconstruction projects: artificial neural networks versus regression, *J. Constr. Eng. Manage.*, 129(4), 405-411 (2003).
6. Georgy, M.E., Chang, L., Zhang, L.: Prediction of engineering performance: a neurofuzzy approach, *J. Constr. Eng. Manage.*, 131(5), 548-557 (2005).
7. Sousa, V., Almeida, N. M., Dias, L., Branco, F. A.: Risk-Informed Time-Cost Relationship Models for Sanitation Projects, *J. Constr. Eng. Manage.*, 140(5), 06014002:1-5 (2014).
8. Behmardi, B., Doolen, T., Winston, T.: Comparison of Predictive Cost Models for Bridge Replacement Projects, *Journal of Management in Engineering*, 31(4), 04014058:1-7 (2015).
9. Son, H., Kim, C., Kim, C.: Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables. *Journal of Automation in Construction*, 27, 60-66 (2012).
10. Williams, T. P., Gong, J.: Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in construction*, 43, 23-29 (2014).
11. Lee, Y. S.: Suggestions for Improvement of Construction Project Awarding Methods in Public Sector. Korea Research Institute for Human Settlements (KRIHS), *Construction Economy Articles* (winter), 75, 43-54 (2013).

12. Williams, T. P., Miles, J. C., Moore, C. J.: Predicted cost escalations in competitively bid highway projects. *Proceedings of the Institution of Civil Engineers. Transport*, 135, 195-199 (1999).
13. Skitmore, R. M., Ng, S. T.: Forecast models for actual construction time and cost. *Building and Environment*, 38, 1075-1083 (2003).
14. Gkritza, K., Labi, S.: Estimating Cost Discrepancies in Highway Contracts: Multistep Econometric Approach. *J. Constr. Eng. Manage.*, 134(12), 953-962 (2008).
15. Jähren, C. T., Ashe, A. M.: Predictors of Cost-Overrun Rates. *J. Constr. Eng. Manage.*, 116(3), 548-552 (1990).
16. Ling, F. Y. Y., Chan, S. L., Chong, E. L. P.: Predicting Performance of Design-Build and Design-Bid-Build Projects. *J. Constr. Eng. Manage.*, 130, 75-83 (2004).
17. Polikar, R.: Ensemble based systems in decision making. In: *IEEE circuits and systems magazine*, pp. 21-45, IEEE Press, New York (third quarter 2006).
18. Gaines, B. R., Compton, P.: Induction of ripple-down rules applied to modeling large databases, *J. Intell. Inf. Syst.* 5 (3) 211–228 (1995).
19. Cleary, J. G., Trigg, L. E.: K*: An Instance-based Learner Using an Entropic Distance Measure. In: *Proceedings of 12th International Conference on Machine Learning*, pp. 108-114. Morgan Kaufmann Publishers, San Francisco (1995).
20. Witten, I. H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco (2005).

Time Series Forecasting applying Data Transformation and Neural Networks Ensembles

German Gutierrez¹, Paz Sesmero², Araceli Sanchis³

Computer Science Department
Carlos III University of Madrid, Spain
{ggutierrez¹, msesmero², masmas³}@inf.uc3m.es

Abstract. This paper is focused in univariate time series forecasting. In this work, we present a framework which includes alternative transformations to the original time series observations data to carry out the forecasting task. This method is based on transforming the original observation time series which turns into a time series of the difference between two consecutive values, or the one that indicates if there is an increment or a decrement between two consecutive values. A specific and different model that maps the future values and past values is obtained and applied for each of the time series, the original observations and the transformed time series. And finally, the answer given by each of the models is merged to get the final one step ahead forecasted value. To fit a model between independent variables (present and past known values) and dependent variables (future unknown values), the Artificial Neural Networks can accomplish suitable results. Each of the models committed for the different original and transformed time series can be made-up of a single neural net or a combination of several nets, i.e. an Ensemble of neural nets. This contribution shows the ongoing experimentation performed to evaluate if the system with different ensembles for original and transformed time series gets a better result than applying single nets as model to forecast the original observations time series.

Keywords: Artificial Neural Networks, Time Series Forecasting, Taguchi's method, Ensembles.

1 Introduction

The univariate time series forecasting task lies in computing the unknown future values of a measure by the application of a model f . The inputs (independent variables) of the model are the k known previous values of the measure to time t , and it will obtain h values ahead (horizon).

$$[y_{t+h}, \dots, y_{t+1}] = f(y_t, y_{t-1}, \dots, y_{t-k-1}) \quad (1)$$

The Artificial Neural Networks (ANNs) is a well-known and reliable technique to map functions, and have been widely applied for system modelling and carry out predictions. In time series forecasting, ANNs is considered an advance method as they provide a mapping approach for nonlinear relationships. In particular, when ANN are applied to

adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

univariate time series forecasting, the ANN inputs (dependent variables) are the k known present and past values of a measure got throughout the time. And the output is the h unknown future values ahead from time t . That is, the ANN is an approach to model the function f in eq(1).

There are also statistical techniques to model the relationship between past and future values. Some of these techniques take not only the original observations (y_t) to accomplish a model, for instance, they take the residuals, the error on the forecast value ($e_t = \hat{y}_t - y_t$) into the model as in autoregressive integrated moving average (ARIMA) model [1]. Also, some other models include the relation between two consecutive values of the original time series ($y_t - y_{t-1}$), so the differencing and the original data can be merged into the model.

In this work, we apply the original observations time series and two additional transformed time series data: differencing (*dif*) and increment (*inc*) data, to generate a forecast on the one step ahead time series value (y_{t+1}). For each of the three time series *raw* (i.e. no transformation), *dif* and *inc*, a different forecasting model is obtained through a supervised learning process, and then combine their answer to forecast the future values for original observations. Each of the models is focused only on its particular time series. In this approach, up to now, the models for any of the time series involved (*raw*, *dif*, and *inc*) are based in Artificial Neural Networks.

The purpose of this paper is to present an ongoing experimentation performed on this framework to show that it is an alternative method to carry out time series forecasting. This is bringing out evaluating the performance of a single ANN for *raw* data, the combination of single ANN for each of *raw*, *dif*, and *inc*, and finally, the combination of different Ensembles again for each of *raw*, *dif*, and *inc* time series data.

2 Approach Description

The approach developed in this work lies in three phases, see **Fig. 1**. In the first phase: the original observations time series (y_t^{orig}), i.e. the raw data (y_t^{raw}), is transformed in differencing (y_t^{dif}) data and increment (y_t^{inc}) data.

$$\begin{aligned} y_t^{raw} &= y_t^{orig} \\ y_t^{dif} &= y_t^{orig} - y_{t-1}^{orig} \\ y_t^{inc} &= \begin{cases} 1 & , \text{ if } y_t^{orig} > y_{t-1}^{orig} \\ 0 & , \text{ if } y_t^{orig} = y_{t-1}^{orig} \\ -1 & , \text{ if } y_t^{orig} < y_{t-1}^{orig} \end{cases} \end{aligned} \quad (2)$$

In second phase, a different model is obtained for each of the *raw*, *dif* and *inc* time series data. In each of these models, for *raw*, *dif* and *inc* independently, the relation between the past values (input of the model) and the future values (output of the model) is mapped. So, each model, *raw*, *dif* and *inc*, means an approximation for its function f^{raw} , f^{dif} and f^{inc} respectively, eq. (3),

$$\begin{aligned}
y_{t+1}^{raw} &= f^{raw}(y_t^{raw}, y_{t+1}^{raw}, \dots, y_{t-kr+1}^{raw}) \\
y_{t+1}^{dif} &= f^{dif}(y_t^{dif}, y_{t-1}^{dif}, \dots, y_{t-kd+1}^{dif}) \\
y_{t+1}^{inc} &= f^{inc}(y_t^{inc}, y_{t-1}^{inc}, \dots, y_{t-ki+1}^{inc})
\end{aligned} \tag{3}$$

In the third and final phase, the answers given by the model of each unit in phase 2 (after a simple post-process to get the answer on original observations data) are combined to get the final answer. Following some details for phase 2 and phase 3 are indicated.

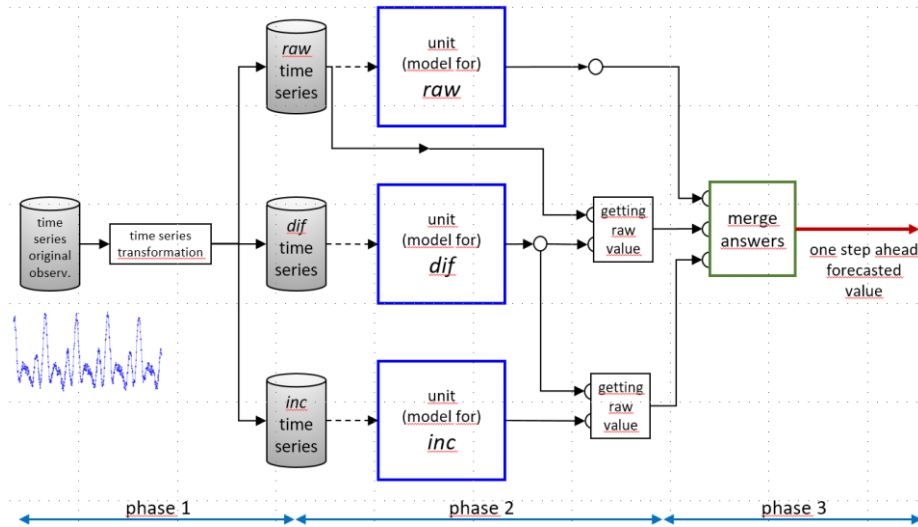


Fig. 1. Schema of the system and its three phases.

An essential issue when an ANN is applied is its hyper-parameter (other than connection weights, which are computed through learning algorithms based on backpropagation) setting to maximize its performance. We can find within the literature several methods to manage this issue, but they involve a kind of optimization or search process which implies a high computational cost. In this work, we apply as hyper-parameter setting procedure the Taguchi's method, a technique from Design of Experiments (DoE) [2], instead of the full-factorial design. The factor (parameter) and level (parameter value) combination given by DoE means 32 combinations for 5 hyper-parameters: number of inputs, hidden neurons, learning algorithm (resilient backpropagation and scaled conjugate gradient backpropagation), learning rate and training cycles. The full factorial design means 1024 combinations.

In this contribution, each model for the second phase will be an ensemble of ANN, i.e. a different ensemble of ANN is obtained as the model for each of the *raw*, *dif* and *inc* units. These Ensembles comprise some or all of the elements of the set of nets obtained through the factor-level combination applying the Taguchi's method.

As final step in phase 2, of course, the ANN which is focused on *dif* or *inc* data needs the previous raw value to give forecast value for original observations, eq (4). In this

equation the parameter r means if the output of the model is rounded just to 1 or +1, or not taking just the output of the model (e.g. 0.789)

$$\begin{aligned} y_{t+1}^{dif} &= dif_{t+1} + y_t \\ y_{t+1}^{inc} &= round(inc_{t+1}, r) * abs(y_{t+1}^{dif}) + y_t \end{aligned} \quad (4)$$

The aim of the third phase is to consider the three different units to forecast the *raw*, *dif* and *inc* time series data obtained in phase 2 and mix their answers to give a final forecasted value. The third phase is only executed for the unknown future values to be forecasted (e.g. the test subset, or the final values in a time series competition). However, the combination of the three models from phase 2, eq. (3) is established based on the error for validation subset (as generalization abilities), which is part of the known time series values.

$$\hat{y}_t = \sum_{j=1}^n (w_j \times y_t^j), \quad n = 3 \quad (5)$$

The weight of each model depends on its own performance and the performance of the other units on validation subset. The formula for the weights is shown in eq. (4), where n is the number of units (in this work $n = 3$), and e_val is the error on validation subset.

$$w_j = \frac{1}{n-1} \times \frac{(\sum_1^n e_val_i) - e_val_j}{\sum_1^n e_val_i}, \quad n = 3 \quad (6)$$

3 Experimental Procedure and Results

In this section, first we are going to explain the experimental setup, and then show the experimental results.

3.1 Experimental Setup

The framework shown in this work is evaluated in four time series widely applied in the literature: Mackey-Glass, Dow-Jones, Quebec, and Temperature. The last three time series has been collected from Hyndman's Time Series Data Library [3]. For each time series, the whole data is split in three subsets: train, validation, and test **Table 1**. Train subset is used in phase 2 by the learning algorithm to modify and fixed the ANN weights. Validation subset allows to state, in phase 2, the generalization ability of the ANN result from learning process. Test subset is never seen by any ANN in the learning process, so it is incorporated to test the generalization ability of the system, and taking into account only in the third phase.

Table 1. Time Series, percentages and number of values for train, valid and test subsets.

Time Series	Training (60%)	Validation (20%)	Test (20%)
Dow-Jones	441	146	146
Mackey-Glass	97	30	30
Quebec 584	438	146	146
Temperature	144	48	48

In this work the accuracy measure to evaluate each ANN or Ensemble and the whole system is the root mean square error (*rmse*). Additionally to accuracy measures, to assess whether differences in accuracy between two models (single ANN, Ensembles or any combination of them combination) are statistically significant we use the Diebold-Mariano test [4]. This test analyzes if the difference of expected losses between two models is zero against the alternative that one model is better. If the absolute value obtained for the Diebold-Mariano test (*dm*) for the forecasted values given by two different models is greater than 1.96 ($|dm| > 1.96$) the forecast of both models are statistically different.

In this work, the benchmark model is a single ANN for the original time series observations (raw data), in particular the best ANN (best *rmse* for validation subset) obtained from the 32 different combinations of hyper-parameters for the raw time series data.

As indicated above, to set the ANN hyper-parameters within each unit in phase 2, instead a full factorial combination of the parameters selected (input nodes, hidden nodes, learning algorithm, learning rate and training cycles), a factor-level combination based on Latin-squares is taking into account which means a much smaller (about 1.5 order of magnitude lower) number (32) of combinations from the full factory (1024).

3.2 Experimental Results

In this contribution we show the results obtained for the four time series indicated above for three different options setting the units (raw, dif and inc) in phase 2: i) the best single net for each unit ii) an Ensemble of the 32 nets obtained applying Taguchi Method for hyper-parameter setting (note that the elements of the ensemble for each unit from phase 2 are different and focus on its specific forecast task, raw, dif or inc data), see **Table 2** for the 32 factors-levels combinations, to get the hyper-parameters; iii) for each unit (*raw*, *dif*, *inc*) an Ensemble of any 4 ANN of the 32 nets obtained applying Taguchi Method for hyper-parameter setting.

Due to limited space, for one of the time series, we show the *rmse* (root mean square error) for validation and test subsets, obtained by two options to set the element(s) of each units at phase 3 (rounding or not the output of unit *inc*, and what units are combined *wO*, 1 for *raw*, 2 for *dif* and 3 for *inc*). These results are shown in **Table 3**

To evaluate the performance of the system bearing in mind the three options i), ii) and iii) already indicated for phase 2, we can take into account the following options for their combination in phase 3:

- (a) The best net from the 32 different nets obtained applying Taguchi's Method for raw (original observations) data. This is the reference benchmark is this study.

- (b) The combination of the best net of each unit (*raw*, *dif*, *inc*) from the 32 nets applying Taguchi's Method.
- (c) The net of the unit selected (*raw*, *dif*, *inc*) with lowest *rmse* error for validation subset. The forecast on test subset for the system would be the answer of this unit on test subset.
- (d) The combination of the three units, comprises each of them by a different Ensemble.
- (e) The ensemble from the unit selected (*raw*, *dif*, *inc*) with lower *rmse* error for the validation subset. The forecast on test subset for the system would be the answer of this unit on test subset.

Table 2. List of the 32 factors-levels combinations, to get the hyper-parameters

<i>in</i> : inputs; <i>hn</i> : hidden nodes; <i>la</i> : learning algorithm; α : learning rate; <i>trc</i> : training cycles.																	
<i>run</i>	<i>in</i>	<i>hn</i>	<i>la</i>	α	<i>trc</i>	<i>run</i>	<i>in</i>	<i>hn</i>	<i>la</i>	α	<i>trc</i>	<i>run</i>	<i>in</i>	<i>hn</i>	<i>la</i>	α	<i>trc</i>
01	05	05	05	0,20	05	12	15	20	05	0,10	05	23	10	35	06	0,01	05
02	05	10	06	0,10	10	13	20	05	05	0,01	10	24	10	40	05	0,05	10
03	05	15	05	0,05	20	14	20	10	06	0,05	05	25	15	25	06	0,05	50
04	05	20	06	0,01	50	15	20	15	05	0,10	50	26	15	30	05	0,01	20
05	10	05	06	0,10	20	16	20	20	06	0,20	20	27	15	35	06	0,20	10
06	10	10	05	0,20	50	17	05	25	05	0,20	05	28	15	40	05	0,10	05
07	10	15	06	0,01	05	18	05	30	06	0,10	10	29	20	25	05	0,01	10
08	10	20	05	0,05	10	19	05	35	05	0,05	20	30	20	30	06	0,05	05
09	15	05	06	0,05	50	20	05	40	06	0,01	50	31	20	35	05	0,10	50
10	15	10	05	0,01	20	21	10	25	06	0,10	20	32	20	40	06	0,20	20
11	15	15	06	0,20	10	22	10	30	05	0,20	50	-	-	-	-	-	-

In (c) and (e), a selection based on validation subset performance is applied as combining method to get the output of the system. In (b) and (d) eq. (3) and eq. (4) are applied to get the answer of the system in phase 3.

To evaluate the framework shown in this contribution we have carried out different comparisons, which are shown in **Table 4**. Note that the options (d) and (e) are the ones that comprises ensembles in each unit in phase 2.

The results for these comparisons are shown in **Table 5**. The columns *dm1* to *dm6* are the Diebold-Mariano test value for comparisons c1 to c6 of **Table 4**. The columns *T1* to *T6* indicates if the forecasted values are statistically different for each of the comparisons. The "+" means that the forecast made by the option related with ensembles ((d) or (e)) is statistically different (its absolute value for *dm* test is greater than 1.96) and its *rmse* on test data is better (lower) than the option for single nets ((a), (b) and (c)). The "=" means that the forecast made by the option related with Ensembles ((d) or (e)) is statistically the same than (a), (b) and (c). And "-" means that gets a statistically different and worse result. The column *rO* means if the output of the unit *inc* is round or not to 1 or -1. The column *wO* means which are the answers combined in phase 3 (1: *raw*; 2: *dif*; 3: *inc*).

Table 3. For Dow-Jones (dj), results for the best single ANN from 32 combinations of the parameters in Taguchis's method, and for the ensemble comprises by the 32 nets obtained in Taguchis's method.

			single best ANN for each unit (raw, dif, inc)								
param			rmse - validation				rmse - test				
ts	rO	wO	raw	dif	inc	comb	raw	dif	inc	comb	
dj	0	12	26,199	24,949	26,321	23,041	149,917	28,814	27,803	78,606	
dj	0	13				24,240				80,140	
dj	0	123				23,578				57,906	
dj	1	12			26,347	23,041			28,522	78,606	
dj	1	13				24,109				80,652	
dj	1	123				23,400				58,118	
			Ensemble of 32 nets for each unit (raw, dif, inc)								(a) vs (d)
dj	0	12	30,548	27,253	26,811	25,802	152,116	27,148	26,965	78,120	=
dj	0	13				25,730				77,012	+
dj	0	123				25,534				57,612	=
dj	1	12			26,755	25,802			27,144	78,120	=
dj	1	13				25,787				77,154	+
dj	1	123				25,543				57,726	=

Table 4. Comparisons carried in the experiments on test subset forecast

c1: (a) vs (d)	c4: (b) vs (e)
c2: (a) vs (e)	c5: (c) vs (d)
c3: (b) vs (d)	c6: (c) vs (e)

The result shown in the final columns of **Table 5** indicates than 46% of the configurations through the four time series in this experimentation using Ensembles as model within the units in phase 2 get a better result than using single nets, almost half of the times (52%) of time gets a similar result, and only 1.4 % get a worse results. Additional comparisons have been carried out when the number of elements of the ensembles are limited (e.g. 4, 8, 16 and 20 nets) where similar results have been obtained.

4 Conclusions and future works.

In this work, we show an approach to intend an alternative to the ANN (or model) that learns only from the raw time series data. This alternative mix the model obtained for original observation (raw) time series data, the differential (*dif*) time series data and the increment (*inc*) time series data. Also, we evaluate if by means of ensembles as specific model for *raw*, *dif*, and *inc* unit gets better results. In fact, more than 98% of the times the system with Ensembles as model for each unit gets better (40%) or similar results (58%).

Table 5. Result for comparisons (dmi and Ti are for ci in Table 4) of best net for each unit in vs Ensemble with the 32 nets obtained from Taguchi's Method.

nan means that the selected unit is raw. DJ:(Dow-Jones), MG (Mackey-Glass), Q (Quebec), T(Temperature)																Sum through Comparisons		
ts	rO	wO	dm1	dm2	dm3	dm4	dm5	dm6	T1	T2	T3	T4	T5	T6		(+)	(=)	(-)
DJ	0	12	5,220	5,055	0,323	5,220	0,323	5,055	+	+	=	+	=	+		4	2	0
DJ	0	13	5,197	5,056	3,267	5,197	3,267	5,056	+	+	+	+	+	+		6	0	0
DJ	0	123	5,214	5,060	0,277	5,214	0,277	5,060	+	+	=	+	=	+		4	2	0
DJ	1	12	5,220	5,055	0,323	5,220	0,323	5,055	+	+	=	+	=	+		4	2	0
DJ	1	13	5,205	5,047	2,618	5,205	2,618	5,047	+	+	+	+	+	+		6	0	0
DJ	1	123	5,217	5,055	0,462	5,217	0,462	5,055	+	+	=	+	=	+		4	2	0
MG	0	12	6,861	6,905	5,339	nan	1,097	nan	+	+	+	=	=	=		3	3	0
MG	0	13	6,840	6,953	6,092	nan	1,097	nan	+	+	+	=	=	=		3	3	0
MG	0	123	6,504	6,614	2,949	nan	1,097	nan	+	+	+	=	=	=		3	3	0
MG	1	12	6,861	6,905	5,339	nan	1,097	nan	+	+	+	=	=	=		3	3	0
MG	1	13	6,808	6,900	5,432	nan	1,097	nan	+	+	+	=	=	=		3	3	0
MG	1	123	6,446	6,515	3,161	nan	1,097	nan	+	+	+	=	=	=		3	3	0
Q	0	12	2,985	2,097	1,983	nan	3,344	nan	+	+	+	=	+	=		4	2	0
Q	0	13	3,236	2,420	0,812	3,236	0,812	2,420	+	+	=	+	=	+		4	2	0
Q	0	123	0,943	0,174	0,573	nan	1,026	0,174	=	=	=	=	=	=		0	6	0
Q	1	12	2,985	2,097	1,983	nan	3,344	nan	+	+	+	=	+	=		4	2	0
Q	1	13	2,982	1,946	0,746	2,982	0,746	1,946	+	=	=	+	=	=		2	4	0
Q	1	123	0,329	-0,632	-0,468	nan	3,344	nan	=	=	=	=	+	=		1	5	0
T	0	12	0,760	-0,583	0,871	nan	2,137	nan	=	=	=	=	+	=		1	5	0
T	0	13	0,469	-0,629	1,004	nan	2,137	nan	=	=	=	=	+	=		1	5	0
T	0	123	-1,150	-2,090	0,468	nan	2,137	nan	=	-	=	=	+	=		1	4	1
T	1	12	0,760	-0,583	0,871	nan	2,137	nan	=	=	=	=	+	=		1	5	0
T	1	13	0,375	-0,687	1,065	nan	2,137	nan	=	=	=	=	+	=		1	5	0
T	1	123	-1,216	-2,193	0,310	nan	2,137	nan	=	-	=	=	+	=		1	4	1
																67	75	2

5 Conclusions and future works.

In this work, we show an approach to intend an alternative to the ANN (or model) that learns only from the raw time series data. This alternative mix the model obtained for original observation (raw) time series data, the differential (*dif*) time series data and the increment (*inc*) time series data. Also, we evaluate if by means of ensembles as specific model for *raw*, *dif*, and *inc* unit gets better results. In fact, more than 98% of the times the system with Ensembles as model for each unit gets better (40%) or similar results (58%).

Among the future works, we must extend the experimentation applying a method to get a candidate for best single model at all (brute force, or a metaheuristic search as evolutionary computation), and to an additional number of time series to endorse the result from this works. Also, the same framework could be applied but in this case with different computational intelligence or machine learning techniques, for instance Support Vector Machines, to get a model of one or each of the units. And additionally, we could apply an additional machine learning (artificial neural networks, or support vec-

tor machines) to learn how to combine the model (Stacking). Also, different transformations of the original observation can be added to the framework, which means both additional units in phase 2 and new answers to merge in phase 3.

Acknowledgement

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness under the following projects: TRA2015-63708-R, and TRA2016-78886-C3-1-R

References

1. Hyndman, R.J. and Athanasopoulos, G. (2013) Forecasting: principles and practice. <http://otexts.org/fpp/>. Accessed on April 2017
2. NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/pri/pri.htm>, Section 5.5.6. What are Taguchi designs? Last access April 2017
3. Hyndman, R.J. "Time Series Data Library", <http://data.is/TSDLdemo>. Accessed on May 2017
4. Diebold, F.X. and R.S. Mariano (1995), "Comparing predictive accuracy", Journal of Business & Economic Statistics, 13, 253-263.

Dynamics of Memory in Investor Attention to Energy Market

Ravi Prakash Ranjan* and Malay Bhattacharyya

Indian Institute of Management Bangalore, 560076, India

Abstract. In this article, we investigate the correlation structure of the time series of investor attention as measured by relative search query volume of stocks in Google. Specifically, we explore - i) Whether the time series has a power law correlated dependence (long range memory) and how does it evolve over time? ii) How does this dependence vary with frequencies of sampled data? iii) Does a cross-correlation dependence exist between local and global investor attention? iv) What happens to this memory structure in case of volatility clustering periods of price and volume? We perform detrended fluctuation analysis and detrended cross-correlation analysis of the time series of investor attention of top 20 energy companies (by their market capitalization). The results confirm the existence of long range dependence in investor attention. The memory dynamics are characterized by persistent and mean-reverting behavior. There is a reasonably high positive cross-correlation dependence between local and global investor attention. Finally, we observe that volatility clustering has little effect on long range dependence structure of investor attention.

Keywords: Investor Attention, Google Trends, Fluctuation Analysis, Power law dependence

1 Introduction

When the New York times reported the breakthrough in cancer research on 3rd May 1998, the stock price of EntreMed's surged by 300 % [1]. Although the article was there in the journal *Nature* and some other newspapers five months back, the market remained under reacted till it appeared in *Times*. This news not only affected EntreMed but other biotechnology firms witnessed a considerable increase in their stock price as well. This suggests that mere an availability of information does not get reflected in prices unless enough attention is paid to it by the relevant people (like investors). Hence, the investor attention must play a crucial role in determining market movement and efficiency. Furthermore, attention is a limited cognitive resource available to us [2]. So, even if there is a huge volume of information available, investors have no choice but to select only specific set of information and make their investment decisions.

* Corresponding Author, Email: ravi.ranjan14@iimb.ernet.in

In this article, we study the memory associated with the time series of investor attention. An investigation on whether the time series has noise, short memory or long memory shall have direct implications while modeling the relationships of investor attention and other variables. A number of studies are available on long run memory characteristics of stock market variables like - stock prices [3], [4], stock returns [5], [6], stock volume [7], [8], stock volatility [9], [10] and conditional variance of stock returns [11]. In case the time series has no memory i.e. it's a pure noise, the series cannot be used for any kind of predictive modeling. The existence of short memory in a time series implies that the effect of exogenous variable or shock to the series is short lived and dissipates very fast [12]. Long memory in a time series implies that its autocorrelations decay slowly, making it efficacious for modeling and analyzing the relationship with other variables. Understanding of long memory is important and special because it is often absent in most of the stochastic processes [13]. Existing literature primarily covers the impact of investor attention on other stock market variables, volatility and returns predictability [14], [15], [16]. In a very recent article, Xiaoquian Fan et. al analyzed Baidu search engine based investor attention index and its cross correlations with trading volume and volatility [17]. In this article we analyze noise and long range memory structure for investor attention based on the relative volume of Google search queries.

The main focus of this article is to carry out an in depth analysis of this dependence structure rather than predictions. With respect to memory in investor attention time series, specifically we explore the following - **a) Existence & Dynamics:** Whether the time series has a power law correlated dependence (long range memory) and how does it change over time? **b) Sampling Frequency:** How does the dependence structure vary with frequency of sampled data? **c) Local Vs Global Investor Attention:** Does a cross correlation dependence exist between local and global investor attention? **d) Volatility Clustering:** What happens to the memory structure in case of volatility clustering periods of price and volume? A better understanding of memory in investor attention shall have important implications to value at risk computation, volatility modeling, analyzing market efficiency, risk diversification and policies in energy market.

2 Memory Detection in Time Series

In this section we briefly discuss the notion of 'memory' in a time series and statistical methods for its detection. Let X_t be a sequence of IID random variables such that $E(X_t^2) < \infty$ and $var(X_t)$ is independent of t . Let $\lambda_u = Cov(X_t, X_{t+u})$. The time series is said to have [18] **no memory** if $\lambda_u = 0$ for all $u \neq 0$. It has a **short memory** if λ_u decays faster or has an exponential decay. In a less stringent sense, X_t has a short memory if $\sum_{u=-\infty}^{u=\infty} |\lambda_u| < \infty$. A **long Memory** exists if λ_u decays slowly or has a power law decay. Again using the mild definition, X_t has a long memory if $\sum_{u=-\infty}^{u=\infty} |\lambda_u| = \infty$.

We analyze the memory structure in the time series using detrended fluctuation and cross correlation analysis. To outline the algorithmic steps involved in these methods, let $\{x_t\}$ and $\{y_t\}$ be two time series with $t = 1, 2, 3, \dots, N$. We denote $m_x = \frac{1}{N} \sum_{i=1}^N (x_i)$ and $m_y = \frac{1}{N} \sum_{i=1}^N (y_i)$. A cumulative sum function (called ‘profile’) for the given time series $\{x_t\}$ and $\{y_t\}$ is constructed as: $X_t = \sum_{i=1}^t (x_i - m_x)$, $Y_t = \sum_{i=1}^t (y_i - m_y)$. To perform **detrended cross correlation analysis**, a fluctuation function is obtained using these steps: [19]: a) Partition X_t and Y_t into $[T_b = \frac{N}{l}]$ non overlapping segments of size l from beginning to end of X_t and Y_t . If the series N is not divisible by l , some points at the end of the series may be left out. Hence another portion of size l is done $[T_e = \frac{N}{l}]$ from end to beginning on both series. b) Enumerate the partitions as $i = 1, 2, 3, \dots, 2T = (T_b + T_e)$. For each partition i in $1 < i < 2T$, a least square line is fitted (denoted by $X_{i,t}^{ols}$ and $Y_{i,t}^{ols}$). The detrended covariance is computed as -

$$\psi_i^2(l) = \frac{1}{l} \sum_{j=1}^l ([X_{(i-1)l+j} - X_{i,j}^{ols}] [Y_{(j-1)l+j} - Y_{i,j}^{ols}])$$

for $i = 1, 2, 3, \dots, T_b$

$$\psi_i^2(l) = \frac{1}{l} \sum_{j=1}^l ([X_{N-(i-T_e)l+j} - X_{i,j}^{ols}] [Y_{N-(i-T_e)l+j} - Y_{i,j}^{ols}])$$

for $i = T_b + 1, T_b + 2, \dots, 2T$. The detrended cross correlation analysis (DCCA) fluctuation function is given by $\psi_{DCCA}^2(l) = \left\{ \frac{1}{2T} \sum_{i=1}^{2T} \psi_i^2(l) \right\}^{\frac{1}{2}}$. If only one of the time series is considered, the detrended covariance reduces to detrended variance. The **detrended fluctuation analysis** (DFA) function is given: $\psi_{DFA}^2(l) = \left\{ \frac{1}{2T} \sum_{i=1}^{2T} \psi_i^2(l) \right\}^{\frac{1}{2}}$. However this method was developed earlier by Peng et. al [20]. The main essence of detrended fluctuation function is the fact it follows power law [21]: $\psi_{DFA}^2(l) \propto l^\alpha$. If the individual series x_t and y_t are power law correlated then $\psi_{DCCA}^2(l) \propto l^\beta$ [19]. Using DFA and DCCA exponents, detrended cross correlation coefficient (ρ_{DCCA}) for series x_t and y_t is computed as -

$$\rho_{DCCA}(l) = \frac{\psi_{DCCA}^2(l)}{[\psi_{DFA}^2(l)]_{\{x_i\}} [\psi_{DFA}^2(l)]_{\{y_i\}}} \quad (1)$$

The idea of both DFA and DCCA has its root in a method known as ‘Hurst Rescaled Analysis’ [22]. The associated exponent (known as Hurst exponent, H) could be affected by non-stationaries [4], while DFA and DCCA exponents (α and β) works well on non stationary series as well. In our analysis we only use DFA and DCCA to analyze the memory structure of the time series. The interpretations of H and α are similar [23]. For a stationary process the value of H lies between 0 and 1. For $H = 0.5$, the series is just a white noise (random walk) and has no memory. For $H < 0.5$ the series is anti persistent while $H > 0.5$

shows persistent behavior of the series. The series becomes non-stationary when H crosses 1, however till $H=1.5$, it exhibits a mean reverting behavior. $H = 1.5$ reflects that the series is Brown Noise (Brownian motion). When H exceeds 1.5 it represents an explosive process. β is a measure of nature of cross correlation between two series. For a given value of β the primary nature of series remains same but for the interpretation. For example - if $0 < \beta < 0.5$, there is anti persistent cross correlation i.e. increase in one series is marked by a decrease in another. $\rho_{DCCA} = -1, 0, 1$, imply that there complete negative, zero, positive cross correlation respectively between the two series.

3 Data & Investor Attention Measure

For our analysis we chose 20 largest energy companies by market capitalization [24] listed at New York Stock Exchange. We quantify investor attention using Google search queries for these particular stock. The key idea is that if an investor is searching for a query in Google, this means (s)he is paying attention to it. So Google search could be a revealed measure of attention [25]. Zhi Da et. al showed that the investor attention as measured by relative search volume of stock ticker symbols correlates with existing measures of investor attention [25]. Their results also suggest that search based investor attention is more real time. Amal Aouadi et al. [26] analyzed France stock market and showed that Google search based investor attention is correlated with trading volume of the stock and could be used to model volatility.

To quantify investor attention we use relative search volume time series of ‘stock name’ instead of ‘stock ticker symbol’ because the later is likely to capture more of retail investor attention [25]. Further, in our analysis a ‘company name’ as search query is much more relevant than ticker symbol. For example - stocks like CNOOC Limited has its symbol as ‘CEO’, so looking at search query time series of ‘CEO’ gives little information about investor attention to the stock. In fact while looking for time series of a particular stock name, Google gives suggestions whether the entered stock is just a search term or a corporation. We select the time series of the stock name corresponding to corporation. The obtained series is the relative search volume of the stock with respect to the total search volume worldwide over time scaled from 0 to 100. Let R_t be the relative search volume of stock at time t . We define the measure of investor attention (\mathbb{I}_t) as $\log(1 + R_t)$.

We collect the data for R_t for each of the stock using public web facility of Google called “Google Trends”. For a given stock, we collect dataset for R_t classified into following categories - **a) Longer time duration, searched locally:** In this case R_t consists of weekly data from second week of April 2012 to last week of April 2017 where search location is restricted to the country of origin for the stock, **b) Longer time duration, searched globally:** In this case R_t is same as above but the search location is worldwide now. **c) Shorter time duration, searched locally:** In this case R_t consists of daily data from 26th

Jan 2017 to 24th April 2017 where search location is restricted to the country of origin for the stock. **d) Shorter time duration, searched globally:** Again, in this case R_t is same as above except the search location is now modified to worldwide now. The key idea behind this classification is to understand the memory dependence structure when investor attention data is sampled at a low & high frequency as well as to see the cross correlations between local and global investor attention for a given stock.

4 Analysis & Results

4.1 Memory in Investor Attention: Existence

To check the existence of memory in investor attention, we perform detrended fluctuation analysis of the time series for both long and short duration. Based on the estimated coefficients we conclude on the existence of long range memory in the series. For each stock we obtain the investor attention by $\mathbb{I}_t = \log(1 + R_t)$. We denote $\mathbb{I}_{t,d}$ and $\mathbb{I}_{t,w}$ as investor attention of stock with underlying time series frequency as daily and weekly respectively. The length of time series of $\mathbb{I}_{t,d}$ is 89 while for $\mathbb{I}_{t,w}$ it is 261. In this case, we limit our analysis to investor attention obtained using global search for the stocks (since local search volume could be zero if the language of query entered is non - English). For example, we observe that for Sinopec (a Chinese firm) relative volume of local search query is often zero while globally it has non zero and significant large search volumes (Figure 1). This means investors in China use Chinese search queries or a different search engine (like Baidu).

For any given stock we first compute $\mathbb{I}_{t,d}$ and $\mathbb{I}_{t,w}$. Using the steps outlined in section 2, we obtain the fluctuations (ψ_{DFA}) as a function of window size (l). Using equation (3), we assume a constant k_i for a stock i such that - $\psi_{DFA}^2(l) = k_i l^{\alpha_i}$. Therefore,

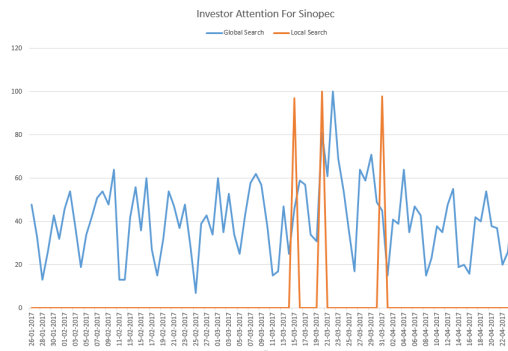


Fig. 1. Local and Global Search Trends For Query “Sinopec”

$$\log(\psi_{DFA}^2(l)) = \log(k_i) + \alpha_i * \log(l) \quad (2)$$

To obtain α_i for a given stock, we fit a linear model using ordinary least squares between $\log(\psi_{DFA}^2(l))$ as dependent variable and $\log(l)$ as independent variable. Since, we are interested in whether the $\mathbb{I}_{t,d}$ or $\mathbb{I}_{t,w}$ has memory or its just a noise, we do an hypothesis testing to check whether obtained α_i is statistically different from 0.5 (case when it is a pure noise). For this we use a null hypothesis $H_0 : \alpha_i = 0.5$ and alternate hypothesis as $H_A : \alpha_i \neq 0.5$. Wald statistic for this test is defined as: $W = \left(\frac{\alpha_i - 0.5}{\sigma_{\alpha_i}} \right)^2$, where σ_{α_i} is the standard error of α_i . Figure 2 shows the plot of logarithm of fluctuation function vs logarithm of window size for the stock Sinopec. To obtain the fluctuation function we vary window size from 5 to 85 for $\mathbb{I}_{t,d}$ and 5 to 250 for $\mathbb{I}_{t,w}$. It is evident (from Table 1 & Table 2) that for both the series the null hypothesis of pure noise is rejected. We observe that the DFA exponent for $\mathbb{I}_{t,d}$ is 0.34, suggesting that the series is anti-persistent i.e. it exhibits a mean reverting behavior. However $\mathbb{I}_{t,w}$ has DFA exponent as 0.90 (Table 2) indicating a long term dependence in the investor attention and is near to the edge of non-stationarity. We carry out the same analysis for global investor attention of all the stocks. It is evident that more often than not Wald test rejects null of pure noise in investor attention. This confirms the existence of power law correlated structure implying a long term memory in the time series of investor attention.

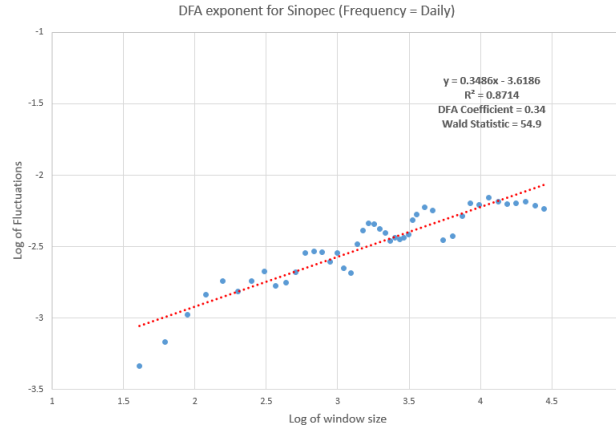


Fig. 2. DFA coefficient estimation for investor attention (for 3 months) for “Sinopec”

4.2 Memory in Investor Attention: Dynamics

To explore the dynamics of memory across time we carry out a rolling window analysis for both 90 day & and 5 years series of investor attention based on global

Stock Name	DFA Exponent	$R_{Squared}$	Sd_{Error}	Wald Statistic	P Value
Exxonmobil	0.4359	0.9291	0.0184	12.1934	0.0011
Royal Dutch	0.3395	0.9244	0.0148	117.5224	0.0000
Chevron	0.3452	0.9192	0.0156	98.3467	0.0000
Petrochina	0.2864	0.7039	0.0283	56.8750	0.0000
Total SA	0.6054	0.8817	0.0338	9.7144	0.0033
Schlumberger	0.2335	0.8325	0.0160	278.5484	0.0000
British Petroleum	0.3486	0.8714	0.0204	54.9230	0.0000
Sinopec	0.4825	0.8429	0.0318	0.3045	0.5839
Petrobras	0.4473	0.8945	0.0234	5.0672	0.0295
Conco Phillips	0.3716	0.5840	0.0478	7.2129	0.0102
ENI	0.1779	0.5513	0.0245	173.1554	0.0000
Enterprise Products	0.4601	0.8230	0.0325	1.5019	0.2270
Statoil	0.8162	0.8936	0.0429	54.2219	0.0000
EOG Resources	0.3235	0.8580	0.0201	77.3600	0.0000
CNOOC Limited	0.5332	0.8713	0.0313	1.1292	0.2939
Suncor Energy	0.3534	0.9106	0.0169	75.4474	0.0000
Kinder Morgan	0.4800	0.7820	0.0386	0.2672	0.6079
Occidental Petroleum	0.6091	0.9398	0.0235	21.5404	0.0000
Halliburton	0.2469	0.8833	0.0137	341.8434	0.0000
Phillips 66	0.7697	0.9690	0.0210	164.8613	0.0000

Table 1. DFA Exponents For Investor Attention (90 day Period)

Stock Name	DFA Exponent	$R_{Squared}$	Sd_{Error}	Wald Statistic	P Value
Exxonmobil	0.8085	0.9543	0.0179	298.0700	0.0000
Royal Dutch	0.9214	0.9303	0.0255	273.3864	0.0000
Chevron	1.1181	0.9745	0.0183	1146.7347	0.0000
Petrochina	0.9053	0.9155	0.0278	212.9451	0.0000
Total SA	0.8237	0.9669	0.0154	441.5677	0.0000
Schlumberger	0.9949	0.8210	0.0469	111.2002	0.0000
British Petroleum	0.9665	0.9277	0.0272	293.0681	0.0000
Sinopec	0.9011	0.9308	0.0248	261.3366	0.0000
Petrobras	1.1102	0.9628	0.0220	766.0020	0.0000
Conco Phillips	0.9839	0.9563	0.0212	518.9868	0.0000
ENI	0.6126	0.8886	0.0219	26.4150	0.0000
Enterprise Products	0.6503	0.8818	0.0241	39.0325	0.0000
Statoil	1.0556	0.9380	0.0274	410.5673	0.0000
EOG Resources	0.7491	0.9179	0.0226	121.0644	0.0000
CNOOC Limited	0.6927	0.9512	0.0158	147.8458	0.0000
Suncor Energy	0.6855	0.8659	0.0272	46.3254	0.0000
Kinder Morgan	0.8919	0.9472	0.0213	339.4692	0.0000
Occidental Petroleum	0.8680	0.9266	0.0247	222.3952	0.0000
Halliburton	1.1109	0.8871	0.0400	232.9367	0.0000
Phillips 66	0.6761	0.8182	0.0322	29.9242	0.0000

Table 2. DFA Exponents For Investor Attention (5 years Period)

search volume for the stock. For a 90 day period the rolling window consists of 22 days and for a 5 year period, 24 quarters are taken. For each stock we take 65 and 217 rolling windows for 90 days and 5 year period respectively. We compute DFA coefficients for each rolling window using the method discussed in section 2. The dynamics of power law dependence is shown in Figure 3 for a subset of stocks. The dynamics of this dependence structure is observed to be persistent and short lived in nature. This means as we progress across rolling windows for a given stock, a large change is followed by a large change and small change is followed by a small change. From the plot, it is clear that direction of changes DFA exponent varies rapidly thereby changing the extent of dependence quickly. This characteristic brings down the predictability of investor attention which could lead to higher efficiency in the market. Similar pattern is observed for both 90 days and a 5 year period.

4.3 Sampling Frequency & Dependence Structure

We have considered the investor attention at two different frequencies. As mentioned earlier, for a short term investor attention we consider 90 days data measured daily and for a long term investor attention we consider 5 years data measured weekly.

We delved deeper into obtained DFA exponents to spot any differences in pattern or values for $\mathbb{I}_{t,d}$ & $\mathbb{I}_{t,w}$. We partitioned the estimated DFA exponents into four intervals - a) **(0 - 0.4)**: Anti-persistent, b) **(0.4 - 0.6)**: Almost Pure Noise, c) **(0.6 - 1.0)**: Persistent & d) **(1 - 1.5)**: Non Stationary. DFA exponent for each rolling window for a given stock falls exactly in one of the partitions. For both $\mathbb{I}_{t,d}$ & $\mathbb{I}_{t,w}$, we compute the probability of a rolling window falling into one of these partitions using the relative frequency approach. From the computed probabilities we observe that for a 5 year period $Prob(Persistent)$ is consistently higher than $Prob(Antipersistent)$ for all stocks. This suggests that at low frequency (i.e. weekly), the investor attention has a long range dependence with near non stationary structure making the predictability difficult and thereby boosting market efficiency. However at a high frequency (i.e. daily), $Prob(Antipersistent)$ is relatively higher than $Prob(Persistent)$ for almost all the stocks. This means for most of the rolling windows the series is stationary and mean reverting indicating higher predictability and lesser efficiency in the market. The results remain same when estimated DFA exponents are compared for full time period [Figure 4]. For nearly all stocks low frequency investor attention is closer to 1 while it is less than 0.5 for high frequency investor attention.

4.4 Local and Global Investor Attention

To investigate the cross correlation structure between local and global investor attention we compute $\rho_{DCCA}(l)$ as defined by equation 4. We perform this analysis on $\mathbb{I}_{t,w}$ for a five year period. Local investor attention is the time series based on search queries for the stock at country of origin as the geographical location

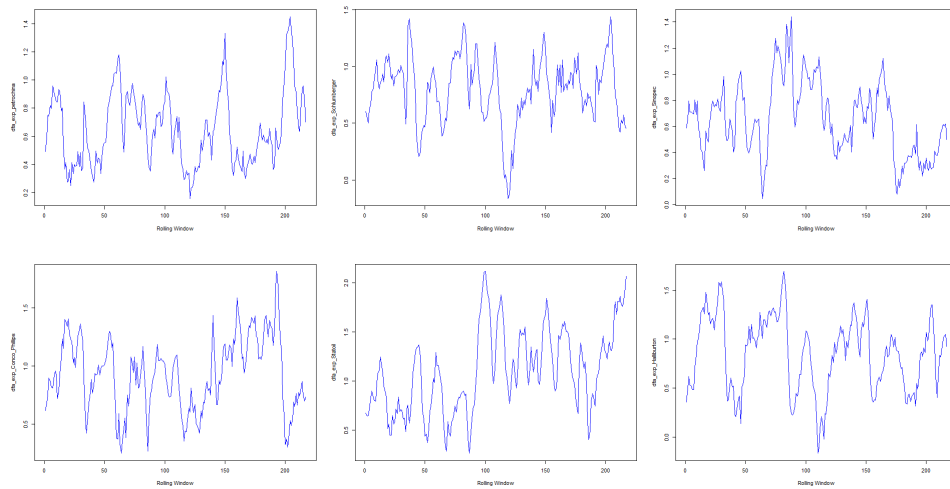


Fig. 3. Memory Dynamics For Investor Attention For Stocks (5 year period) (Left to Right, Top : Petrochina, Schlumberger, Sinopec & Bottom: Conoco Phillips, Statoil, Halliburton)

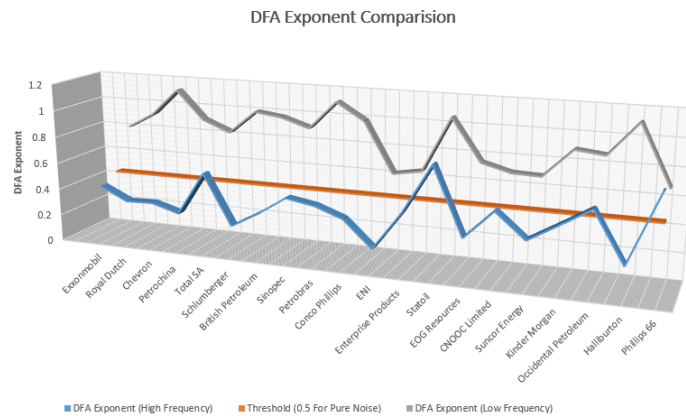


Fig. 4. Dependence Structure At High and Low Frequency

while for global investor attention the location is chosen to be worldwide. An important point to note here is that $\rho_{DCCA}(l)$ is calculated at a given scale. We have a total of 261 observations and we chose $l = 20$. One may calculate $\rho_{DCCA}(l)$ at different scales and then average it out. This value will only be slightly different. We expect the cross correlations to be positive and should be reasonably high. This is because an important news related to the stock draws local and global investors attention. However depending on the stock and its

Stock Name	P(Antipersistent) (90 days)	P(Antipersistent) (5 years)	P (Persistent) (90 days)	P (Persistent) (5 years)	ρ_{DCCA}
Exxonmobil	0.4308	0.1198	0.0923	0.3364	0.86
Royal Dutch	0.3231	0.0507	0.2154	0.4194	0.36
Chevron	0.5385	0.0138	0.0615	0.4378	0.92
Petrochina	0.6462	0.2028	0.0462	0.4286	NA
Total SA	0.4154	0.0507	0.2000	0.5346	0.79
Schlumberger	0.7385	0.0783	0.0000	0.5161	0.35
British Petroleum	0.3231	0.2212	0.2308	0.4700	0.54
Sinopec	0.3846	0.2074	0.2615	0.4424	NA
Petrobras	0.5077	0.0092	0.2000	0.3088	0.99
Conco Phillips	0.6462	0.0553	0.1692	0.4009	0.86
ENI	0.9077	0.0507	0.0000	0.4562	0.94
Enterprise Products	0.7077	0.1429	0.0923	0.3456	0.93
Statoil	0.3538	0.0230	0.2769	0.3318	0.46
EOG Resources	0.5231	0.1751	0.0462	0.2719	0.98
CNOOC Limited	0.6000	0.3502	0.0923	0.3226	NA
Suncor Energy	0.4154	0.0876	0.0769	0.5392	0.94
Kinder Morgan	0.5231	0.1060	0.1077	0.4147	0.81
Occidental Petroleum	0.3231	0.0691	0.2154	0.5023	0.80
Halliburton	0.7846	0.1429	0.0000	0.3180	0.94
Phillips 66	0.3231	0.0599	0.2462	0.3364	0.78

Table 3. DFA Exponents & ρ_{DCCA} For Investor Attention

importance of information related to stock, the intensity of attention may vary. In Table 3 (last column), we enlist all the cross correlation values. As expected the correlations are positive, some of them are high and most of them are above 0.5. For a few stock correlations cannot be computed because search volume is very small (due to non English search queries). In our case, all three happens to be Chinese stocks indicating the investors in China uses queries in ‘Chinese’ to collect stock information.

4.5 Volatility Clustering and Investor Attention

From the estimated DFA exponents and rolling window analysis we have seen that the long range memory has persistent and short lived nature. We also observed that the extent of dependence is changing rapidly across rolling windows. In this part we analyze if the dependence structure changes during returns or volume volatility clustering periods. Given the dynamics of memory of investor attention, the long range dependence should not have much variation under such periods and intrinsic memory structure should be retained. However, it is very much possible investor attention can affect the returns or volume volatility (as discussed by Daniel Andrei and Michael Hasler [27]).

For a given window, we measure returns volatility by taking standard deviation of log returns and log volumes. We observe that memory structure is retained during volatility clustering periods. To validate this proposition we check corre-

lations between volatility between DFA exponents and volatility for all stocks. The results suggest that there is a small negative correlation (~ 0.2) between the two for most of the stocks. To confirm this further we carry out Granger causality tests with lag 3 and check for both ways causality. The null hypothesis that volatility doesn't Granger cause dependence structure (or vice versa) is failed to get rejected in almost of all the cases. Hence, the results are in favor of the proposition that volatility clustering has little effect on long dependence of investor attention.

5 Conclusions

In this article we investigated the long range dependence of investor attention for top 20 stocks from energy market. Google search queries are revealed measure of attention and we used the relative search query volume to quantify the investor attention. Our results suggest that investor attention is indeed power law correlated and has long term dependence in its time series at both high and low frequencies. Further we observed that at high frequencies, investor attention is stationary and anti-persistent indicating a higher predictability. Dynamics of long range investor attention indicates that extent of dependence is changing rapidly and is short lived and persisting in nature. Detrended cross correlation analysis reveals that there is a reasonably high cross correlations between local and global investor attention. Finally, by using Granger Causality tests we see that the returns and volume volatility clustering has little effect on long range dependence structure of investor attention time series.

References

1. G. Huberman and T. Regev, "Contagious speculation and a cure for cancer: A nonevent that made stock prices soar," *The Journal of Finance*, vol. 56, no. 1, pp. 387–396, 2001.
2. D. Kahneman, "Attention and effort prentice hall englewood cliffs," *NJ Google Scholar*, 1973.
3. W. Willinger, M. S. Taqqu, and V. Teverovsky, "Stock market prices and long-range dependence," *Finance and stochastics*, vol. 3, no. 1, pp. 1–13, 1999.
4. A. W. Lo, "Long-term memory in stock market prices," tech. rep., National Bureau of Economic Research, 1989.
5. Z. Ding, C. W. Granger, and R. F. Engle, "A long memory property of stock market returns and a new model," *Journal of empirical finance*, vol. 1, no. 1, pp. 83–106, 1993.
6. Y.-W. Cheung and K. S. Lai, "A search for long memory in international stock market returns," *Journal of International Money and Finance*, vol. 14, no. 4, pp. 597–615, 1995.
7. I. N. Lobato and C. Velasco, "Long memory in stock-market trading volume," *Journal of Business & Economic Statistics*, vol. 18, no. 4, pp. 410–427, 2000.
8. B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley, "Cross-correlations between volume change and price change," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22079–22084, 2009.

9. F. J. Breidt, N. Crato, and P. De Lima, "The detection and estimation of long memory in stochastic volatility," *Journal of econometrics*, vol. 83, no. 1, pp. 325–348, 1998.
10. T. Bollerslev and H. O. Mikkelsen, "Modeling and pricing long memory in stock market volatility," *Journal of econometrics*, vol. 73, no. 1, pp. 151–184, 1996.
11. N. Crato and P. J. de Lima, "Long-range dependence in the conditional variance of stock returns," *Economics Letters*, vol. 45, no. 3, pp. 281–285, 1994.
12. S. R. Souza, B. M. Tabak, and D. O. Cajueiro, "Long-range dependence in exchange rates: the case of the european monetary system," *International Journal of Theoretical and Applied Finance*, vol. 11, no. 02, pp. 199–223, 2008.
13. G. Samorodnitsky *et al.*, "Long range dependence," *Foundations and Trends® in Stochastic Systems*, vol. 1, no. 3, pp. 163–257, 2007.
14. N. Vlastakis and R. N. Markellos, "Information demand and stock market volatility," *Journal of Banking & Finance*, vol. 36, no. 6, pp. 1808–1821, 2012.
15. N. Vozlyublennaya, "Investor attention, index performance, and return predictability," *Journal of Banking & Finance*, vol. 41, pp. 17–35, 2014.
16. J. Li and J. Yu, "Investor attention, psychological anchors, and stock return predictability," *Journal of Financial Economics*, vol. 104, no. 2, pp. 401–419, 2012.
17. X. Fan, Y. Yuan, X. Zhuang, and X. Jin, "Long memory of abnormal investor attention and the cross-correlations between abnormal investor attention and trading volume, volatility respectively," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 323–333, 2017.
18. P. M. Robinson, "Modeling memory of economic and financial time series," *The Singapore Economic Review*, vol. 50, no. 01, pp. 1–8, 2005.
19. B. Podobnik and H. E. Stanley, "Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series," *Physical review letters*, vol. 100, no. 8, p. 084102, 2008.
20. C.-K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, M. Simons, and H. Stanley, "Finite-size effects on long-range correlations: Implications for analyzing dna sequences," *Physical Review E*, vol. 47, no. 5, p. 3730, 1993.
21. C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of dna nucleotides," *Physical review e*, vol. 49, no. 2, p. 1685, 1994.
22. H. E. Hurst, R. P. Black, and Y. Simaika, *Long-term storage: an experimental study*. Constable, 1965.
23. L. Kristoufek, "Power-law correlations in finance-related google searches, and their cross-correlations with volatility and traded volume: Evidence from the dow jones industrial components," *Physica A: Statistical Mechanics and its Applications*, vol. 428, pp. 194–205, 2015.
24. R. Rapier, "The 25 biggest oil and gas companies in the world," *Financial Times*, March 2016.
25. Z. Da, J. Engelberg, and P. Gao, "In search of attention," *The Journal of Finance*, vol. 66, no. 5, pp. 1461–1499, 2011.
26. A. Aouadi, M. Aroui, and F. Teulon, "Investor attention and stock market activity: Evidence from france," *Economic Modelling*, vol. 35, pp. 674–681, 2013.
27. D. Andrei and M. Hasler, "Investor attention and stock market volatility," *Review of Financial Studies*, p. hhu059, 2014.

Sparse Granger-Causal Network Learning via the Depth Wise Group LASSO

An Application of ADMM for Large Vector Autoregressions

R. J. Kinnear and R. R. Mazumdar

Department of Electrical and Computer Engineering, University of Waterloo, Canada
Ryan@Kinnear.ca

Abstract. In pursuit of learning Granger-causality networks from time series data we derive, via application of the Alternating Direction Method of Multipliers (ADMM), a method to fit large and multi-lag vector autoregressive models with a “depth-wise” grouped sparsity pattern. Our grouped sparsity algorithm is an extension of the widely applied LASSO (directly applicable to VAR(1) models) to VAR(p) models with a structure specifically tailored to causality networks. We apply our so called DWGLASSO algorithm on a large system consisting of temperature data from 165 Canadian weather stations in order to provide some empirical validation.

Keywords: Granger causality, time series, sparsity, LASSO, vector autoregression, proximal methods, ADMM.

1 Introduction and Related Work

Since the work of Granger[1] in the latter half of the 20th century there has been considerable interest in the discovery of causal relations in time series data. The idea behind Granger’s definition of causality for stochastic processes, “Granger-causality”, is that if the past of process j provides information about the future of process i , that is not available anywhere else, then there must be more than a merely correlative connection between processes i and j . This definition induces a natural graph structure among the processes, generally referred to as a causality graph or causality network, and it is the modeling and analysis of this graph that is the primary goal of time series analysis in the context of Granger-causality.

In purely autoregressive models, Granger-causality has a particularly simple structure and statistical methods for testing Granger-causality in these models were established by Geweke [2] [3]. Along similar lines, [4] [5] have extended the methods of Geweke to state space models. The crux of these approaches is in asymptotic analysis of the statistics of classical estimation procedures for time series models.

Aside from asymptotic analysis, many researchers have pursued methods for estimating or learning causality graphs through convex optimization. For example, [6] provides a simulation study of a number of different estimation techniques (including sparsity inducing regularization), and [7] provides an approach to VAR(1) models that enforces stability. Inspired by the success of LASSO type estimators in regression problems, we follow a similar line of work here. One of the core requirements underlying the theory of sparsity inducing regularizers is that the underlying system is in fact sparse, hence we stress that the techniques presented here are useful primarily for large systems having a small number of causal interactions.

In this paper we derive a convex optimization algorithm for fitting large VAR(p) models with a grouped sparsity pattern along the edges of the causality graph. The algorithm is derived in section (3.2), but we begin with some preliminaries in section (2) and a review of the ADMM algorithm in section (3).

Applications of Granger Causality have been explored in diverse areas including medical imaging [8], neuroscience [9], finance [10], and others. In section (4) of this paper we demonstrate application of our technique on Canadian temperature data obtained from the Canadian weather energy and engineering data sets (CWEEDS) [11]. Our concluding remarks are given in section (5).

2 Notation and Preliminaries

2.1 Granger Causality

Let $x(t) = (x_1(t), \dots, x_n(t))$ be a column vector of real valued discrete time ($t \in \mathbb{Z}$) wide sense stationary (WSS) stochastic processes with bounded second moments, that is $x_i(t) \in L_2(\Omega, \mathcal{F}, \mathbf{P})$, the Hilbert space of square integrable random variables and $E[x_i(t)x_j(s)] = E[x_i(|t-s|)x_j(0)]$. Let

$$\mathbb{H}_t = \text{cl} \left\{ \sum_{\tau=1}^{\infty} \sum_{i=1}^n b_i^{(\tau)} x(t-\tau) \mid b_i^{(\tau)} \in \mathbb{R} \right\} \quad (1)$$

denote the Hilbert space of random variables generated by the (strict) past of $x(t)$ and

$$\mathbb{H}_t^{-j} = \text{cl} \left\{ \sum_{\tau=1}^{\infty} \sum_{i \neq j} b_i^{(\tau)} x(t-\tau) \mid b_i^{(\tau)} \in \mathbb{R} \right\} \quad (2)$$

the space generated by all but component j .

The notation $\hat{E}[x_i(t) \mid \mathbb{H}_t]$ will be used to denote the *unique* projection of $x_i(t)$ onto the Hilbert space \mathbb{H}_t , which is the causal linear minimum mean square error (LMMSE), or Wiener, estimate of $x_i(t)$ given the strict past of $x(t)$. That is,

$$\hat{E}[x_i(t) \mid \mathbb{H}_t] = \underset{z \in \mathbb{H}_t}{\text{argmin}} E[|x_i(t) - z|^2] . \quad (3)$$

And, the expected squared error of the estimate:

$$\hat{\xi}[x_i(t) | \mathbb{H}_t] = E[(\hat{E}[x_i(t) | \mathbb{H}_t^V] - x_i(t))^2] . \quad (4)$$

Note that since the processes are wide sense stationary, the aforementioned quantities do not vary with time. The notion of Granger-causality is captured in the following definition.

Definition 1. *If*

$$\hat{\xi}[x_i(t) | \mathbb{H}_t] < \hat{\xi}[x_i(t) | \mathbb{H}_t^{-j}] , \quad (5)$$

then we say that x_j Granger-causes x_i (conditional on x), and write $x_j \longrightarrow x_i$.

In contrast to the covariance between variables $E[x_i(t)x_j(t)]$, Granger-causality is measured with a *strict* time difference. Furthermore, Granger-causality is a joint measure amongst all processes, whereas covariance is only pairwise. And finally, the covariance is an undirected measurement, and Granger-causality emphasizes a direction from one process to another.

Some of the intuition behind this definition is mentioned in section 1, and is greatly expanded upon in [12]. We also point out that this notion of causality has little in common with the notion of causality popularized by Pearl [13].

2.2 Autoregressive Modeling

Recall that $x(t)$ is an n -vector of WSS processes. The Wold decomposition theorem tells us that there is some square-summable (in $\|\cdot\|_F$ norm) sequence of real valued $n \times n$ matrices $A(\tau)$, a white noise sequence $\epsilon(t)$, and a perfectly predictable sequence $u(t)$ such that

$$x(t) = \sum_{\tau=0}^{\infty} A(\tau)\epsilon(t-\tau) + u(t) . \quad (6)$$

This is a moving average representation of $x(t)$, and exists for every WSS L_2 process. In practice, the predictable term $u(t)$ should be removed by detrending, and so we simply take $u(t) = 0$. In order to obtain an autoregressive representation, the LSI filter given by $A(\tau)$ must be invertible, and a sufficient condition for this invertibility is that there is some $c > 0$ such that the spectral density matrix $S_x(\lambda)$ of $x(t)$ satisfies $c^{-1}I \preceq S_x(\lambda) \preceq cI$ for λ almost everywhere in $[-\pi, \pi)$. Given this condition, we have again a square summable sequence $B(\tau)$ such that

$$x(t) = \sum_{\tau=1}^{\infty} B(\tau)x(t-\tau) + e(t) , \quad (7)$$

where $e(t)$ is uncorrelated in time, but not necessarily across its own components. Finally, availability of only finite quantities of data necessitates that we restrict ourselves further by assuming that $x(t)$ is generated by the Markovian vector autoregressive VAR(p) model

$$x(t) = \sum_{\tau=1}^p B(\tau)x(t-\tau) + e(t) . \quad (8)$$

A natural perspective is to view this model as a graph having nodes $x_i(t)$ and edges given by the linear shift-invariant (LSI) filter $\tilde{B}_{ij}(z) = \sum_{\tau=1}^p B_{ij}(\tau)z^{-\tau}$ whose coefficients are arranged into a column vector $\tilde{B}_{ij} = (B_{ij}(1), \dots, B_{ij}(p))$. In this model, Granger-causality has a particularly simple characterization:

Proposition 1. *If $x(t)$ is an n dimensional wide sense stationary L_2 stochastic process generated by the VAR(p) model (8) then $x_j(t)$ Granger-causes $x_i(t)$ if and only if $\tilde{B}_{ij} \neq 0$.*

Proof. The condition for Granger-Causality $x_j \rightarrow x_i$ is given as

$$\hat{\xi}[x_i(t) | \mathbb{H}_t] < \hat{\xi}[x_i(t) | \mathbb{H}_t^{-j}] . \quad (9)$$

Since $e(t)$ is temporally uncorrelated, the Hilbert space projections are given by the model's true parameters, so (9) is equivalent to

$$E|x_i(t) - \sum_{\tau=1}^{\infty} \sum_{k=1}^n B_{ik}^{(\tau)} x_k(\tau)|^2 < E|x_i(t) - \sum_{\tau=1}^{\infty} \sum_{k \neq j} B_{ik}^{(\tau)} x_k(\tau)|^2 .$$

Now, if there were no τ_0 such that $B_{ij}^{(\tau_0)} \neq 0$ then the above strict inequality would in fact be an equality, a contradiction. Conversely, since $B_{ik}^{(\tau)}$ provides the best linear estimate of $x_i(t)$ from $x(t)$, if there is some τ_0 such that $B_{ij}^{(\tau_0)} \neq 0$ then the above strict inequality must hold, otherwise $B_{ij}^{(\tau_0)} = 0$ would provide an equivalent or superior prediction, contradicting either the uniqueness of projections in Hilbert space, or the optimality of the projection. \square

2.3 Estimating VAR Model Coefficients

Given a finite sample of $T+p$ data points: $x(-p+1), x(-p+2), \dots, x(T)$, there are a wide variety of methods available to produce an estimate $\hat{B}(\tau)$ of the coefficients $B(\tau)$ in the model (8). Classical methods revolve around solving the Yule-Walker equations with finite data estimates of covariance sequences, and indeed, this is the approach put forth by Geweke in [3]. A similar approach is taken in [14]. Another is the simple ordinary least squares estimate

$$\underset{B(\tau)}{\text{minimize}} \quad \frac{1}{2T} \sum_{t=1}^T \|x(t) - \sum_{\tau=1}^p B(\tau)x(t-\tau)\|_2^2 , \quad (10)$$

which is our starting point in this paper. This can be viewed as either a maximum likelihood estimate in the case for which $e(t)$ is Gaussian, or as an asymptotically valid estimate of the LMMSE estimator.

When data is abundant for each component of $x(t)$ (e.g. when $T \gg n^2p$), either of the aforementioned methods are perfectly adequate. However, many applications do not satisfy this requirement. Indeed, the underlying graphical structure induced by $B(\tau)$ only becomes interesting when n is of at least modest size. In this case, the variance of traditional or OLS estimates of $B(\tau)$ is so large as to render the estimates entirely useless.

Standard methods to deal with this issue is to accept some bias in the estimation process and add regularizing terms. By appropriately arranging coefficients, we can consider the following problem:

$$\underset{B}{\text{minimize}} \quad \frac{1}{2T} \|Y - ZB\|_F^2 + \lambda [\alpha \|B\|_F^2 + (1 - \alpha) \Gamma(B)] \quad , \quad (11)$$

where $Y = [x(T) \dots x(1)]^\top$ is $(T \times n)$ formed directly from the vectors $x(t)$, $Z = [z(T-1) \dots z(0)]^\top$ is $(T \times np)$ where the rows $z(t)$ are formed from stacking $x(t), \dots, x(t-p+1)$, and $B = [B(1) B(2) \dots B(p)]^\top$ is the $(np \times n)$ coefficient matrix. The term $\lambda \geq 0$ is a tuning parameter for the amount of regularization, and $\alpha \in [0, 1]$ trades off between the regularizer Γ and $\|\cdot\|_F^2$.

Different choices of Γ in the problem (11) lead to different estimates of $B(\tau)$ and hence allow for a great deal of flexibility in the modeling process. One obvious drawback in this approach is that the resulting estimates are not guaranteed to yield a stable system. This is a big problem if the model is to be used for forecasting, but when we are interested only in the underlying graphical structure induced by \hat{B} , it is not of any great consequence whether the resulting system is stable or not. That being said, Granger-causal analysis can be applied as a model selection procedure; the estimated \hat{B} matrix need not be the ultimate result of the modeling process.

3 Structured Grouping for Causal Inference

3.1 DWGLASSO

Common regularizers in the context of regression are the squared ℓ_2 Frobenius norm ($\alpha = 1$), referred to as Tikhonov regularization, or a simple ℓ_1 norm $\Gamma_1(B) = \|B\|_1 \triangleq \sum_{ij} |B_{ij}|$ (with $\alpha = 0$), which is the well known sparsity inducing ‘‘LASSO’’ regularizer [15].

The LASSO regularizer, which results in an unstructured sparsity pattern in the B matrix, can be extended to the grouped LASSO (GLASSO) in which we take a sum of unsquared Euclidean norms $\Gamma_G(B) = \sum_{g \in G} \|B_{[g]}\|_2$ on groups in the B matrix, where $B_{[g]}$ denotes a vector of B coefficients in group $g \subseteq \{1, 2, \dots, n\}$. It was shown by Yuan et al. [16] that this leads to a sparsity pattern in which each of the coefficients in $B_{[g]}$ are jointly zero or non-zero.

Inspired by the characterization of proposition 1, the proposal of this paper, in a vein similar to [6] and [17] is to use

$$\Gamma(B) = \sum_{i=1}^n \sum_{j=1}^n \|\tilde{B}_{ij}\|_2 \quad , \quad (12)$$

which forms groups along each edge of the underlying causality graph. The matrix B is formed from stacking (the transposes of) the lagged coefficient matrices as $B = [B(1) B(2) \dots B(p)]^T$, but it is also natural to imagine stacking into or out of the page (“depth wise”) the matrices $B(\tau)$ to form an $(n \times n \times p)$ array, analogous to the adjacency matrix of the underlying graph, so that looking through this array in location ij gives the coefficients $B_{ij} \in \mathbb{R}^p$ of the LSI filter from process j to process i . It is for this reason that we refer to this structured regularizer as the depth-wise group LASSO (DWGLASSO) regularizer. The adjacency matrix G of the causality graph induced by B is given simply by checking these depth wise filters:

$$G_{ij} = \begin{cases} 1 & ; \|\tilde{B}_{ji}\|_2 > 0, i \neq j \\ 0 & ; \text{otherwise} \end{cases} .$$

In the case where $\alpha = 0$ (no Frobenius term), it is known that co-linearity in the data leads to inconsistent estimates in the sense that if x_j and $x_{j'}$ provide similar information about x_i the LASSO estimate will tend to select only one or the other. This is a serious problem when we want to infer a causality graph. Adding in the ℓ_2 norm term with $\alpha \in (0, 1)$ is referred to as the elastic net [18] and eliminates this problem; the estimator will blend together the influences from x_j and $x_{j'}$.

3.2 ADMM for DWGLASSO

In this section we derive an algorithm to solve 11. Our algorithm derives from the the alternating direction method of multipliers (ADMM) [19] [20], a fast and flexible technique well suited to our needs. Given two closed, proper, convex, though not necessarily differentiable functions f and g , the ADMM algorithm minimizes over B the objective $f(B) + g(B)$ and dictates that we perform the following updates (after initialization to 0):

$$\begin{aligned} B_x^{k+1} &\leftarrow \text{prox}_{\mu f}(B_z^k - B_u^k) , \\ B_z^{k+1} &\leftarrow \text{prox}_{\mu g}(B_x^{k+1} + B_u^k) , \\ B_u^{k+1} &\leftarrow B_u^k + B_x^{k+1} - B_z^{k+1} , \end{aligned} \tag{13}$$

where

$$\text{prox}_{\mu \phi}(V) = \underset{X \in \mathbb{R}^{n \times n}}{\text{argmin}} \left(\phi(X) + \frac{1}{2\mu} \|X - V\|_2^2 \right) \triangleq \underset{X \in \mathbb{R}^{m \times k}}{\text{argmin}} P_\phi(X) , \tag{14}$$

is the proximity operator of some function $\phi : \mathbb{R}^{m \times k} \rightarrow \mathbb{R}$. The parameter μ tunes the convergence of the algorithm, but it’s careful selection is not of paramount importance — we note simply that ad-hoc tuning is sufficient, and that μ should be “small”.

For our purposes, we use the functions

$$f(B) = \frac{1}{2T} \|Y - ZB\|_F^2 + \lambda\alpha \|B\|_F^2, \quad (15)$$

$$g(B) = \lambda(1 - \alpha) \sum_{i=1}^n \sum_{j=1}^n \|\tilde{B}_{ij}\|_2, \quad (16)$$

keeping in mind that the notation \tilde{B}_{ij} refers to the depth-wise grouping of the coefficients of B .

The ADMM algorithm guarantees that $\frac{1}{n^2p} \|B_x^k - B_z^k\|_F^2 \rightarrow 0$ as $k \rightarrow \infty$, and that the value of the objective function $f(B_x^k) + g(B_z^k)$ converges towards the minimum achievable value. Since the objective (11) is strongly convex (as long as we require $\lambda > 0$ and $\alpha \in (0, 1]$), this implies that ADMM is guaranteed to find the unique global minimizer of our problem.

Proposition 2 (Proximity Operator of $f(B) = \frac{1}{2T} \|Y - ZB\|_F^2 + \lambda\alpha \|B\|_F^2$).

$$\text{prox}_{\mu f}(V) = \left(\frac{1}{T} Z^\top Z + \frac{1 + 2\mu\lambda\alpha}{\mu} I \right)^{-1} \left(\frac{1}{T} Z^\top Y + \frac{1}{\mu} V \right). \quad (17)$$

Proof. Since this objective is differentiable and unconstrained, we can easily solve (14).

$$\frac{\partial P_f}{\partial B}(B) = \frac{1}{T} (Z^\top Z B - Z^\top Y) + 2\alpha\lambda B + \frac{1}{\mu} (B - V).$$

Applying the first order optimality condition

$$\frac{\partial P_f}{\partial B}(B^*) = 0 \implies B^* = \left(\frac{1}{T} Z^\top Z + \frac{1 + 2\alpha\lambda\mu}{\mu} I \right)^{-1} \left(\frac{1}{T} Z^\top Y + \frac{1}{\mu} V \right),$$

and since the objective is strongly convex, we have obtained the unique global minimizer defining the proximity operator. \square

Proposition 3 (Proximity Operator of $g(B) = \lambda(1 - \alpha) \sum_{i,j} \|\tilde{B}_{ij}\|_2$).

$$\text{prox}_{\mu g}(V) = \left[P(1) P(2) \dots P(p) \right]^\top \in \mathbb{R}^{np \times n}, \quad (18)$$

where

$$P(\tau)_{ij} = \left(1 - \frac{\mu\lambda(1 - \alpha)}{\|\tilde{V}_{ij}\|_2} \right)_+ \tilde{V}(\tau)_{ij}, \quad (19)$$

and $(x)_+ = \max\{0, x\}$.

Recall that \tilde{B}_{ij} denotes the coefficients of the LSI filter from x_j to x_i . The notation $\tilde{V}_{ij}(\tau)$ denotes the τ^{th} component of the analogous arrangement. The operation in (19) is referred to as group soft thresholding.

Proof. The objective function separates along \tilde{V}_{ij} , so we need only establish (19). To this end, let $\phi(x) = \lambda(1 - \alpha)\|x\|_2$ so that $g(B) = \sum_{ij} \phi(\tilde{B}_{ij})$. The Fenchel conjugate $(\mu\phi)^*$ of $\mu\phi$ is the convex indicator function of the Euclidean ball having radius $\mu\lambda(1 - \alpha)$. That is,

$$(\mu\phi)^*(x) = \begin{cases} 0 & ; \|x\|_2 \leq \mu\lambda(1 - \alpha) \\ \infty & ; \text{otherwise} \end{cases}.$$

Thence we obtain,

$$\text{prox}_{(\mu\phi)^*}(\tilde{V}_{ij}) = \begin{cases} \tilde{V}_{ij} & ; \|\tilde{V}_{ij}\|_2 \leq \mu\lambda(1 - \alpha) \\ \frac{\mu\lambda(1 - \alpha)\tilde{V}_{ij}}{\|\tilde{V}_{ij}\|_2} & ; \text{otherwise} \end{cases}, \quad (20)$$

which is simply a projection onto the aforementioned Euclidean ball¹. A fundamental property of the proximity operator is the Moreau decomposition: $\text{prox}_{\mu\phi}(x) = x - \text{prox}_{(\mu\phi)^*}(x)$, application of which yields (19). \square

Additional details for these types of derivations can be found in [20].

3.3 Computational Considerations

There are a few things to note in regards to practical implementation. Firstly, the matrix inverse in (17) should not be carried out literally, an LU (or Cholesky) factorization of $(\frac{1}{T}Z^T Z + \frac{1+2\mu\lambda\alpha}{\mu}I)$ can be cached and used throughout in solving the system of equations. Secondly, the matrices $Z^T Z$ and $Z^T Y$ can be formed from the pairwise covariances of each x_i, x_j pair at the lags from 0 to p , further savings can be had by making use of the block toeplitz structure of $Z^T Z$. Finally, the matrix B_z^k is formed from the soft-thresholding in (19), and hence will be the sparse solution the algorithm should output upon convergence. The time complexity is $O(n^2 p^2)$ per iteration, with $O(n^3 p^3 + n^2 p T)$ at initialization. Storage complexity is also on the order of $O(n^2 p^2)$.

For large np , it may be prudent to add two additional parameters $\sigma \geq 0$ and $\delta \in [0, 1)$ to shrink and regularize the covariance matrix estimate $\widehat{Z^T Z} = (1 - \delta)Z^T Z + \delta Z^T Z + \sigma I$. However, this is not strictly necessary for the algorithm to work.

4 Example Application with Canadian Weather Data

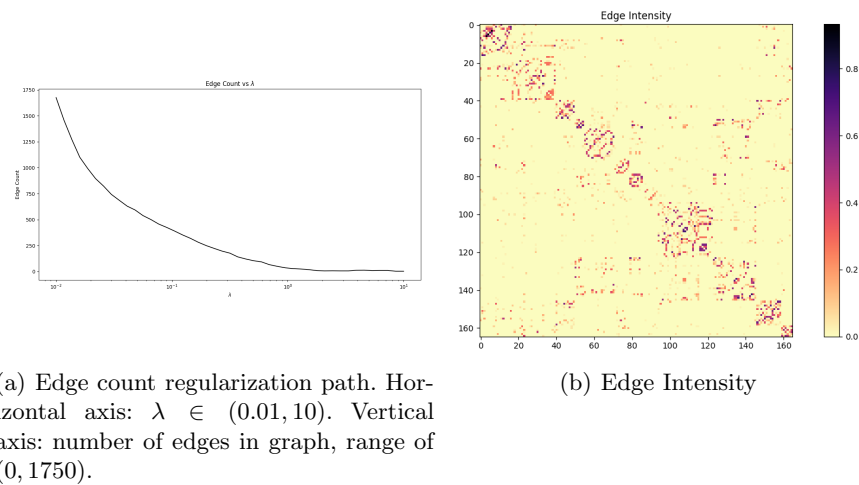
As an example application we have used DWGLASSO to infer a Granger-causality graph from hourly temperature data (from the CWEEDS dataset [11]) between various Canadian locations. We have used $n = 165$ time series of hourly temperature readings of length $T = 1600$ starting on January 1st 1980. We have chosen to use temperature data because geographic considerations can give some intuition

¹ This is typical, the proximity operator for the indicator function of a set is the projection onto that set.

about what the “true” Granger-causality graph should look like, yet the data is still more realistic than using a synthetic dataset. All of the computational tools we have used are a part of Python’s scientific computing stack [21].

After interpolating a small number of missing datapoints, and ensuring that the time stamps of each series are properly aligned with a universal time, we preprocessed all of the data by filtering out the predictable yearly and daily temperature variations (and harmonics thereof) via `scipy.signal.iirnotch`. This preprocessing step is important as each series should be free from perfectly predictable variations $u(t)$ as noted in section 2.2. The parameters $p = 2, \alpha = 0.1, \mu = 0.001$ are held fixed.

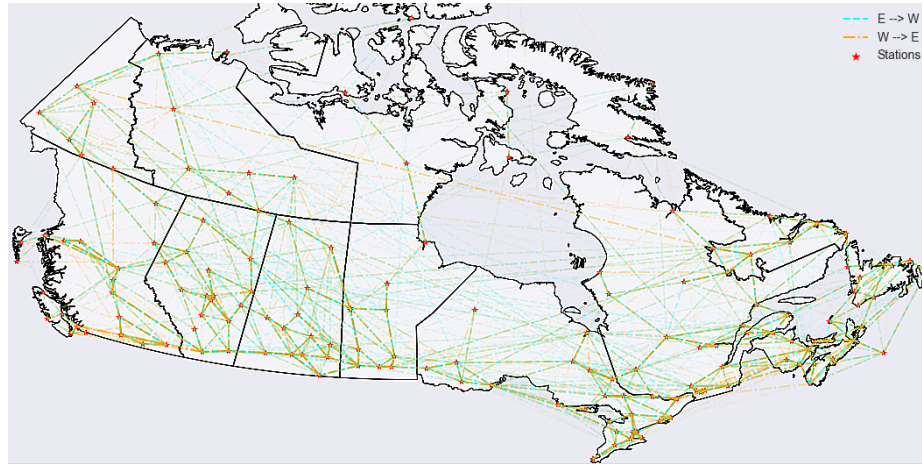
Fig. 1: λ Sweep from 0.01 to 10



The choice of λ is the key parameter for tuning the sparsity of the resulting adjacency matrix G_λ . We have found it to be clear through simulation on synthetic data, as well as with our present application, that choosing λ based on cross validation against mean-squared 1-step ahead prediction error leads to a G_λ matrix which is dramatically denser than is reasonable. As an alternative to choosing a fixed λ we have swept λ through $N = 45$ points on a logarithmic scale in $[10^{-2}, 10]$ and calculated the “edge intensity” of G_λ which we define by the matrix $\frac{1}{N} \sum_\lambda G_\lambda$. The result is shown in figure 2(b). This edge intensity matrix is an attempt to quantify the importance of each edge and provides one possibility for weighting each edge as inferred via DWGLASSO.

Further, we note that as λ increases there is a tendency for G to become sparser, though this regularization path need not be monotonic. This can be seen in the figure (2(a)) where we plot the edge count against λ on a logarithmic axis.

Fig. 2: Inferred causality graph. Direction of each edge from west (left) to east (right) or from east to west is indicated by color and line style. The transparency of each edge is weighted by the edge intensity.



Finally, the inferred causality graph is shown in figure 2 where the transparency of the edges is weighted by the edge intensity. It is clear that there is a great deal of spurious edges with a low intensity (corresponding to small λ) but edges with a high intensity correspond to weather stations in close proximity, as would be expected.

5 Conclusions and Further Outlook

The DWGLASSO algorithm of this paper provides an effective means of exploring the interactions between processes that generate time series data. As apposed to a pairwise testing strategy, our approach considers all n processes jointly, and the nature of LASSO type regularizers means that our algorithm can naturally handle a large number of processes, even when given only short samples of data². For models with long lags (large p) additional within group sparsity would be desirable and can in principle be achieved by further adding a $\|B\|_1$ term to the problem formulation of (11). However, this necessitates evaluating the proximity operator of the norm $\|\cdot\|_1 + \Gamma(\cdot)$, for which there is no closed form. But, as shown by [22], it is still possible to efficiently perform the needed calculation.

The key sparsity inducing parameter of our algorithm is λ and the best method of fixing this parameter is unclear. What is clear however is that choosing λ in the standard way via cross validation on the one step ahead prediction task is not appropriate for inferring a causality graph. But, based on the roughly

² We remark however that for systems involving 2 or only a handful of processes that classical methods and statistical tests are more appropriate.

monotonic decrease in the number of edges as λ increases has lead us to suggest sweeping λ over a range and inspecting the resultant edge intensity, as described in section (4).

Combined with the vast literature which seeks to analyze the qualitative properties of large graphs (e.g. [23]) DWGLASSO may provide a fruitful approach to investigating large interacting systems in biology, finance, or other areas, as it is often the qualitative behaviour of such complex systems which are of ultimate interest. The most obvious downside of our approach in this context is that the interactions of these systems can be highly nonlinear and time variant. Although we made no consideration of these issues in this paper, Granger causality is robust to some non-stationarities, as long as the underlying causality graph remains constant. Our final remark is that the modeling power of the DWGLASSO algorithm can be extended via kernel methods in a fairly straightforward way, though the selection of additional hyperparameters then becomes a point of significant additional complexity.

References

1. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**(3) (1969) 424–438
2. Geweke, J.: Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association* **77**(378) (1982) 304–313
3. Geweke, J.F.: Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association* **79**(388) 907–915
4. Solo, V.: State space methods for granger-geweke causality measures. *arXiv preprint arXiv:1501.04663* (2015)
5. Barnett, L., Seth, A.K.: Granger causality for state-space models. *Physical Review E* **91**(4) (2015) 040101
6. Haufe, S., Müller, K.R., Nolte, G., Krämer, N.: Sparse causal discovery in multivariate time series. In: *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6, JMLR. org* (2008) 97–106
7. He, Y., She, Y., Wu, D.: Stationary-sparse causality network learning. *Journal of Machine Learning Research* **14**(1) (2013) 3073–3104
8. David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C., Depaulis, A.: Identifying neural drivers with functional mri: an electrophysiological validation. *PLoS Biol* **6**(12) (2008) e315
9. Barnett, L., Seth, A.K.: Detectability of granger causality for subsampled continuous-time neurophysiological processes. *Journal of Neuroscience Methods* **275** (2017) 93 – 121
10. Billio, M., Getmansky, M., Lo, A.W., Pelizzon, L.: Econometric measures of systemic risk in the finance and insurance sectors. *Working Paper 16223, National Bureau of Economic Research* (July 2010)
11. Canada, E.: Canadian weather energy and engineering datasets (cweeds) (2005)
12. Granger, C.: Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* **2** (1980) 329 – 352
13. Pearl, J.: *Causality*. Cambridge university press (2009)
14. Bach, F.R., Jordan, M.I.: Learning graphical models for stationary time series. *IEEE transactions on signal processing* **52**(8) (2004) 2189–2199

15. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996) 267–288
16. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1) (2006) 49–67
17. Syamantak Datta Gupta, R.R.M.: A frequency domain lasso approach for detecting interdependence relations among time series. In: *Proc. International Work-Conference on Time Series (ITISE, Granada, Spain, June 2014)*. (2014)
18. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2) (2005) 301–320
19. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3**(1) (2011) 1–122
20. Parikh, N., Boyd, S., et al.: Proximal algorithms. *Foundations and Trends in Optimization* **1**(3) (2014) 127–239
21. Jones, E., Oliphant, T., Peterson, P., et al.: *SciPy: Open source scientific tools for Python* (2001–)
22. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research* **12**(Jul) (2011) 2297–2334
23. Chung, F.R.: *Spectral graph theory*. Volume 92. American Mathematical Soc. (1997)

Development of a Routing Procedure to Assist an Earth Systems Model with Long Term Coastal Discharge Predictions

Josefine Wilms¹ and Marcus Thatcher²

¹ Council for Scientific and Industrial Research, Stellenbosch 7600, South Africa,
jwilms@csir.co.za,
<http://www.csir.co.za>

² Oceans and Atmosphere, Commonwealth Scientific and Industrial Research
Organisation, Aspendale, Australia, marcus.thatcher@csiro.au,
<http://www.csiro.au>

Abstract. Nearest neighbour searches, scaling, and a flow accumulation method were applied to improve predictions for freshwater deposits from land surfaces to the ocean for an earth systems model. Runoff, generated by the Conformal Cubic Atmospheric Model (CCAM), was read at a coarse resolution and downscaled, whereas digital elevation- and accumulation values were obtained from the HydroSHEDS database and upscaled. The accumulation, digital elevation, and runoff values were matched using a KDTree nearest neighbour algorithm. Starting from a zero-valued initial water body, CCAM runoff was routed to neighbouring cells. Flow direction was determined with a maximum flow accumulation method whereas the Manning equation was used to calculate the discharge rate. Inland reservoirs and coastal waters were removed and added to an outflow term. Mass conservation checks confirmed that the proposed procedure conserves mass and a 25-year simulation shows that the relative discharge rates, river routes, and outflow locations were sufficiently predicted.

Keywords: Long Term Forecasting, Runoff Routing, Earth Systems Model, K-d tree, Scaling, Manning Equation

1 Introduction

Runoff is the residual water from precipitation after evapotranspiration. This moisture, not absorbed by soil or plants, results in continental freshwater discharges into the ocean. Water evaporates from the ocean's surface, is transported back to the land as atmospheric moisture and reaches the land surface as precipitation. This process is known as the land-ocean water cycle [3].

Evapotranspiration and precipitation vary spatially but the return of runoff into the ocean is mostly concentrated at the world's largest river mouths. This significant freshwater discharge at mouth locations results in the salinity of ocean-water to be less within these regions. Salinity differentials, in turn, result

in regional changes of the ocean's density [11]. Estimates of freshwater fluxes into the ocean is therefore needed to study oceanic freshwater budgets. Since stream discharge can be measured quantitatively, such estimates are also important to check that earth systems models (ESMs) are adhering to mass conservation and performing with reasonable accuracy.

The performance of mapping- and prediction methods for the routing of terrestrial runoff fields are constantly improved upon. Fekete et al. [6] used river discharge information from gauging stations from the World Meteorological Organization Global Runoff Data Centre (GRDC) to calculate annual inter-station runoff. In addition, they simulated river discharge with a water balance model (WBM), driven by long-term mean monthly climate data. These WBM simulation results were then weighted by multiplying the value of each simulation point with the ratio of its discharge to the observed runoff of the corresponding inter-station region from the GRDC data. Using this method, a set of spatially distributed runoff fields were created at a 0.5° resolution.

Recently, Mizukami et al. [10] developed the routing tool, mizuRoute. MizuRoute can use both small- and large scale runoff outputs from land-surface models as input and produces a spatially distributed streamflow at various spatial scales. It can use both grid- and vector based river networks and applies two different river routing schemes: kinematic wave tracking and impulse response function-unit-hydrograph routing [10].

Another, widely used, routing method is a river network model termed the Routing Application for Parallel Computation of Discharge (RAPID). It was developed for the National Hydrography Dataset Plus river network for which lateral inflow is obtained from a land surface model. A matrix-based version of the Muskingum method is applied to calculate flow and water volumes in all reaches [4].

However, for water budget analyses in ESMs, only the discharges at coastal river mouths are of interest. To estimate continental discharge with runoff fields, a river transport model that routes the terrestrial runoff into the correct river mouths is required. This study therefore follows a similar approach to that of Dai et al. [3] who used a river transport model (RTM), developed by Branstetter et al. [2], to route surface runoff to the ocean. The RTM they used, implemented a linear advection scheme at a resolution of 0.5° .

The objective of this work is to update and improve the continental discharge estimates of CCAM [9] and, in so doing, improve its water budget approximations and forecasting capabilities. However, ESMs such as CCAM are run at coarse scales and currently operate at a resolution of 1° , or marginally higher, for global simulations [1]. It would be computationally too expensive to run the entire globe at a resolution high enough to incorporate smaller scale events.

Finding and incorporating accurate but computationally cheap methods for up- and downscaling as well as mapping non-matching grids therefore form an integral part of our study: Runoff, generated by CCAM, is downscaled and mapped to upscaled elevation- and flow accumulation data that are obtained from HydroSHEDS [8]. A matrix based RTM is then used to route the terrestrial runoff

to the river mouths. In so doing, the freshwater fluxes and locations can be mapped back to the CCAM grid and used to drive its ocean model.

In the current work, the developed model is applied to the African continent only, but it should be noted that the procedures, discussed here, can be used for any landmass.

2 Methods

The modeling procedure, used in this study, is done in three stages and all code is written in Python 2.7: During the preprocessing stage, a digital elevation model (DEM)- and flow accumulation (FA) data are read and upscaled to a target resolution at which routing will be done. Runoff data, obtained from CCAM's NetCDF output at a coarse resolution are, in turn, downscaled to the target resolution and subsequently mapped to the new FA and DEM data locations. Flow direction is then determined from the upscaled accumulation set.

The second stage entails the routing of the water by utilising the flow direction information, obtained during the preprocessing procedure.

Postprocessing is done in the third stage: At the end of each simulation month, water budget results, obtained during the second stage, are written to a NetCDF file and visualisation of the results is done by opening this file in an open source, integrated data viewer called Panoply.

A detailed discussion on the procedures underlying each of the aforementioned modelling stages is given in Sections 2.1 to 2.3.

2.1 Preprocessing: Scaling and Mapping

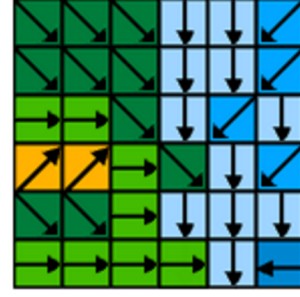
During preprocessing, FA- and DEM data are obtained from the HydroSHEDS [8] website. The DEM and FA data coincide spatially and is provided as binary interleaved (bil) format at a 30'' (1/120°) resolution.

FA and DEM data are read as two-dimensional arrays. The value of each location in the FA array denotes the number of locations that donate water to it. FA data are derived from DEMs by implementing a method developed by Jenson and Domingue [7] which calculates flow direction as the direction of steepest descent from DEM locations. For clarity, an example of such a process is shown in Figures 1a and 1b. For each (i, j) location in the FA array, the accumulation value is determined by summing the number of surrounding positions that deposit water into the particular location. An example of the relationship between flow direction and an FA is shown in Figures 1b and 1c.

FA and DEM arrays are upscaled to a 0.25° resolution. This is done by utilising a Python implementation of the Geospatial Data Abstraction Library (GDAL) that is specifically tailored to read and interpret binary geospatial data. Each bil file is accompanied by a header file that contains information about the number of rows, the number of columns, the coordinates of the upper left corner, and the step size between points. For the example used here, Africa, the upper left coordinates are given by $(-18.99583, 37.99583)$, the step size, for both

78	72	69	71	58	49
74	67	56	49	46	50
69	53	44	37	38	48
64	58	55	22	31	24
68	61	47	21	16	19
74	53	34	12	11	12

(a) Elevations.



(b) Flow directions.

0	0	0	0	0	0
0	1	1	2	2	0
0	3	7	5	4	0
0	0	0	20	0	1
0	0	0	1	24	0
0	2	4	7	35	1

(c) Flow accumulation.

Fig. 1. An example of a flow accumulation derivation from DEM values.

directions, is given as $0.008\bar{3}^\circ$ and the number of rows and columns are denoted by 8760 and 8880, respectively.

The information within the header files is used to construct an array of latitudes and longitudes. These are subsequently upsampled from $1/120^\circ$ to 0.25° by extracting every 30^{th} value. The resulting array lengths are then used to split the original accumulation matrix into equally sized sub-matrices, of which each contains 30×30 entries that signify the values within a 0.25° spacial range.

Upscaling of the accumulation matrix is done by following [5] and applying the Network scaling algorithm (NSA) by using a maximum value operator to aggregate each of the 30×30 grid values, i.e. the maximum accumulation value within each of the blocks is kept. An example case of the NSA method is illustrated in Figure 2 for upscaling a 6×6 to a 3×3 grid.

The DEM is upsampled in a similar manner, but, instead of keeping the maximum value, the average of each of the 30×30 elevation values is calculated.

Downscaling of routing values begins by reading runoff values in NetCDF format. Again, a Python library, NetCDF4, is utilised to obtain the data in the correct format. Each file contains the runoff values for one month at six-hourly intervals. The average runoff value for each day of the month is computed in order to have a single daily runoff matrix.

0	0	0	0	0	0
0	1	1	2	2	0
0	3	7	5	4	0
0	0	0	20	0	1
0	0	0	1	24	0
0	2	4	7	35	1

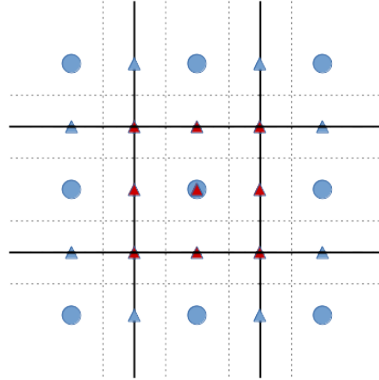
(a) Original flow accumulation.

1	2	2
3	20	4
2	7	35

(b) Upscaled flow accumulation.

Fig. 2. An example of the NSA to scale a 6×6 to a 3×3 grid.

Python's PySAL library [12] is used to construct a k-d tree from the up-scaled target latitudes and longitudes. The k-d tree is used to map each runoff value, in equal parts, to its nine nearest neighbours in the FA- and DEM arrays. The original runoff values are of such a nature that it coincides to one of the accumulation value coordinates. For clarity, the refinement procedure is shown in Figure 3.

**Fig. 3.** An example of grid refinement from 0.5° to 0.25° resolution.

The preprocessing stage concludes with runoff-, FA-, and DEM arrays that coincide spatially at a resolution of 0.25° .

2.2 Routing

The routing procedure is initiated by determining out- and inflow locations for each coordinate value on the 0.25° resolution grid. The indices of each point's 8

neighbours are determined and the outflow location is set to the indices of the neighbour with the highest accumulation value. Each grid location can therefore deposit its contents to a single neighbouring location only.

Subsequently, it is determined whether a location will receive water. A two dimensional inflow array is initialised with a size equal to that of the outflow-location matrix and all values are set to False. If a neighbour's outflow-location is equal to the indices of this cell, then the inflow value is changed to True. Once the Truth condition of each point has been determined, the True values are isolated by adding their indices to a list of inflow-locations.

At this stage, it is therefore known to where a point deposits its contents and if it will receive content from neighbouring positions.

The routing procedure starts with initialising a two-dimensional zero-valued water body array. At the start of each day, runoff-volumes are added to the water body array. It is assumed that water is equally distributed within the $0.25^\circ \times 0.25^\circ$ area and the water level within each cell is therefore determined by dividing the runoff volume V by the area of its host cell.

For each cell the water body slope to its outflow location is calculated as

$$S = \frac{(z_{DEM}^{host} + z_{WL}^{host}) - (z_{DEM}^{out} + z_{WL}^{out})}{\Delta x}, \quad (1)$$

where S denotes the slope, z_{DEM}^{host} and z_{DEM}^{out} are the elevations of the host and its outflow location, respectively, whereas, z_{WL}^{host} and z_{WL}^{out} denote the water levels within these locations. The Haversine distance between the cell centres is given by Δx .

The discharge from each cell is given by the Manning equation,

$$Q_{out} = \left(\frac{1}{n}\right) AR^{2/3} \sqrt{S}, \quad (2)$$

where Q_{out} is the discharge rate in m^3s^{-1} and n is Manning's roughness coefficient, which, following [13], is set to 0.025. The width of the domain through which water travels is approximated as the square root of its area, $D = \sqrt{A}$, which allows the hydraulic radius to be expressed as $R = (Dz_{WL}^{host}) / (D + 2z_{WL}^{host})$.

The water volume that is discharged during a time step Δt is calculated as $V_{out} = Q_{out} \Delta t$. The amount of water deposited from any location on the grid within Δt is therefore known at this stage.

To calculate the volume of water V_{in} that a position will receive, the neighbours of each location, recorded in the inflow-locations list, are examined: For each neighbour it is determined whether its outflow corresponds to the inflow-location. If it does, the discharge from this neighbour is added to the inflow volume of the point that is being analysed.

Once the inflow to- and outflow from each point have been calculated, the water body values can be updated as

$$V(t + \Delta t) = V(t) - V(t)_{out} + V(t)_{in}. \quad (3)$$

Finally, the updated water body is examined and water is removed from locations that either receive water but does not have an outflow location or are located on the coast.

At the end of each month discharge, water level, water volume, and the total amount of water that is deposited as outflow are written to binary files.

The routing algorithm is shown in Figure 4.

2.3 Post-Processing

For visualisation purposes, the binary output files are converted to a single Network Common Data Form (NetCDF) file. and an open source software, Panoply, is used to view results.

The discharge values can then be mapped back to the CCAM grid and used as input to CCAM's ocean procedure.

3 Results

A simulation was done for the African continent for a 24-year period spanning from 1981/01/01 to 2005/03/01. Courant numbers, discharge rates, and water volumes were recorded for each grid location on a daily basis. To orientate the reader a map, illustrating the actual locations of the largest African rivers, is shown in Figure 5. Courant numbers Co for 2005/03/01 are shown in Figure 3 and were calculated as

$$Co = \frac{|u|\Delta t}{\Delta x}, \quad (4)$$

where $|u| = Q_{out}/A$ is the velocity at which water is discharged at time t . Figure 3 shows the discharge rates for 2005/03/01 that were calculated using Eq. 2.

Adherence to mass conservation is checked daily by calculating the total mass of water on land using two methods: Firstly, a landmass water volume is determined by the summation of all values within the water body array, located on land. Secondly, a land mass water volume is calculated by computing the amount of runoff that has entered the system up to this point in time and subtracting the amount of water that has left the system as output. The aforementioned output includes locations that have no outflow but contains water (inland reservoirs) as well as water within cells that are located on the coast.

Values obtained from these two computations are then subtracted and should be close to zero if mass conservation is adhered to. Results for mass conservation are shown in Figure 7.

The solution method used in this model is explicit and therefore the Courant numbers should be significantly smaller than one. Figure 3 shows that the Courant number satisfies the aforementioned CFL condition in that the maximum value is $7.7e - 4$. Visual inspection of the discharge rates show that the Congo yields the highest discharge rate followed by the Nile, Niger, and the Zambezi. In reality, the Niger should dominate the Nile and the Zambezi. When comparing the simulation results with the actual positions of the major African

rivers, shown in Figure 5, it is concluded that the locations of the largest rivers and their mouths are located at approximately the correct locations.

Output from the original CCAM routing algorithm is shown in Figure 8.

Visual inspection confirms that the updated routing procedure is an improvement on the original since the rivers are now more clearly defined and the severe growth of inland water bodies has been subdued.

The simulation is computationally cheap: A decade was simulated in less than two hours on a single processor of a Lenovo laptop with 7.7 GiB Memory and an Intel Core i7-6500U CPU @ 2.50GHz.

4 Conclusions

A model has been presented for determining the locations and discharges of rivers into the ocean on a 0.25° grid and was applied to the African continent. The accuracy of predictions for river- as well as river-mouth locations have been improved upon when the discharge results are compared to the original discharge output from CCAM, shown in Figure 8. It should be noted that there is a slight increase in volume of the total water body over a period of 24 simulation years. Since it is assumed that the density of water is constant, this amounts to a slight increase in mass. However, the total water volume on land for which a maximal increase of 0.017 m^3 is recorded, is of the order $10e+11$. The increase is therefore comparatively small.

The discharge of the Niger relative to that of the Nile and Zambezi has been under-predicted. This could be due to the fact that the simulation is started with a zero water level in that it is forced solely by runoff and the rivers may therefore not have stabilised after 24 simulation years. The discharge discrepancies could also be the result of using an incorrect roughness factor in the Manning equation. Currently, the roughness coefficient is kept constant for all rivers. This may be an unrealistic assumption and the method used for discharge prediction warrants further investigation.

Although the algorithm does not take long to run, parallelisation of the procedure would allow the authors to test its performance for longer simulation periods at finer spatial and temporal scales.

Acknowledgments

The authors wish to thank the CSIR for funding and CSIRO for providing the ESM.

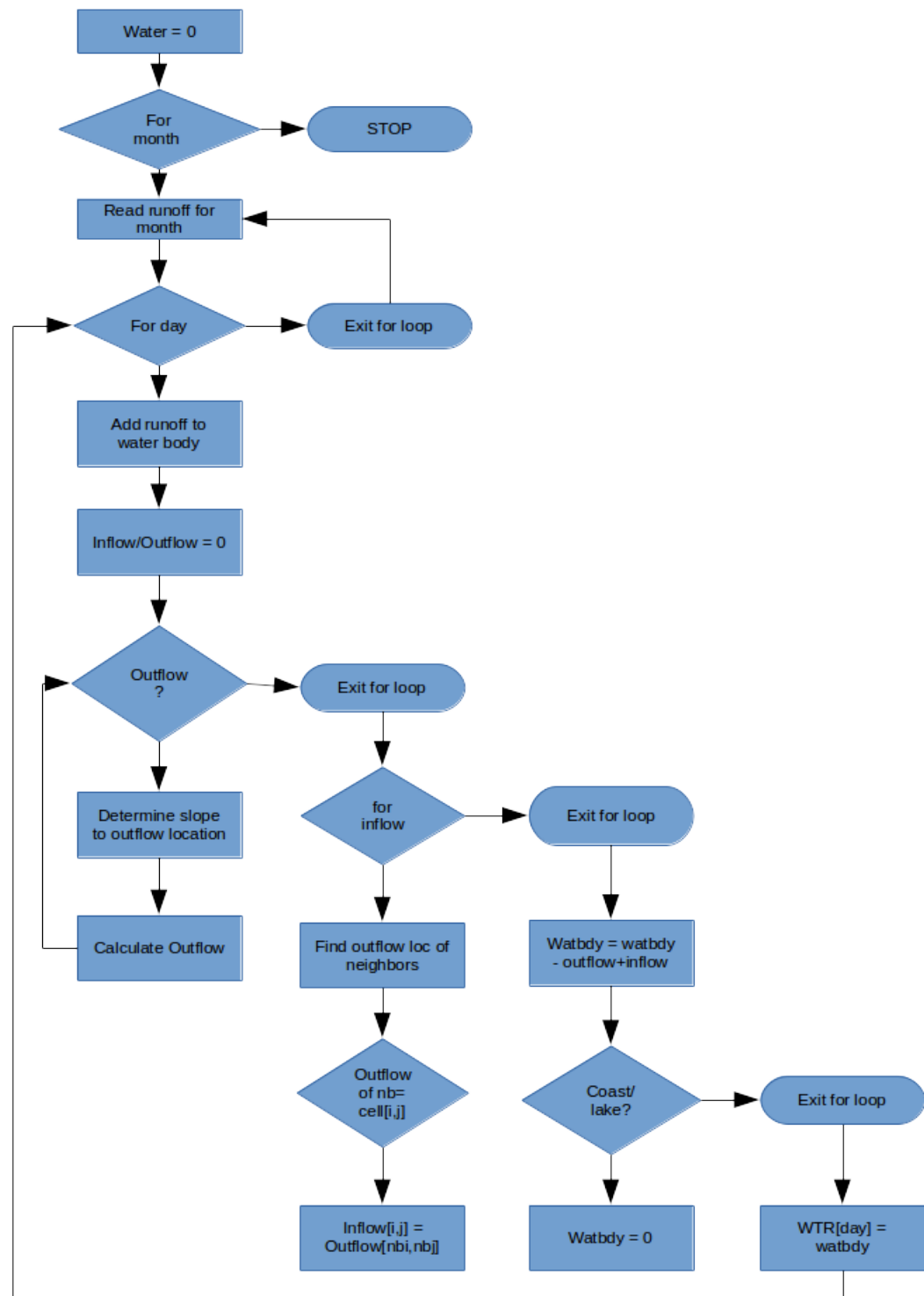


Fig. 4. Flowchart for routing.



Fig. 5. African rivers.

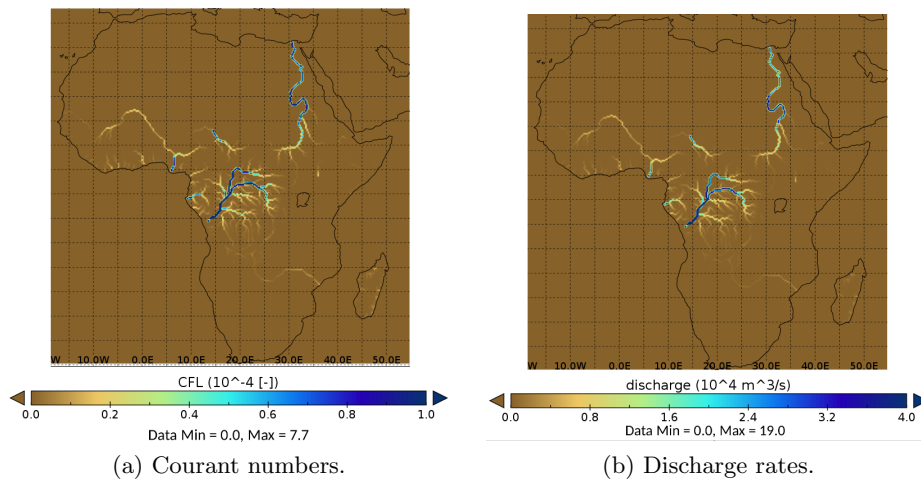


Fig. 6. Courant numbers and discharge rates for 2005/03/01.

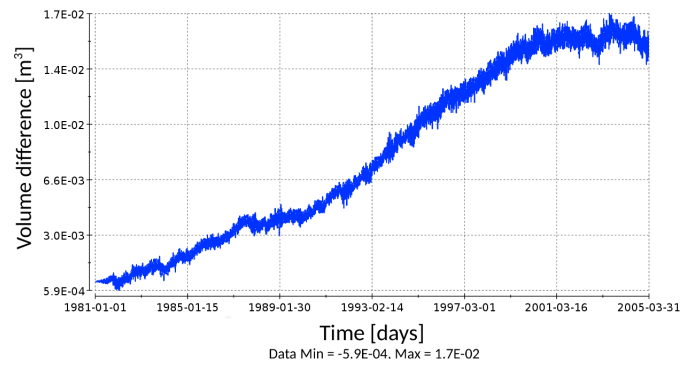


Fig. 7. Total Volume (Mass) conservation check results for 1981/01/01-2005/03/01.

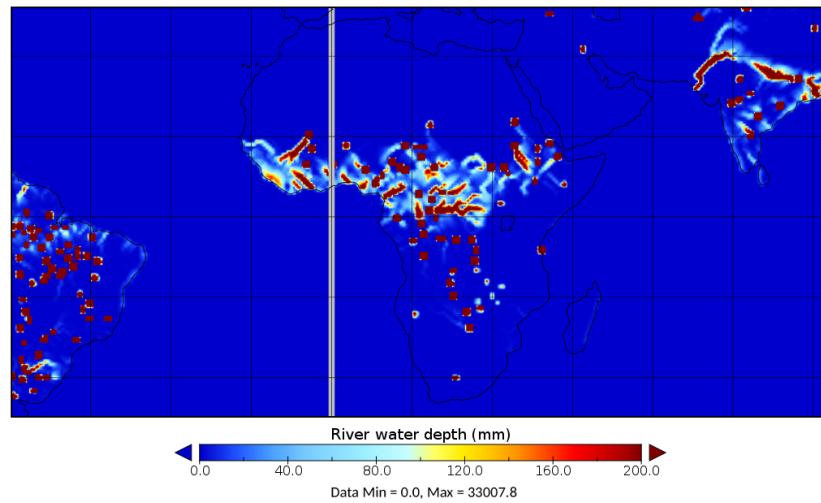


Fig. 8. Original routing output from CCAM simulation.

References

1. Encyclopedia of Sustainability Science and Technology, chap. Coupled Climate and Earth System Models. Springer-Verlag New York (2012)
2. Branstetter, M.: Development of a parallel river transport algorithm and applications to climate studies. Ph.D. thesis, University of Texas, Austin (2001)
3. Dai, A., Trenberth, K.: Estimates of freshwater discharge from continents: Latitudinal and seasonal variations. *Journal of Hydrometeorology* 3, 660–687 (Jul 2002)
4. David, C., Maidment, D., Niu, G.Y., Yang, Z.L., Habets, F., Eijkhout, V.: River network routing on the nhdplus dataset. *Journal of Hydrometeorology* 12, 913–934 (Mar 2011)
5. Fekete, B., Vorosmarty, C., Lammers, R.: Scaling gridded river networks for macroscale hydrology: Development, analysis, and control of error. *Water Resources Research* 37, 1955–1967 (Jul 2001)
6. Fekete, B., Vorosmarty, C., Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances. *Global Biogeochemical Cycles* 16(3), 15–1–15–10 (2002), <http://dx.doi.org/10.1029/1999GB001254>
7. Jenson, S., Domingue, J.: Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric Engineering and Remote Sensing* 54, 1593–1600 (1988)
8. Lehner, B., Verdin, K., Jarvis, A.: New global hydrography derived from spaceborne elevation data. *Transactions, AGU* 89, 93–94 (2008)
9. McGregor, John L. and Dix, M.R.: An Updated Description of the Conformal-Cubic Atmospheric Model, pp. 51–75. Springer New York, New York, NY (2008), http://dx.doi.org/10.1007/978-0-387-49791-4_4
10. Mizukami, N., Clark, M., Sampson, K., Nijssen, B., Mao, Y., McMillan, H., Viger, R., Markstrom, S., Hay, L., Woods, R., Arnold, J., Brekke, L.: mizuroute version 1: a river network routing tool for a continental domain water resources applications. *Geosci. Model Dev* 9, 2223–2016 (Jun 2016)
11. Nakamura, M.: Effects of ice albedo and runoff feedbacks on the thermohaline circulation. *Journal of Climate* 9, 1783–1794 (Aug 1996)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011)
13. Zhao, Y.: Modeling of River-flow Routing Using a Muskingum-and-Manning Method and Application in Basin of Seine. Ph.D. thesis, Pierre and Marie Curie University, Paris (2006)

Short-term Stream Flow Forecasting at Australian River Sites using Data-driven Regression Techniques

Melise Steyn^{1,2}, Josefine Wilms², Willie Brink¹, and Francois Smit¹

¹ Applied Mathematics, Stellenbosch University, Stellenbosch, 7600, South Africa, {melisedt,wbrink,fsmit}@sun.ac.za, <http://appliedmaths.sun.ac.za>

² Council for Scientific and Industrial Research, Stellenbosch, 7600, South Africa, jwilms@csir.co.za, <http://www.csir.co.za>

Abstract. This study proposes a computationally efficient solution to stream flow forecasting for river basins where historical time series data are available. Two data-driven modeling techniques are investigated, namely support vector regression and artificial neural network. Each model is trained on historical stream flow and precipitation data to forecast stream flow with a lead time of up to seven days. The Shoalhaven, Herbert and Adelaide rivers in Australia are considered for experimentation. The predictive performance of each model is determined by the Pearson correlation coefficient, the root mean squared error and the Nash-Sutcliffe efficiency. The performance of our data-driven models are compared to that of a physical stream flow forecasting model currently supplied by Australia's Bureau of Meteorology. It is concluded that the data-driven models have the potential to be useful stream flow forecasting tools in river basin modeling.

Keywords: Stream Flow Forecasting, Support Vector Regression, Artificial Neural Networks

1 Introduction

Stream flow is an important component in the hydrological cycle and plays a vital role in many hydraulic and hydrological applications. Research on model-generated stream flow is used by river engineers and scientists for the study of various hydro-environmental aspects, such as the increasing international concern of riverine pollution and the growing flood stages of rivers [5]. The devastating consequences of natural disasters, such as floods, can be lessened or even prevented through accurate stream flow forecasts [15].

Two main types of stream flow forecasting models can be distinguished, based on available information: physical and empirical. A physical model consists of governing partial differential equations that describe the physical laws of a specific system. Empirical or data-driven models are based on observed data that characterize the system [16].

A physical rainfall-runoff model can be used to transform rainfall estimations to runoff by modeling the hydrologic processes within a catchment, such as interception, evaporation, overland and subsurface flow [8]. According to Perrin *et al.* [14], it can be challenging to choose an appropriate model structure and complexity for accurate simulation of hydrological behavior at catchment scale.

During the past few decades, considerable progress has been made in the study of data-driven models to simulate the rainfall-runoff relationship [16]. Various processes within a river basin are characterized by measurable state variables, such as stream flow, precipitation, temperature and humidity. A river basin for which historical time series data are available is therefore a good candidate for the implementation of data-driven models.

In this paper the practicality of data-driven models for stream flow forecasting with a lead time of up to seven days are investigated. In particular, two supervised machine learning models are constructed, namely support vector regression (SVR) and artificial neural network (ANN). Australian river sites are considered, mainly because of a sufficient amount of available historical stream flow and precipitation data.

The Bureau of Meteorology (BOM), Australia's national weather and climate agency, provides a forecasting service that supplies stream flow predictions at more than 100 locations across Australia. These forecasts are determined by a computer based system which uses a rainfall-runoff model known as GR4H as its main component.³ It determines the total amount of rainfall in a specific catchment, the fraction of rainfall that ends up as runoff, and the accumulation of that runoff in downstream rivers [14]. Forecasts are given for a lead time of up to seven days, and are used for several water management purposes. The predictive capabilities of our data-driven models will be compared to the BOM rainfall-runoff model.

2 Overview of SVR and ANN

We proceed with a cursory theoretical overview of the two data-driven prediction methods considered in this paper.

2.1 Support Vector Regression

Support vector machines were originally developed to solve classification problems, but have been extended to the task of regression and time series prediction in the form of support vector regression (SVR). Many hydrological prediction problems have been addressed using SVR [15].

Consider a training set of n real-valued data pairs $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, where \mathbf{x}_i is an input vector in some space X , with corresponding output value y_i . A generalized continuous-valued target function $f(\mathbf{x})$ is fit to the training set, such that a deviation of at most ϵ is obtained between each true

³ <http://www.bom.gov.au/water/7daystreamflow/about.shtml>

output and its corresponding predicted value, and that $f(\mathbf{x})$ is as flat as possible [6]. Assuming f to be linear, we may write

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b, \quad (1)$$

where $\mathbf{w} \in X$, $b \in \mathbb{R}$, and $\langle \cdot, \cdot \rangle$ denotes a dot product in X . In order to get f as flat as possible, the orientation parameter (or weight) \mathbf{w} should be minimized. Some of the data pairs might exceed the ϵ margin of error and cause the optimization problem to be infeasible. We introduce slack variables, denoted as ξ and ξ^* , to indicate the vertical distance from each data pair above and below the ϵ margins. The convex optimization problem is solved by minimizing

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*). \quad (2)$$

The positive penalty parameter C determines the tolerated deviations larger than ϵ . The minimization of (2) is a standard constrained optimization problem and can be solved by applying Lagrangian theory [4]. The weight vector is derived as

$$\mathbf{w} = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \mathbf{x}_i, \quad (3)$$

where α_i and α_i^* represent Lagrangian multipliers associated with the training points above and below the regression line, respectively. The value of b in equation (1) is computed by exploiting the Karush-Kuhn-Tucker conditions [7, 10], as explained by Granata *et al.* [6].

In many applications the relationship between inputs and outputs in the training data might show complex nonlinear behavior. A kernel function can be introduced to implicitly map the training points from the original input space X to a higher dimensional feature space $\Phi(X)$, such that a linear relationship between the variables exist in $\Phi(X)$. The support vector expansion of the target function for linear regression is then applicable in the feature space. Equation (1) changes to

$$f(\mathbf{x}) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) k(\mathbf{x}_i, \mathbf{x}) + b, \quad (4)$$

where

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle. \quad (5)$$

The radial basis function (RBF) is a widely used kernel in hydrological prediction applications [6], and is defined as

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2), \quad (6)$$

where $\gamma > 0$ is a kernel-specific hyperparameter. Choosing an optimal value for λ , as well as for ϵ and C , is important when training an SVR model to fit a given dataset [6].

2.2 Artificial Neural Network

Artificial neural networks (ANNs) are especially suitable when the underlying functions that describe complex phenomena are unknown [20], and have been used extensively for hydrological modeling purposes [11].

An ANN contains a set of interconnected nodes that receive, process and send information to one another over weighted connections. These nodes are grouped in different layers. Input values enter the model through the first layer (the input layer). The data is then fed forward through successive hidden layers until it reaches the final layer (the output layer). The hidden layers enable the ANN to learn complex relationships between input and output data [16]. An ANN can be single layered, bilayered or multilayered, depending on the number of hidden layers.

ANNs are further classified as feed-forward or recurrent, based on the direction of information flow and processing between nodes. Feed-forward ANNs allow information to travel only from the input layer to the output layer, while recurrent ANNs allow information to travel in both directions. For each node, an output is determined by calculating the sum of its weighted input nodes and applying a nonlinear activation function. According to Maier and Dandy [11], the sigmoidal-type and logistic sigmoidal-type (such as tanh) activation functions are frequently used in hydrological applications:

$$\text{sigmoidal-type: } g(z) = \frac{2}{1 + \exp(-2z)} - 1, \quad (7)$$

$$\text{tanh: } g(z) = \frac{1}{1 + \exp(-z)}, \quad (8)$$

where z represents the weighted sum of a particular node's inputs. This result is then used as input for the nodes in a succeeding layer. A linear activation function is considered for the final hidden layer of regression models [11].

Training is achieved by finding an optimal set of connection weights that minimize the estimated error between the true output values and the output values that are determined by the network.

3 Methodology

A description of the procedures to construct SVR and ANN models for stream flow forecasting at specific river sites follows.

3.1 Study Area and Data

High quality time series of daily stream flow and precipitation data for the Australian river sites under study were obtained from the Australian Bureau of Meteorology's Hydrologic Reference Stations (HRS) and Climate Data Online (CDO) services, respectively. The HRS network consists of over 200 river sites that are mostly unaffected by water-related systems, such as dam construction

and irrigation services, and located in different hydro-climatic regions across Australia. CDO provides access to precipitation records from the Australian Data Archive for Meteorology.

Three Australian river sites are considered for this study: the Shoalhaven River at Fossickers Flat in New South Wales, the Herbert River at Abergowrie in Queensland, and the Adelaide River at Railway Bridge in Northern Territory. The Shoalhaven River is located in a temperate climate region and has a catchment size of 4660 km². The stream flow data for this site were obtained from gauging station 215207 (150.18° E, 34.82° S) and the corresponding precipitation from station 068085, 5.3 km away from station 215207. The Herbert River is in a subtropical climate region and has a catchment size of 7488 km². Its stream flow data were obtained from gauging station 116006B (145.92° E, 18.49° S) and the precipitation data from station 032091, 8.7 km away from station 116006B. The Adelaide River is in a tropical climate region and has a catchment size of 638 km². Its stream flow data were obtained from gauging station G8170002 (131.11° E, 13.24° S) and the precipitation data from station 014237, 3.3 km away from station G8170002.

Only uninterrupted time series data were used for training: data from 1 January 2000 to 31 December 2014 for training the data-driven models at the Shoalhaven and Herbert rivers, and data from 1 January 2008 to 31 December 2012 for the Adelaide river. For all three river sites, data from 5 February 2017 to 5 May 2017 were used as test data.

3.2 Input Selection, Data Preprocessing and Cross Validation

A moving time window is considered for the generation of input and output data pairs. For each measured stream flow value (which is considered as an output value), a corresponding input vector contains the precipitation and stream flow values of the preceding p -day and q -day time windows, respectively. For this study, p ranges from 0 to 2 and q from 2 to 5. P represents precipitation, Q represents stream flow, t refers to the current day and d refers to the forecasting lead time. An output value Q_{t+d} then has an input vector $\{P_t, P_{t-1}, \dots, P_{t-p}, Q_t, Q_{t-1}, \dots, Q_{t-q}\}$. For each model that forecasts with a lead time of d days, an exhaustive search is followed during training to find optimal values for p and q .

Data preprocessing is implemented by normalizing the values in the dataset to a range of $[0, 1]$. This ensures that the influence of large feature values (like stream flow) does not dominate that of smaller feature values (like precipitation) during the training process.

As discussed in Section 3.1, the available datasets are split into a training set and a test set. In order to obtain a model that generalizes well to unseen data, 10-fold cross validation is introduced, i.e. the full training dataset is split into 10 folds of equal size. Each fold is considered as a validation set once, while the remaining 9 folds are combined to form a training set. Ultimately, the model with the lowest average validation error on all 10 trials is used for forecasting purposes, and tested on the test set [16].

3.3 Model Performance Evaluation

Three quantitative indices are considered to evaluate the performance of the SVR and ANN models, and to compare them to the physically based BOM model. These are the Pearson's correlation coefficient (r), the root mean squared error (RMSE) and the Nash-Sutcliffe efficiency (NSE):

$$r = \frac{\sum_{i=1}^m (y_i - \bar{y})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^m (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^m (f_i - \bar{f})^2}}, \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - f_i)^2}, \quad (10)$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^m (y_i - f_i)^2}{\sum_{i=1}^m (y_i - \bar{y})^2}, \quad (11)$$

where y_i and f_i represent each of the m true and forecasted outputs in the test set, respectively. The average of all true outputs is represented by \bar{y} and the average of all forecasted outputs by \bar{f} .

Pearson's correlation coefficient gives the extent to which the input and output values are linearly correlated, and ranges between -1 and 1 . A value close to -1 or 1 shows a strong linear relationship between the two variables, whereas values close to zero show little to no linear relationship. If the predicted values of the model increase as the input values increase, a positive r -value is obtained. If the predicted values decrease as the input values increase, a negative r -value is obtained.

The RMSE measures the difference between a model's predicted outcomes and the true outcomes from the system that is being modeled. The smaller the RMSE value, the better the performance of the model.

The NSE is used to assess the predictive power of a model and is always less than or equal to 1 . A model with an NSE of 1 corresponds to a perfect match of predicted outcomes to true outcomes. An NSE of 0 indicates that the model's predictive capability is the same as considering the mean true outcome value as a predictor. An NSE less than 0 occurs when the mean true outcome value would have been a more reliable predictor than the model [9]. According to Noori and Kalin [13], a model can be considered "good" if the NSE is above 0.5 , and "very good" if it is above 0.7 .

3.4 SVR Hyperparameters

The SVR model with an RBF kernel is considered for this study. Three parameters have to be selected, namely C , ϵ and γ . We pick possible C values ranging from 1 to 10^4 , ϵ values from 10^{-3} to 10^{-1} , and γ values from 10^{-4} to 1 . An exhaustive grid search is performed to find the combination of parameters with optimal performance during training and cross validation.

3.5 ANN Architecture

According to Maier and Dandy [11], a one hidden layered feed-forward neural network provides suitable complexity to reproduce the nonlinear behavior of hydrological systems and has been suitable for forecasting hydrological variables in various studies.

It can be challenging to choose an appropriate number of hidden nodes within the hidden layer, as too few might result in a network that cannot capture the complex relationship between input and output, while too many may cause overfitting. This study uses two different methods as bounds for the number of hidden nodes, as proposed by Belayneh and Adamowski [1]. Wanas *et al.* [18] determined that the optimal performance of a neural network is obtained with $\log(n)$ hidden nodes, where n is the number of training samples. Mishra and Desai [12] showed that optimal results are obtained with $2N+1$ hidden nodes, where N is the number of input nodes. Following Belayneh and Adamowski [1], a trial and error approach can be implemented during training to find the optimal number of hidden nodes ranging from $\log(n)$ to $2N+1$.

As discussed, the sigmoidal-type and logistic sigmoidal-type activation functions, given in equations (7) and (8), have been used frequently in hydrological applications. We implement both, and pick the one that achieves the lowest error during training and cross validation.

4 Results and Discussion

Results for the optimal input features, hyperparameter combinations for SVR and architecture for ANN are discussed in the following subsections. The predictive capabilities of our data-driven models are also evaluated, based on the criteria listed in Section 3.3.

4.1 Parameter Selection

Different lead times are considered for stream flow forecasting, ranging from 1 day to 7 days in advance. As stated in Section 3.2, the preceding time windows for stream flow and precipitation that provide an optimal model are found separately during training for each of the different prediction lead times. For SVR, an optimal combination of hyperparameters is also determined, whereas for ANN, an optimal number of hidden nodes and the choice of activation function. Results are listed in Tables 1 and 2.

It can be observed that, when considering different prediction lead times, the preceding time windows for stream flow and precipitation and the combination of model parameters vary. It is also noticeable that only the optimal ANN and SVR models for 7 day lead time forecasting of the Shoalhaven river site do not consider any rainfall values. Apart from this particular case, it appears that rainfall is an important input to the data-driven models for the three considered river sites. Furthermore, each ANN model achieved the lowest error during training and cross validation when considering the tanh activation function.

Table 1. Optimal input features and hyperparameters in the SVR models for the three gauging stations (C , ϵ and γ are SVR parameters; the model uses precipitation data from days $t - p$ to t and stream flow data from days $t - q$ to t to predict stream flow on day $t + d$, with d the lead time).

Lead time (d)	Shoalhaven					Herbert					Adelaide				
	p	q	C	ϵ	γ	p	q	C	ϵ	γ	p	q	C	ϵ	γ
1 day	3	2	100	0.001	0.1	5	2	100	0.001	0.1	5	2	1	0.001	1
2 day	2	1	10	0.001	0.1	5	1	1000	0.001	0.1	4	2	1000	0.001	0.01
3 day	2	2	1	0.001	0.1	4	1	10000	0.01	0.1	5	1	10000	0.01	0.01
4 day	3	2	100	0.001	0.001	4	1	10000	0.01	0.1	5	2	10000	0.01	0.01
5 day	3	2	100	0.001	0.001	2	2	1000	0.001	0.001	5	2	10000	0.01	0.01
6 day	2	2	100	0.001	0.01	2	1	10000	0.01	0.1	5	2	10000	0.01	0.01
7 day	2	0	10	0.001	0.1	2	1	10000	0.01	0.1	2	1	10000	0.01	0.1

Table 2. Optimal input features and architecture (number of nodes in the hidden layer, h) in the ANN models for the three gauging stations.

Lead time (d)	Shoalhaven			Herbert			Adelaide		
	p	q	h	p	q	h	p	q	h
1 day	3	2	9	5	2	12	3	2	4
2 day	3	1	5	3	1	8	4	2	6
3 day	5	1	11	5	2	5	5	2	3
4 day	5	2	10	4	1	4	5	2	3
5 day	5	2	10	4	1	4	3	2	8
6 day	2	2	4	4	1	4	5	1	4
7 day	4	0	3	3	1	3	5	2	3

4.2 Performance Evaluation

The efficiency criteria used in this study are the Pearson correlation coefficient, the root mean squared error and the Nash-Sutcliffe efficiency. Based on these performance indices, the SVR and ANN models that performed optimally on the training and validation sets were applied to the (as yet unused) test sets of the three river sites under study. Results are shown in Tables 3 to 5. For comparison, prediction accuracies made by the Bureau of Meteorology's stream flow forecasting model are also given.

ANN outperforms the SVR and BOM models for stream flow predictions with a lead time of 1 to 2 days at the Shoalhaven river site. The base flow as well as the rising and falling limbs of the hydrographs are well represented by the ANN model. However, the peaks are under- and over-predicted. As the prediction lead time increases, the accuracy of each model decreases. Figures 1a and 1b show how the time lag between observed and forecasted peaks increase. Furthermore, Figure 1c shows that the SVR and ANN models fail to forecast the rising limbs of the hydrograph for predictions with a lead time longer than

Table 3. Performance evaluation for stream flow forecasting at the Shoalhaven river station of our trained SVR and ANN models as well as the physically based model used by the Australian Bureau of Meteorology (BOM).

Lead time	r			RMSE			NSE		
	SVR	ANN	BOM	SVR	ANN	BOM	SVR	ANN	BOM
1 day	0.87	0.90	0.85	541	458	866	0.74	0.81	0.34
2 day	0.74	0.81	0.71	807	629	1357	0.43	0.65	-0.61
3 day	0.71	0.65	0.59	948	826	1601	0.22	0.41	-1.22
4 day	0.66	0.46	0.52	1040	969	1446	0.07	0.19	-0.79
5 day	0.54	0.52	0.28	1118	933	3272	-0.07	0.26	-8.13
6 day	0.39	0.33	0.18	1150	1263	5760	-0.11	-0.35	-27.11
7 day	0.32	0.26	0.18	1195	1094	3957	-0.20	-0.01	-12.17

Table 4. Performance evaluation for stream flow forecasting at the Herbert river station of our trained SVR and ANN models as well as the physically based model used by the Australian Bureau of Meteorology (BOM).

Lead time	r			RMSE			NSE		
	SVR	ANN	BOM	SVR	ANN	BOM	SVR	ANN	BOM
1 day	0.93	0.92	0.95	1627	1728	1748	0.85	0.83	0.83
2 day	0.79	0.82	0.90	2721	2525	2152	0.59	0.64	0.74
3 day	0.70	0.73	0.82	3067	2952	3138	0.48	0.52	0.45
4 day	0.59	0.60	0.74	3514	3691	3707	0.32	0.25	0.25
5 day	0.52	0.53	0.28	5996	3936	12911	-0.95	0.16	-8.04
6 day	0.42	0.48	0.12	4154	3982	23648	0.08	0.15	-28.97
7 day	0.38	0.40	0.09	4116	4391	21508	0.10	-0.02	-23.49

Table 5. Performance evaluation for stream flow forecasting at the Adelaide river station of our trained SVR and ANN models as well as the physically based model used by the Australian Bureau of Meteorology (BOM).

Lead time	r			RMSE			NSE		
	SVR	ANN	BOM	SVR	ANN	BOM	SVR	ANN	BOM
1 day	0.79	0.85	0.84	975	944	925	0.61	0.63	0.65
2 day	0.63	0.67	0.54	1287	1277	1466	0.33	0.34	0.12
3 day	0.53	0.51	0.40	1376	1965	1671	0.21	-0.62	-0.17
4 day	0.41	0.43	0.25	1533	1983	1870	0.00	-0.67	-0.49
5 day	0.38	0.47	0.13	1561	1980	2036	-0.03	-0.66	-0.75
6 day	0.33	0.35	0.16	1567	2188	1856	-0.03	-1.00	-0.44
7 day	0.27	0.26	0.19	1602	2158	1835	-0.06	-0.92	-0.39

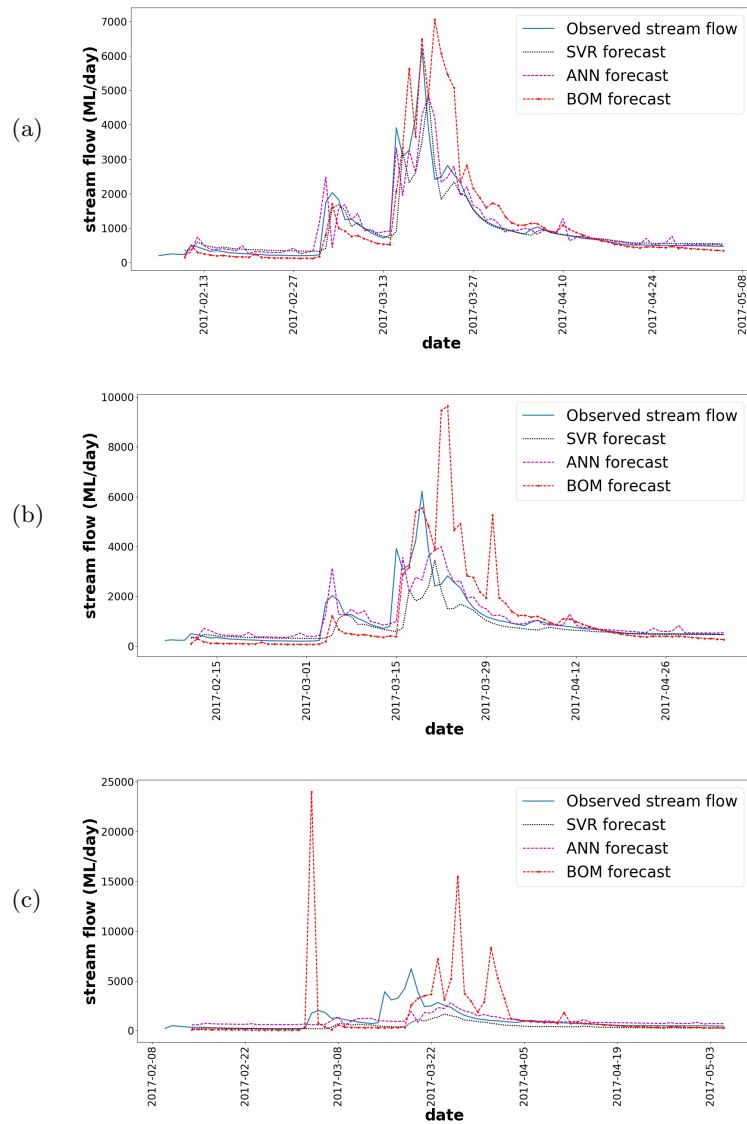


Fig. 1. Daily stream flow predictions for (a) 1 day, (b) 2 day and (c) 5 day lead time forecasts, for the Shoalhaven station.

4 days. This can be attributed to the absence of information (such as rainfall events) when increasing the prediction lead time. For lead times greater than 3 days, the SVR forecasts show the strongest correlation to the observed stream flow, whereas the ANN generally performs better in terms of RMSE and NSE. For 6 and 7 day lead time predictions, the NSE of all three models are negative,

indicating that the mean value of the observed outcomes would have been a more reliable predictor than the forecasting models.

No single model outperforms the rest on the test set of the Herbert river station. For instance, the BOM model obtains the strongest Pearson correlation (0.95) to the observed stream flow when forecasting 1 day in advance, but fails to determine the peaks as accurately as the SVR model. The BOM model does, however, show the better performance in forecasting stream flow with a lead time of 2 days. Similar to the Shoalhaven river models, an increase in prediction lead time causes a decrease in model performance and an increase in lag times between observed peaks and forecasted peaks.

The BOM and ANN models show the better performance on the test set of the Adelaide river station for 1 and 2 day lead time predictions. For instance, as seen in Table 5, comparable r , RMSE and NSE results are obtained for both models. The SVR model shows the better forecasting performance for predictions with a lead time greater than 2 days. Similar to both Herbert and Shoalhaven, the prediction capabilities of all three models worsen with an increase in prediction lead time.

5 Conclusion

This study investigated the ability of data-driven modeling for stream flow forecasting with a lead time of up to 7 days. SVR and ANN models were employed to forecast stream flow at the Shoalhaven, Herbert and Adelaide gauging stations. The predictive capabilities of these data-driven models were compared to that of a physically based rainfall-runoff model. For 1 day lead time forecasts, each data-driven model properly modeled the stream hydrograph shape and the time to peak. However, a noticeable decrease in predictive capabilities with an increase in lead time occurred. The SVR method performed better than the BOM model for the Shoalhaven station, based on the evaluation criteria. For the other stations, no single model outperformed the others.

Based on the results obtained for this study, SVR and ANN models have the potential to be useful tools for short-term stream flow forecasting. They do not require specialized knowledge of physical phenomena, and are therefore especially useful when it is difficult to build a physically based model due to a lack of understanding of the underlying processes. It is also helpful to have modeling alternatives and to validate results obtained from physically based models to that of data-driven models. Furthermore, data-driven models are computationally efficient in the sense that once they are trained, predictions can be made very quickly. Data-driven models could also be combined with physically based models to form even more powerful and accurate hybrid forecasting models.

A limitation of data-driven models are, however, that substantial historical stream flow and precipitation data records should be available. Many of the existing gauging stations have limited available datasets, or a considerable amount of missing data. Developing machine learning techniques to address these problems may be considered in further studies.

References

1. Belayneh, A., Adamowski, J.: Drought Forecasting using New Machine Learning Methods. *Journal of Water and Land Development*. 18, 3-12 (2013)
2. Bureau of Meteorology Hydrologic Reference Stations, <http://www.bom.gov.au/water/hrs/about.shtml>
3. Climate Data Online, <http://www.bom.gov.au/climate/data/?ref=fttr>
4. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2. 121-167 (1998)
5. Falconer, R., Lin, B., Harpin, R.: Environmental Modeling in River Basin Management. *International Journal of River Basin Management*. 3(1), 169-184 (2005)
6. Granata, F., Gargano, R., De Marinis, G.: Support Vector Regression for Rainfall-runoff Modeling in Urban Drainage: a Comparison with the EPA's Storm Water Management Model. *Water*. 8(3), 69 (2016)
7. Karush, W.: Minima of Functions of Several Variables with Inequalities as Side Constrains. Master's Thesis, Department of Mathematics, University of Chicago, Chicago, IL, USA (1939)
8. Knapp, H.V., Durgunoglu, A., Ortel, T.W.: A Review of Rainfall-runoff Modeling for Stormwater Management. Illinois State Water Survey, Hydrology Division (1991)
9. Krause, P.; Boyle, D.P.; Bäse, F.: Comparison of Different Efficiency Criteria for Hydrological Model Assessment. *Advances in Geosciences*. 5, 89-97 (2005)
10. Kuhn, H.W., Tucker, A.W.: Nonlinear Programming. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, pp. 481-492. University of California Press, Oakland, CA, USA (1951)
11. Maier, H.R., Dandy, G.C.: Neural Networks for the Prediction and Forecasting of Water Resources Variables: a Review of Modeling Issues and Applications. *Environmental Modelling & Software*. 15, 101-124 (2000)
12. Mishra, A.K., Desai, V.R.: Drought Forecasting using Feed-forward Recursive Neural Network. *Ecological Modelling*. 198(1-2), 127-138 (2006)
13. Noori, N., Kalin, L.: Coupling SWAT and ANN models for enhanced daily streamflow. *Journal of Hydrology*. 533, 141-151 (2016)
14. Perrin, C., Michel, C., Andréassian, V.: Improvement of a Parsimonious Model for Streamflow Simulation. *Journal of Hydrology*. 279, 275-289 (2003)
15. Raghavendra, S., Deka, P.C.: Support Vector Machine Applications in the Field of Hydrology: a Review. *Applied Soft Computing*. 19, 372-286 (2014)
16. Solomatine, D.P., Ostfeld, A. Data-driven Modeling: some Past Experiences and New Approaches. *Journal of Hydroinformatics*. 10(1), 3-22 (2008)
17. Vapnik, V.: The nature of statistical learning theory. Springer, New York, NY, USA (1995)
18. Wanas, N., Auda, G., Kamel, M.S., Karray, F. On the Optimal Number of Hidden Nodes in a Neural Network. In: *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 918-921. IEEE Press, Michigan (1998)
19. Wang, Y., Guo, S., Xiong, L., Liu, P., Liu, D.: Daily Runoff Model based on ANN and Data Preprocessing Techniques. *Water*. 7, 4144-4160 (2015)
20. Zhang, Z.: A Gentle Introduction to Artificial Neural Networks. *Annals of Translational Medicine*. 4(19), 1-6 (2016)

An Implementation of HMM Classifier in High Dimensions Based on MapReduce

Badreddine Benyacoub ^{*}, Souad El Bernoussi ^{*}, Ismail ELMoudden ^{*},
Abdelhak Zoglat [†]

^{*} Laboratory of mathematics, computer science & applications

[†] Laboratory of applied mathematics

University Mohammed V, Faculty of Sciences

Rabat ,BP 1014, Morocco

Abstract

In the last decades, the amount of data storage has been increased, and data sets repositories which can be displayed is in the range of some five hundred thousand data items and more. However, this information cannot be practically analyzed and utilized on a single commodity computer because these data are generated on a massive scale. For this purpose, the large scale and high dimensionality data that must be acquired and processed requires high-performance computing facilities. The requirement for significant computation imposed by an high performance analytical systems have led to an increasing interest in the use of parallel and distributed environments. This paper develop a distributed environment that can execute the training process for HMM classifier. We find that an application of HMM learning algorithm on such large size datasets using the framework for big data MapReduce of the tool Matlab achieves significant speedups as the system is scaled up in multiprogrammed environments. As results, a parallel processing can be matched against all tuples of data for building HMM classifier. Experimental results show that the parallel version of our developed model could overcome the drawbacks of HMM for large datasets and can efficiently handle data in reasonable time.

keywords

Hidden markov model(HMM). Supervised learning. Classification. Big Data. MapReduce.

1 Introduction

Recently, The processing large size datasets becomes a research opportunities ahead in the field of machine learning [7]. The volume of data being retrieved is large and the processing this amount of information is especially challenging task for learning supervised problems [1,5]. The process of analyzing large amounts of data in order to build new kind of useful models such as implicit relationships between input attributes characterize information and the predefined class labels. Many traditional supervised learning algorithms, such as statistical models, Artificial intelligence methods and other relatives, become computationally infeasible for very large data sets.

The ability to build a classification models using collected large datasets with high speed is a pervasive problem that encompasses many diverse applications. However, such big datasets cannot be practically utilized on a single commodity computer because the data is too large to fit in memory, or takes more time. Thus, the training task for learning classification models requires a high capacity evolutionary computation. The storing, manipulation and analyzing of big data, parallel and distributed architectures allow us to avoid this obstacle and overcome the drawbacks of traditional techniques analysis [9]. Now, big data applications presents a new way to solve some

exploratory data analysis problems have arisen and become popular and useful in many knowledge discovery related fields [3,4].

The subject of classification is also a major research topic in the fields of statistical learning and probabilistic modeling is at the heart of the issue in research. The basic idea is to find subtle relationships in data and infer a Bayesian rules that allow the prediction of results. Conventional probabilistic classifier models have been used as a tool in supervised classification which can predict the probability of different class based on various attributes. However, these classification models cannot manage a large amount of data and the corresponding learning algorithm cannot efficiently do so. The data is so big it affects the types of algorithms we are willing to execute. This paper proposes a method to split incoming data into chunk and build classification model based on parameters resulting from these individual chunks. Our method extends earlier work by introducing a method for adaptively analyzing the chunk. The objective when applying the algorithm in practice is to reduce the running time and memory consumption. It also makes it possible to efficiently optimize the treatment to get results where the analyzed information remains relevant.

Statistical aspects of the analysis and use of high-dimensional data is the major focus of this work. We focus on the adaptation of the HMM estimating techniques with massive amount of information for building the classifier. HMM is the most powerful and stochastically modeling method in pattern recognition [6]. It has been used to model a dynamical system such as speech recognition, handwriting and engineering. In recent years, it was applied for estimating the relationships between inputs attributes and the class labels. It was also proposed as a statistical process for prediction and forecasting such as customer relationship management (CRM) [2,8]. The basic idea of learning HMM algorithm for large datasets is to estimate partially HMM parameters by splitting the data into chunks, learning HMM from each chunk, and combining the estimating of the different parameters to form a overall reliable estimates of all parameters.

HMM learning algorithm for building classification model cannot be directly applied to large data because it is too slow and require too much memory. The whole process is executed serially by one machine. HMM learning algorithm for building classification model cannot be directly applied to large data because it is too slow and require too much memory. The whole process is executed serially by one machine. MapReduce paradigm based distributed approaches can be used as a way to an analyst applying predictive models on large datasets and improve the scalability of the data processed. Thus we improve HMM classifier by using MapReduce. It has many advantages in data processing can reduce the training time and enhance the speed of classification.

The remainder of this paper is organized as follows: In Section 2, we first briefly present the basic elements related to HMM discrete states and discrete observations. Then, we introduce the modeling process of HMM method to build the classification model. Finally, we show experimental limits of training HMM parameters. Subsequently our scalable approach based on Mapreduce distributed environment and the decomposition of data sets into several chunks are provided in Section.3. We present in this section also the experimental results, evaluations and finally, some concluding remarks are given.

2 HMM classifier model

We first describe the basic algorithm that generates the HMM classifier from incoming of data. Then we explain how HMM can be utilized to build a supervised learning for classification.

2.1 Discrete States and Discrete observations for HMM

Consider a input data whose members are characterized by a set of independent or predictor variables called explanatory or exogenous and a set of class labels. Suppose we wish to identify a model that best fits a relationship between the attribute set and class label. This relationship might be presented statistically and estimated from a sequence of observations. The simplest

probabilistic form uses posterior probabilities $P(X|Y)$ and determine the appropriate class of X based on the input data Y . In this work we follow the hidden Markov model (HMM) approach taken in Benyacoub et.al and assume that the posterior probability can be described as a function of observed characteristics Y .

Here we briefly describe a hidden Markov model as given in Benyacoub et.al [2]. In general, the class label could be any one of N values. Formally, we identify the set of N class labels with the N units vectors e_1, e_2, \dots, e_N in R^N , where $e_i = (0, \dots, 1, 0, \dots, 0)'$. We can assume that the finite state of markov chain is defined by the set $S = \{e_1, e_2, \dots, e_N\}$ and we have $X \in S$. In fact, taking inner products notation and turns out that expectations are probabilities in that $P(X = e_i) = E[\langle X, e_i \rangle]$. We suppose there is a vector $Y = (Y^1, Y^2, \dots, Y^p)$ of characteristics which can be observed and each explanatory variable Y^k has a finite state space. Write $S_Y^k = \{f_{1k}, f_{2k}, \dots, f_{m_k k}\}$ the set of unit vectors that identify this finite state space where $f_{ik} = (0, \dots, 1, 0, \dots, 0)'$ in R^{m_k} . Then for each $k = 1, 2, \dots, p$ we have $P(Y^k = f_{jk}) = E[\langle Y^k, f_{jk} \rangle]$ where $Y^k \in S_Y^k$. We should apply a discretization method to quantize a continuous state space. The relationship between the state which represents the class label and each characteristic Y^k is then given by $E[Y^k|X] = C_k X$, where $C = (c_{ji}), 1 \leq i \leq N, 1 \leq j \leq m_k$ is a matrix.

Define $W^k = Y^k - C_k X$, thus, the connection between the class labels and the observed characteristics can be expressed as an observation equation presented in the following:

$$Y^k = C_k X^k + W^k, k = 1, 2, \dots, p$$

Those equations can be summarized in this equation $Y = CX - W$, where $C = (C_1, C_2, \dots, C_p)'$ and $W = (W^1, W^2, \dots, W^p)'$. It can be shown that $E[W^k] = 0$, thus $E[W] = 0$.

Multiply this equation by transpose matrix C^t

$$C^t Y = C^t C X + C^t W$$

Write $H = C^t C$, M is a $N \times N$ symmetric matrix.

If C is a full rank matrix (all columns are linearly independent), we can calculate the matrix inverse H^{-1} .

let

$$H^{-1} C^t Y = X + H^{-1} C^t W.$$

So we have:

$$X = H^{-1} C^t Y - H^{-1} C^t W$$

Now recall that $E[W|Y] = 0$, then taking the conditional expectation of this equation given Y , we obtain

$$E[X|Y] = H^{-1} C^t Y$$

This system of equations presents a combination linear between conditional probability distributions $[\mathbb{1}_{\{X=e_i\}}|Y], i = 1, 2, \dots, N$ and the observed characteristics (Y^1, Y^2, \dots, Y^p) , where the associated coefficients are obtained from the $H^{-1} C^t$.

2.2 Presentation the classifier model

The process defined below show that the hidden state can be estimated as a linear function of the observed vector with corresponding coefficient obtained from $H^{-1} C^t$.

We can use the equation presented below to define the learning algorithm. Firstly, we estimate the coefficients of matrix C_k , where $c_{ji}^k = P(Y = f_{jk} | X_t = e_i), i = 1, 2, \dots, N; j = 1, 2, \dots, m_k$. Therefore, we have a estimation of the parameters of the model the matrix C . Secondly, we determine the parameters of the classifier from the calculation of $H^{-1} C^t$. The above processes show

that the hidden state which represents the class label can be estimated as a linear function of the observed vector. Finally, the conditional expectation of X given the observation Y can be expressed as

$$P(X = e_i|Y) = E[\langle X, e_i \rangle | Y] = \sum_{k=1}^p \sum_{j=1}^{m_k} \beta_{ji}^k \langle Y^k, f_{jk} \rangle, \quad i = 1, 2, \dots, N$$

In summary, the classifier is now illustrated by these posterior probabilities. The observation y is assigned to the class that verify the rule:

$$P(X = e_i|Y = y) = \arg \max_{j=1}^N P(X = e_j|Y = y)$$

where the argmax operator returns the argument i that maximizes the expression $P(X = e_i|Y = y)$.

The parameters in our model are the probabilities $c_{ji}^k, i = 1, 2, \dots, N; j = 1, 2, \dots, m_k$ for $k = 1, 2, \dots, p$. We shall estimate these using the maximum-likelihood method given a sequence of observations $y_1 = (y_{11}, y_{12}, \dots, y_{1p}), \dots, y_n = (y_{n1}, y_{n2}, \dots, y_{np})$. The results formulas (presented in [8]) which maximize the function likelihood are

$$c_{ji}^k = \frac{\sum_{t=1}^n \langle Y_t, f_{jk} \rangle \langle X_t, e_i \rangle}{\sum_{t=1}^n \langle X_t, e_i \rangle}$$

where the sequence of X , as they are observed, is denote by $\{X_t, t = 1, 2, \dots, n\}$. Using the above formula and based on the concept of counting event occurrences (presented in [6]) we can give a method for estimation of the parameters as following:

$$\widehat{c}_{ji}(k) = \frac{\text{expected number of transitions from state } i \text{ and observing symbol } f_{jk}}{\text{expected number of transitions from state } i}$$

This expression will be used to estimate coefficients of matrix C and construct the HMM classifier.

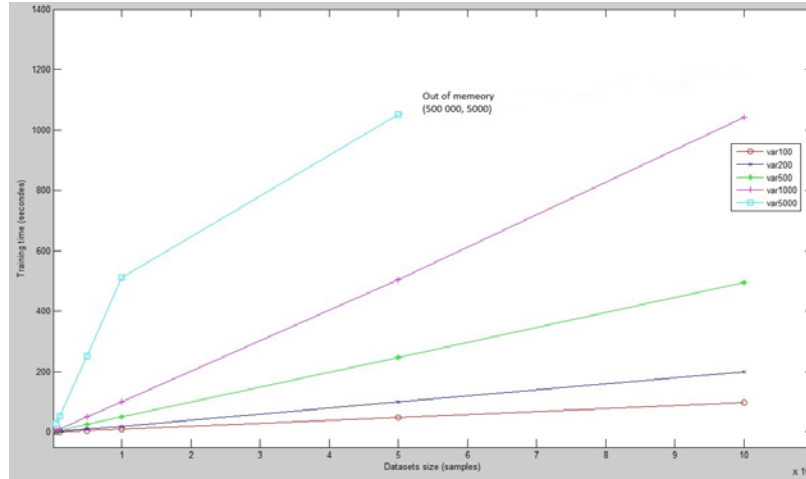


Figure 1: Measurements and performance of the rise in the scale of data sets for HMM.

2.3 The limits of HMM classifier

HMM has made great achievements and proved to be successfully useful in a variety of pattern recognition such as speech recognition, handwriting and engineering. However, HMM has a very limited capacity for recognizing complex patterns when being applied to deal with large scale datasets due to its intensity calculations and consume a lot of computation time for estimating parameters. It seems to be difficult to directly apply the below training process on a single PC to build the classifier because the application of HMM training algorithm spend more time to finish or even cannot be done. In the following, we show the limitations of applying the classifier by testing the speedup of training process on large scale data.

Experiments described in this work were performed on an ordinary computer, with a processor of 2.9GHz Intel CORE i7 7th Gen CPU and 8GB RAM memory, using Windows 10 operating system. Matlab, version 2015a, was used for modeling. Experiments results are obtained by varying the number of observations (n) and the number of observed characteristics (p).

We use the packages PRtools a Matlab toolbox for pattern recognition to simulate several data sets with different size. Five number of variables are chosen to determine data sets. For each number of variables six data sets are generated by a number of samples changed between 5000 and 100000. The scalability of number of samples and size of number of variables is limited and expensive. The large volume of data generated by PRtools it will be limited by its high size, due of the insufficient memory to load the data.

3 MapReduce-Based HMM classifier

Mapreduce is a simple programming tool applied for data processing. It is used to divide the input data in to different blocks. The main task of Mapreduce is to distribute the blocks of data to DataNodes in order to be exploited by the learning algorithm. In this section, we present a technique programming which is suitable for estimating HMM parameters and building classifier based on Mapreduce distributed environment of Matlab. Our methods is divided in three steps. There are described in the following part.

3.1 Implementation of MapReduce with HMM classifier : MR-HMM

Take a dataset including p independent variables $Y = (Y_1, Y_2, \dots, Y_p)$ and a target dependent variable X .

Firstly, we use datastore to process the data in small chunks that individually fit into memory. The large data used in training process is presented separately into many several blocks. Each block goes through a Mapper which formats the data to be processed. Suppose we have a dataset $D_{(n \times p)} = ((y_1, x_1), (y_2, x_2), \dots, (y_n, x_n))$ where $y_i = (y_{i1}, y_{i2} \dots, y_{ip})$.

$$D_{(n \times p)} = \begin{bmatrix} D^1 \\ D^2 \\ D^3 \\ D^4 \end{bmatrix} \quad \text{where : } D^i = D_{n_i \times p}, \quad n_i = \text{block size},$$

The dataset are decomposed to small blocks $D^i = \{(y^i, x^i)\}$ with the same size and delivered together to a Mapper.

Secondly, the technique is composed of a Map phase, which formats the data to be processed and performs a precursory calculation. Every map task takes as parameters block and apply the training algorithm. In fact, each map task receives the sub-datasets as input and compute partially the coefficients of matrix C^i correspond to the used block D^i . The results matrices C^i are associated with the key $\ll Key_i \gg$ and sent them directly to reduce without any treatment.

Table 1: Description statistique des deux bases de donnes.

Data set	sample (n)	feature (p)
MNIST	70 000	784
GISETTE	7000	5000

Thirty, the Reduce phase aggregates all of the results from the Map phase. There is a single reduce task. The inputs are the set of matrices C^i obtained from the map tasks. The matrix C of estimated coefficients for training task is constructed with the intermediate matrix C^i collected from map function. The single C is calculated for this time as follow:

$$C = C^1 + C^2 + C^3 + C^4$$

C is considered to C final delivered which can be used by the training algorithm to build the classifier.

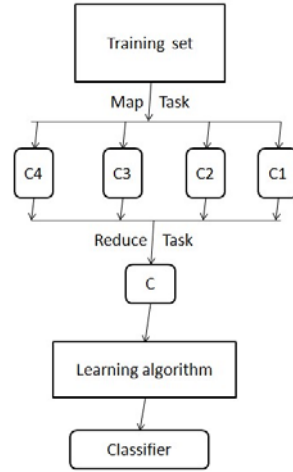


Figure 2: Training HMM classifier using Mapreduce

3.2 Credit Datasets Description

Two benchmark data sets are used in the experiment. The data sets are database of handwritten digits : MNIST and GISETTE. MNIST has a training set images of 60,000 examples, and a test set images of 10,000 examples. It is a subset of a larger set available from NIST. The labels values are 0 to 9. The task of GISETTE is to discriminate between to confusable handwritten digits: the four and the nine. This is a two-class classification problem with sparse continuous input variables. The data set was constructed from the MNIST data. Table shows description information about the four benchmark data sets.

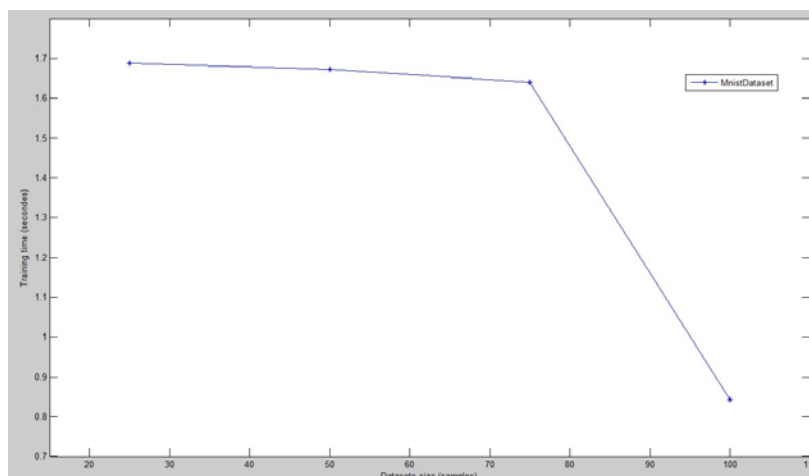


Figure 3: Performance measures of HMM with MapReduce for MNIST Data sets

3.3 Results and Analysis

In this section, we evaluate the performance of HMM-MapReduce on the two data sets described above. The experiments are carried out using the tool MapReduce of Matlab environment. The three steps presented in section 3.1 are implemented and tested. Thus, we conduct two experiments using the two datasets to compare the training performance of our contribution.

Figs 3-4 show plots training time associated to each chunk resulted from MapReduce implementation. As be shown from the two figures, the MNIST data sets are divided to four chunks and the GISETTE data sets are decomposed to three chunks. The number of chunks depend to the size of samples. The time of training process correspond to GISETTE data sets is high than the MNIST samples and depend heavily to the number of variables. Our model are build up with a total time 15.7970s for GISETTE and 5.8438s for MNIST. The experimental results prove the efficiency and scalability of the method over large data. We can conclude that the results presents a significant challenge to increase processing performance of our training HMM parameters.

4 Conclusion

In this paper, we propose a new method for building HMM classifier for large data sets based on Mapreduce architecture. The developed approach is able to extract the information from massive data and estimate the parameters of HMM. We can solve the problem of out memory provided by the scalability of data and accelerate the learning process. Unlike to increase the performance of the pc, we can use MapReduce environment to deal with the problem of limitation of memory and capacity to run the program in reasonable time. Experimental results show that HMM-classifier based on Mapreduce is able to train the learning algorithm over huge amount of data.

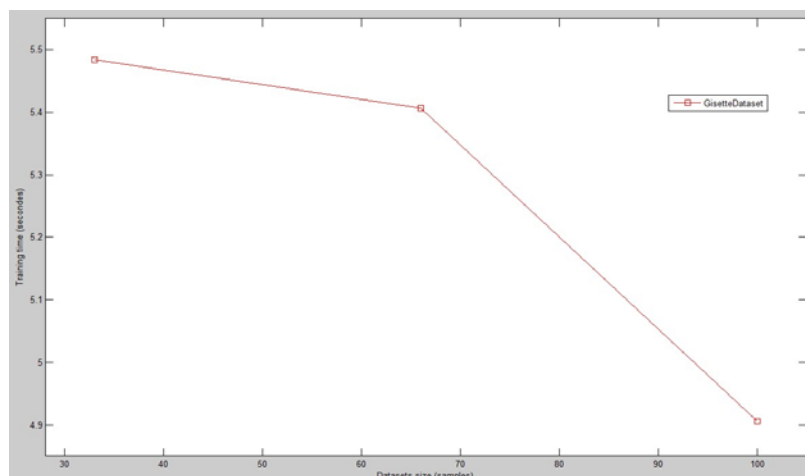


Figure 4: Performance measures of HMM with MapReduce for GIsETTE Data sets

References

- [1] Adjout Rehab. M. and Boufars. F. *Parallel Implementation of Multiple Linear Regression Algorithm Based on MapReduce* Proceedings of the 5th International Conference on Industrial Engineering Operations Management (IEOM 2015), 3-5 March 2015, Dubai, United Arab Emirates UAE. Pages 2493-2497.
- [2] Badreddine. B, Souad. B, Abdelhak. Z, Ismail. E, *Classification with Hidden Markov Model*, Applied Mathematical Sciences, Vol. 8, 2014, no 50,2483-2496.
- [3] Frank. E. T., G. Holmes. G., Kirkby. R.B. and Hall. M.A. *Racing committees for large datasets*, In Proceedings of the International Conference on Discovery Science, Berlin: Springer LNCS 2534, pp 153-164, 2002.
- [4] Marz. N., Warren. J. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, (2013).
- [5] Nick Street. W. and YongSeog Kim. *A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification*, In Proceedings of the 7th ACM SIGKDD inter. Conf. on Knowledge Discovery in Databases and Data Mining, pages 377-382, 2001.
- [6] Rabiner. L.R, *A tutorial on hidden markov models and selected applications in speech recognition*, Proceedings of the IEEE, vol. 77, pp. 257286, 1989.
- [7] Rajaraman. A., Leskovec. J. and Ullman. D. (2010). Mining of Massive Datasets, pp 4-7.
- [8] Robert. J. Elliot, and Alexei Filinkov : *A self tuning model for risk estimation*, Expert Systems with Applications 1692-1697.34, (2008).
- [9] Zhen Niu, Zelong Yin and Huayang Cui. *MapReduce-Based Bayesian Automatic Text Classifier Used in Digital Library*. presented at the Proceedings of the 6th International Symposium on Intelligence Computation and Applications, October 27-28, 121-126 (2012).

Performance Analysis of Time Series Forecasting of Ebola Casualties Using Machine Learning Algorithms.

Manish Kumar Pandey and Karthikeyan Subbiah

Department of Computer Science
Institute of Science, Banaras Hindu University
Varanasi-221005, India.

pandey.manish@live.com

Abstract. There is an emergent concern on our vigilance for controlling the spread of pandemics such as Ebola, Zika etc. Precise and trustworthy prediction incidences of these diseases are obligatory for the health authorities to guarantee the suitable action for the control of the outbreak. The dynamics of epidemic spread in large-scale populations is very complex. Huge data generated in the era of SMAC makes it more complex. Processing of this huge data is very important for effective descriptive, predictive, preventive and prescriptive analytics. Effective planning and response strategies must take these complicated interactions into account. In this paper, we have proposed the use of machine learning techniques for performance evaluation of time series forecasting of casualties in case of Ebola Outbreak. We have conducted experiments on ten different classifiers and selected the better performing random tree classifier for forecasting Ebola casualties. By experimenting without lag creation, we achieved the best results in the MAE of 5.39 %, RMSE value of 42.41 %, and Direction Accuracy of 90.95 %. Thus we can conclude that by using these models for forecasting epidemic spread and developing public health policies leads the health authorities to ensure the appropriate action for the control of the outbreak.

Keywords: *SMAC, Epidemic forecasting, Big data computational epidemiology, Time Series Forecasting, Random Tree*

1. Introduction

Globally infectious diseases are the major cause of human mortality. Moreover, just six infections are there which are deadliest- pneumonia, tuberculosis, diarrhoea, malaria, measles and HIV/AIDS. Key events related to the history of infectious diseases can be traced as far back as the Middle Ages (see Fig. 1 [1]).

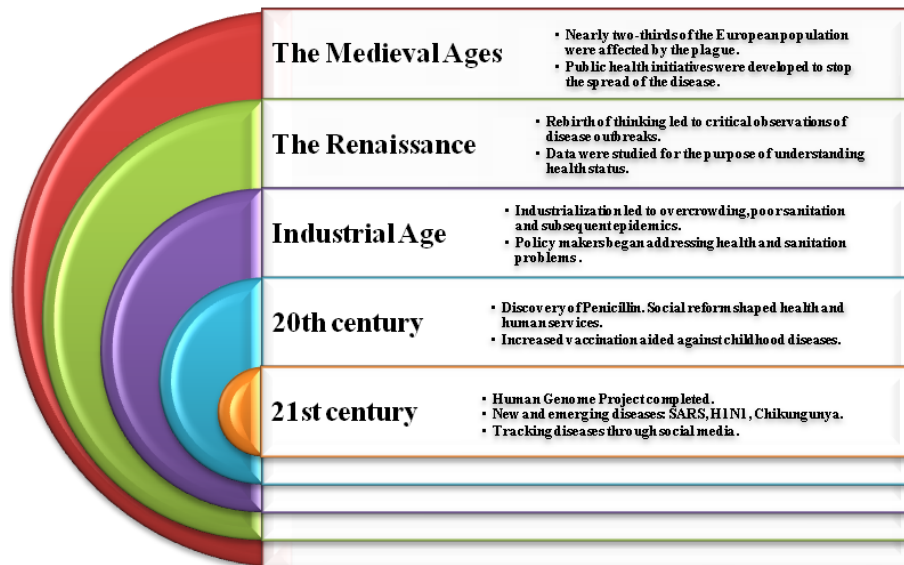


Fig.1 Disease Outbreaks from middle Ages to 21st Century

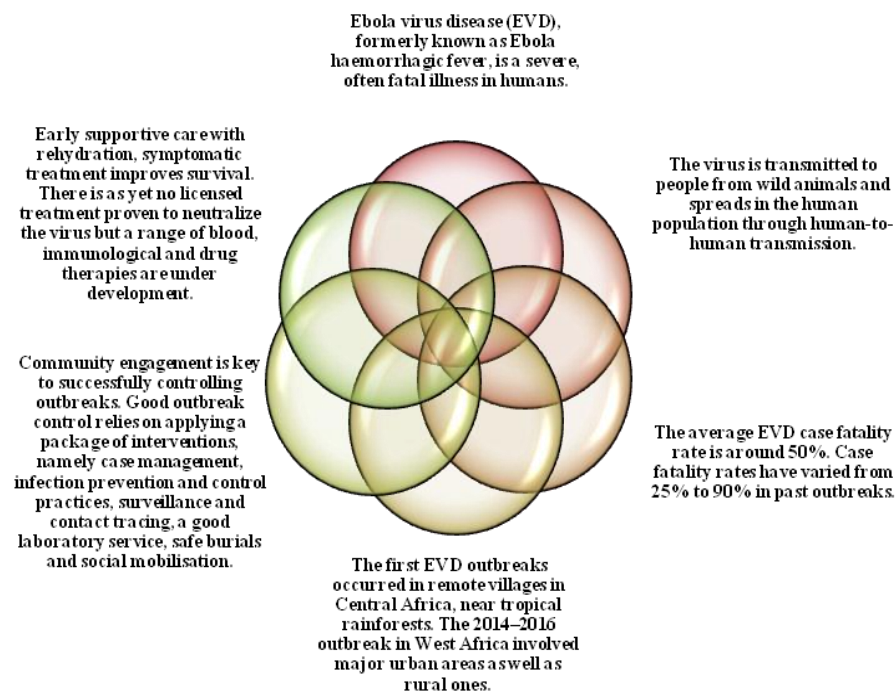


Fig. 2 Key Facts on Ebola Disease

Table 1 Chronology of previous Ebola virus disease outbreaks

Year	Country	Ebolavirus species	Cases	Deaths	Case fatality
2015	Italy	Zaire	1	0	0%
2014	DRC	Zaire	66	49	74%
2014	Spain	Zaire	1	0	0%
2014	UK	Zaire	1	0	0%
2014	USA	Zaire	4	1	25%
2014	Senegal	Zaire	1	0	0%
2014	Mali	Zaire	8	6	75%
2014	Nigeria	Zaire	20	8	40%
2014-2016	Sierra Leone	Zaire	14124*	3956*	28%
2014-2016	Liberia	Zaire	10675*	4809*	45%
2014-2016	Guinea	Zaire	3811*	2543*	67%
2012	Democratic Republic of Congo	Bundibugyo	57	29	51%
2012	Uganda	Sudan	7	4	57%
2012	Uganda	Sudan	24	17	71%
2011	Uganda	Sudan	1	1	100%
2008	Democratic Republic of Congo	Zaire	32	14	44%
2007	Uganda	Bundibugyo	149	37	25%
2007	Democratic Republic of Congo	Zaire	264	187	71%
2005	Congo	Zaire	12	10	83%
2004	Sudan	Sudan	17	7	41%
2003 (Nov-Dec)	Congo	Zaire	35	29	83%
2003 (Jan-Apr)	Congo	Zaire	143	128	90%
2001-2002	Congo	Zaire	59	44	75%
2001-2002	Gabon	Zaire	65	53	82%
2000	Uganda	Sudan	425	224	53%
1996	South Africa (ex-Gabon)	Zaire	1	1	100%
1996 (Jul-Dec)	Gabon	Zaire	60	45	75%
1996 (Jan-Apr)	Gabon	Zaire	31	21	68%
1995	Democratic Republic of Congo	Zaire	315	254	81%
1994	Côte d'Ivoire	Tai Forest	1	0	0%
1994	Gabon	Zaire	52	31	60%
1979	Sudan	Sudan	34	22	65%
1977	Democratic Republic of Congo	Zaire	1	1	100%
1976	Sudan	Sudan	284	151	53%
1976	Democratic Republic of Congo	Zaire	318	280	88%

The word epidemic was derived from the Greek words: epi (upon) and demos (people) meaning “upon people.” It is an event in a population, of cases of a sickness, particular health behaviour or other health-related events in a surplus of what would be normally possible. A pandemic is an epidemic that spans a large portion of the world, such as the H1N1 outbreak in 2009. In contrast, an endemic disease is one wherein new infections are constantly occurring in the population.

The 2014–2016 outbreaks in West Africa were the largest and most complex Ebola outbreak since the virus was first discovered in 1976. It has more confirmed cases and casualties in this outbreak than the rest combined. It has started from Guinea and spread between Sierra Leone and Liberia. Key facts and Chronology of previous Ebola virus outbreaks are given in Figure 2 and Table 1 [2].

The Study and its application of distribution and determinants of the events related to health across specified populations for description, prediction, prevention and prescription of health problems are defined as *Epidemiology* [3]. The Main concern of Epidemiologists is public health which includes the efficient analytics of descriptive public data and maintenance of its collection. They do it by exploring the spatial extent of the outbreak, progress chart of the disease, mode of controlling the disease, the origin of disease and how is it different than the previous outbreaks.

1.1 Big Data Computational Epidemiology

History of Epidemiology goes long back to 1760. In 1760; Daniel Bernoulli [4] has given the first model mathematically and established that inoculation could facilitate an increase in the life expectancy in France. A British physician, John Snow analyzed a cholera outbreak in London in 1854. He credited it to a supply of polluted water [4]. In the current era of SMAC (Social, Mobility, Analytics and Computing) [5,6] platforms, a huge amount of data is getting generated from social networking sites, real time streams of outbreaks etc. This huge data makes the computation in Epidemiology more complex. This calls for *Big Data Computational Epidemiology* which is a rising interdisciplinary field which makes use of computational models and big data for understanding and controlling the spatiotemporal transmission of disease throughout populations. Following are the reason for which big data computational epidemiology is the need of current era.

1. Mathematical models have become increasingly complex for which big data analytics tools are required.
2. The model representing the affected population creates a complex interaction network. These network models are real scenarios based which makes it more computational and data costly. As mentioned in [6, 7], the analysis of such data sets requires powerful computing resources and big data analytics tools
3. New methods of disease surveillance and detection are required for collection of huge data generated. Computational methods for data management, including methods to collect, store, clean, organize, search, fuse, and analyze data, are all important.

4. With the SMAC era, where everyone is connected with the internet, there is a growing demand for developing web-based tools that can be accessed by epidemiologists in a pervasive manner. This clearly indicates the role of big data epidemiology [8].

1.2 Big Data Analytics in Epidemiology

From Big Data computational epidemiology, four basic classes of problems arise based upon the network created that involves the places where the disease has spread.

Descriptive Analytics: This includes characterizing the outbreak size, duration of the epidemic, and other properties of epidemics. Actual visualization of the spread and other related features could be helpful in next step.

Predictive Analytics: This include problems of determining quantities, such as the number of infections over time, or the peak, and identifying the people who might be infected, given partial information of the outbreak until some time. Machine Learning techniques can be used for efficiently forecasting the spread based on the output of the previous step.

Preventive Analytics: As discussed earlier, the networked SIR model is determined by the network, initial conditions, and epidemic model. In general, we may have partial information about some of these components; e.g., edges of the graph might not be completely known, or parameters of disease spread are unknown. With the help of forecasting result obtained in previous step, we can put a check on the spread.

Prescriptive Analytics: This includes problems of controlling the spread of epidemics, e.g., by vaccination or quarantining, correspond to making changes in the node functions or removing edges so that the system converges to configurations with few infections. We could use the result of preventive analytics for an efficient delivery model of production, transportation and distribution of vaccines, doctors and other resources.

We will discuss about predictive analytics in this paper.

1.3 Predictive Analytics

Predictive Analytics includes problems of determining quantities, such as the number of casualties, no of suspected cases, no of infections over time etc. Epidemic Forecasting in terms of casualties, location etc is an important rising topic in big data computational Epidemiology. It involves collecting and combining data from nontraditional sources like Social Media, Wikipedia, and World Health Organization's surveillance systems and processing them with statistical models and machine

learning techniques to now cast and forecast the occurrence of diseases in the host population. Nsoesie et al. [9] reviewed methods for influenza forecasting proposed during previous influenza outbreaks and those evaluated in hindsight. Nishiura [10] have given a discrete time stochastic model and applied as a case study to the weekly incidence of pandemic influenza in Japan. Ohkusa et al. [11] have demonstrated real-time estimation and prediction of the entire course of a pandemic of ILI (influenza-like illness) in Japan. Hall et al. [12] have predicted spread of the H5N1 influenza virus in birds by fitting a mass-action epidemic model to the surveillance data by standard regression analysis. Tizzoni et al. [13] have proposed Global Epidemic and Mobility Model to generate stochastic simulations of epidemic spread worldwide using a Monte Carlo Maximum Likelihood analysis. Shaman et al. [14-16] have used Bayesian ensemble methods to develop surprisingly high-quality forecasts for flu prevalence in US regions. Chakraborty et al. [17] has analyzed the generation of robust quantitative predictions about temporal trends of flu activity, using several surrogate data sources for 15 Latin American countries.

2 Materials and Methods

2.1 Data Set

We have selected Ebola outbreak data from duration 29-08-2014 to 29-12-2015 that is available in the WHO sitrep [18] which provide updated data for countries with an active Ebola outbreak. The Total number of instances are 4112 that contains details about the cumulative no of confirmed Ebola cases and cumulative no of confirmed Ebola Deaths. Fig 3 describes the statistics of the dataset.

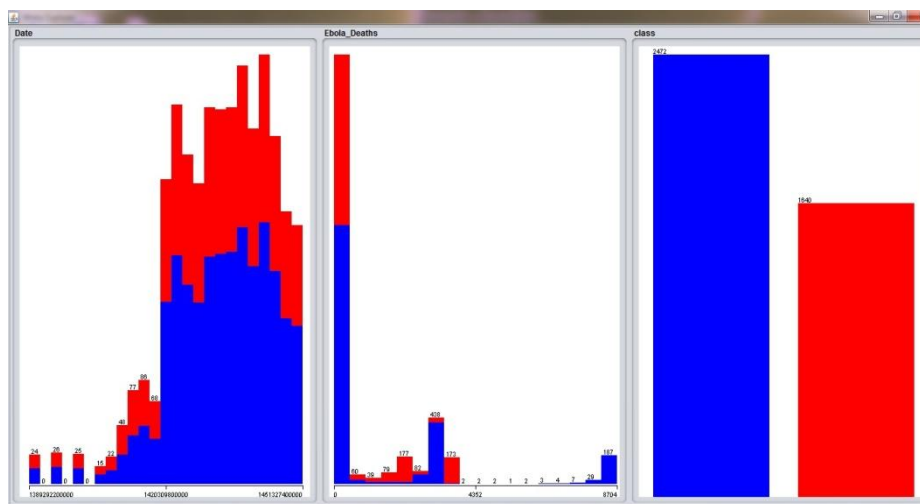


Fig.3.Graphical display of distribution of various attributes between the two classes as Cumulative Confirmed Ebola Cases and Cumulative Confirmed Ebola Deaths

2.2 Proposed Methodologies

Experiments are conducted for time series forecasting using 10 different machine learning algorithms, which are applied for prediction task. A brief description about the best performing algorithm is given in this study under prediction protocol section. The flow diagram of the proposed methodology can be depicted from the Fig 4.

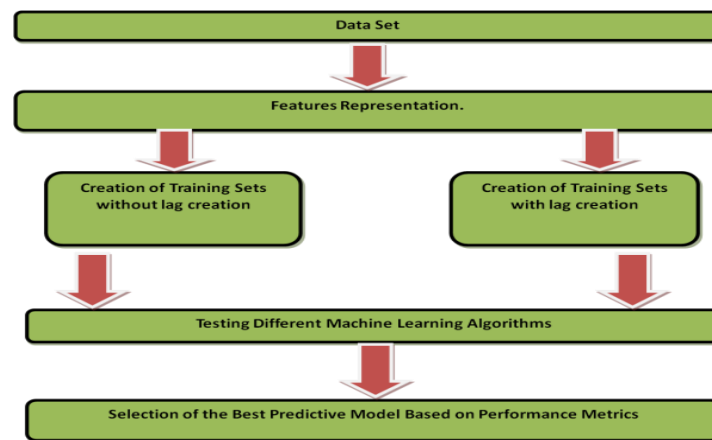


Fig. 4 Flow diagram of the Proposed Methodology

2.3 Prediction Protocol

Random tree was found to be better performing algorithm among all machine learning algorithms for the proposed problem prediction.

2.4 Performance Evaluation Metrics

The relative performance of time series analysis and forecasting through different machine learning algorithms is evaluated by using following three metrics.

Mean absolute error (MAE): The MAE evaluates the standard amount of the errors in a set of forecasts, without taking into account their direction. It gives the accuracy for continuous variables. In other words, it is the average of the total values of the differences of forecast and the matching observation. The MAE is a linear score which indicates that every individual difference is weighted equally in the average.

$$MAE = \frac{\sum |Predicted - Actual|}{N}$$

Root mean squared error (RMSE): It is a rule based on quadratic score that measures the average magnitude of the error. In words, the difference of forecast and matching observations are squared each and then average is calculated over the sample. Lastly, the square root of the average is noted. RMSE gives a relatively high weight to large errors because the errors are squared before averaging it. Thus RMSE is best used for cases where large errors are undesirable.

$$RMSE = \frac{\sqrt{\sum (Predicted - Actual)^2}}{N}$$

Direction Accuracy (DA): It gives the percentage of accurately predicted positive and negative examples with below formula

$$DA = \frac{Count(sign(actual_current - actual_previous))}{N}$$

The relative measures give an indication of how well the forecaster's predictions are doing compared to just using the last known target value as the prediction. They are expressed as a percentage and lower values (not Direction accuracy) indicates that the forecasted values are better predictions than just using the last known target values.

The open source Java based machine learning platform WEKA [19] was used to perform all the experiments in this study.

3 Result and Discussion

We experimented with ten different algorithms, namely: (1) Linear Regression, (2) Multilayer Perceptron, (3) Support Vector Machine for Regression, (4) Ensemble Selection, (5) Bagging with Reptree, (6) Random tree, (7) Random Forest (8) Reptree, (9) Random tree and (10) Random Forest on the training data and the values of different performance metrics for these algorithms are given in Table 2.

Table 1. Algorithms used in the proposed work

Linear-Regression	<i>Class for using linear regression for prediction. Uses the Akaike criterion for model selection, and is able to deal with weighted instances.</i>
--------------------------	--

MLP	<i>A Classifier that uses backpropagation to classify instances.</i>
SMOReg	<i>SMOreg implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer. The most popular algorithm (RegSMOImproved) is due to Shevade, Keerthi et al and this is the default RegOptimizer.[20,21]</i>
Ensemble-Selection	<i>Combines several classifiers using the ensemble selection method.[22]</i>
Bagging	<i>Class for bagging a classifier to reduce variance. Can do classification and regression depending on the base learner. [23]</i>
Random-Forest	<i>Class for constructing a forest of random trees. [24]</i>
RepTree	<i>Fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5).</i>
Random Tree	<i>Class for constructing a tree that considers K randomly chosen attributes at each node. Performs no pruning. Also has an option to allow estimation of class probabilities (or target mean in the regression case) based on a hold-out set (backfitting).</i>

Table 2.Performance metrics of time series forecasting using machine learning algorithms

Machine Learning Algorithms						
	Without Lag Creation			With Lag Creation		
	Mean Absolute Error(MAE)	Root Mean Square Error(RMSE)	Direction Accuracy	Mean Absolute Error(MAE)	Root Mean Square Error(RMSE)	Direction Accuracy
Linear Regression	473.54	685.26	59.18	518.97	684.44	58.51
MLP	164.78	317.63	65.94	145.45	300.95	71.74
SMOReg	300.61	1073.22	66.37	424.76	769.86	60.82
Bagging-Reptree	125.58	300.12	74.18	120.15	262.64	86.27
Bagging-Random Tree	55.92	181.56	84.20	55.60	159.66	82.90
Bagging-Random	71.27	208.12	80.88	70.65	179.25	80.72

Forest						
Ensemble Selection-Forward	120.79	297.20	67.79	130.47	299.72	74.54
RepTree	191.80	397.461	55.52	145.48	339.77	66.74
Random Tree	5.39	42.41	90.95	6.5824	53.22	85.90
Random Forest	54.53	131.43	81.16	47.86	128.31	82.15

We have used both, with Lag formation and without Lag formation to measure the performance of various machine learning algorithms. It was also observed that the performance of Random Tree was superior to the rest of the 9 other machine learning algorithms in terms of the different evaluation parameters. It is clear from the table 2 and figures 5-8 that Random tree has performed superior in both the cases, i.e. with lag creation and without lag creation. The Lag creation allows the user to control and manipulate how lagged variables are created. Lagging is the main method through which the association of past and current values of a set could be encapsulated by propositional learning algorithms. They generate a "window" or "snapshot" above a time period.

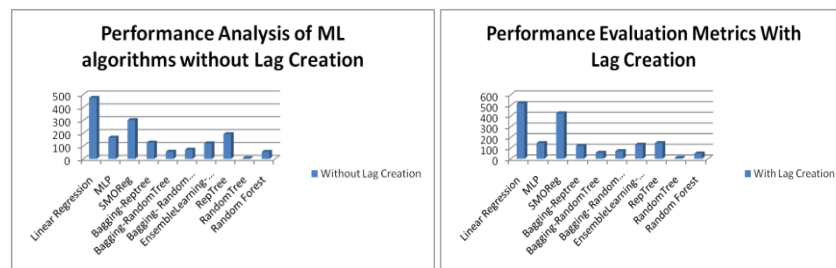


Fig. 5 Mean Absolute Error with/without Lag Creation

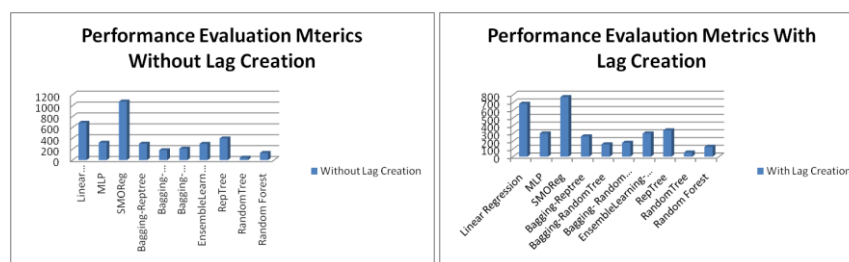


Fig. 6Root Mean Squared Error with/without Lag Creation

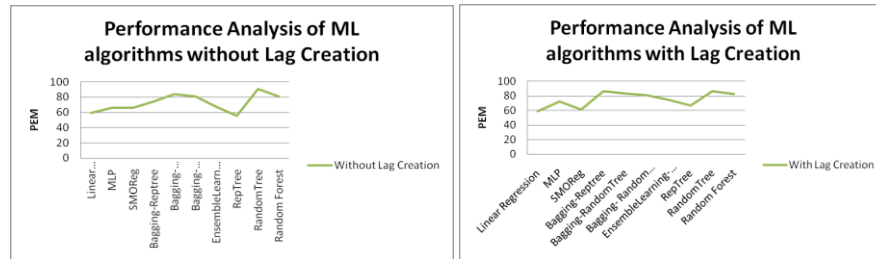


Fig. 7 Direction Accuracy with/without Lag Creation

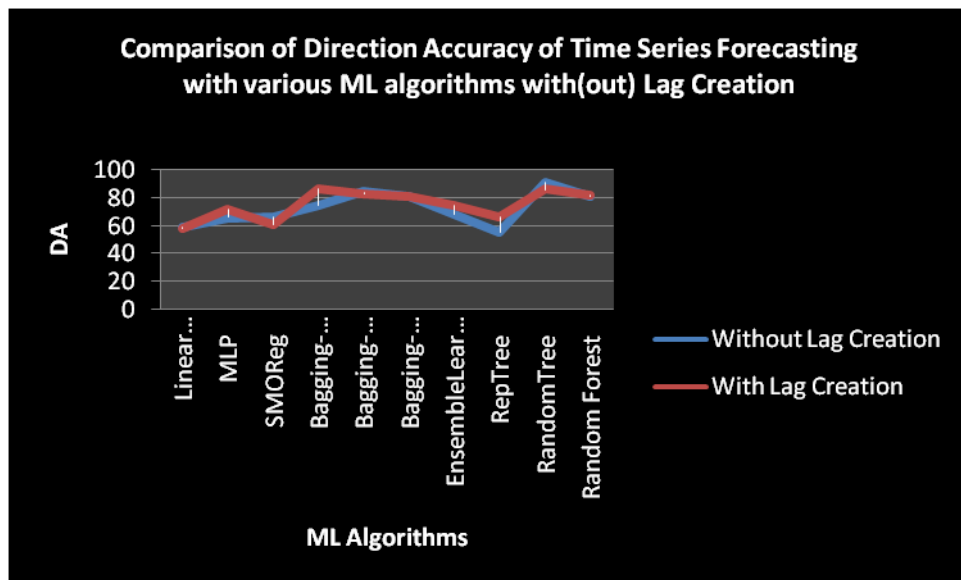


Fig. 8 Comparison of Direction Accuracy with/without Lag Creation

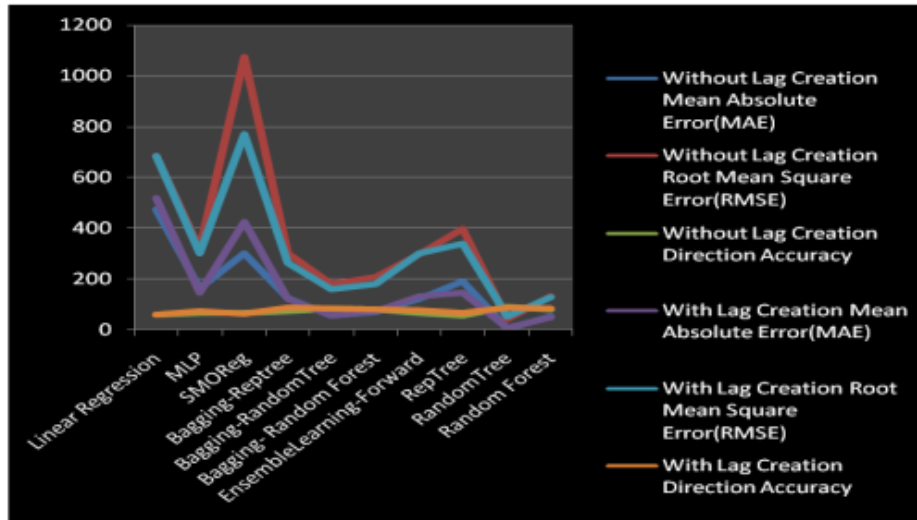


Fig. 9 Comparison of MAE, RMSE and DA with/without Lag Creation

From Fig. 5, 6, 7, it is clear that Random tree has performed in best manner as compared to other algorithms. For the experiments conducted on data set, the value of MAE and RMSE is lowest in without lag creation category as compared to that of with lag creation. Direction Accuracy of Random tree is best in without lag creation category with a value of 90.9545 % while with lag creation Random tree has performed with a value of 85.9014 %. From Fig. 8, it is clear that random tree has best direction Accuracy of 90.9545 % overall in both with and without lag creation category.

Fig. 9 shows a comprehensive comparison of MAE, RMSE and DA in data set with/without lag creation.

4 Conclusion

Big data computational epidemiology is a new and exciting multidisciplinary area with significant challenges of large data and high-performance computing. There are many factors that can affect in achieving the true performance of time series forecasting. Predictive analytics of time series data in epidemiology include problems of determining quantities, such as the number of casualties, no of infections over time, or the peak, and identifying the people who might be infected. In this paper, we have applied Machine Learning techniques in time series forecasting for performance evaluation. With this result, we could say that performance of time series forecasting could be improved with the help of machine learning algorithms. Forecasting of casualties could help health officials in preparing themselves to encounter this outbreak,

supply of medicines, food supply, doctors etc to the location where prediction of casualties are more.

References

1. Saumyadipta Pyne, Anile Kumar S. Vullikanti, Madhav V. Marathe, Big Data Applications in Health Sciences and Epidemiology, Handbook of Statistics, Volume 33, 2015, Pages 171-202, ISSN 0169-7161
2. <http://www.who.int/mediacentre/factsheets/fs103/en/>
3. Last, J., 2001. A Dictionary of Epidemiology, fourth ed. Oxford University Press, New York.
4. Brauer, F., van den Driessche, P., Wu, J. (Eds.), 2008. Mathematical Epidemiology. Lecture Notes in Mathematics 1945. Springer Verlag, Berlin, Heidelberg, New York.
5. <http://www.idc.com/research/Predictions13/downloadable/238044.pdf>
6. <http://www.gartner.com/technology/research/nexus-of-forces/>
7. Bisset, K., Chen, J., Feng, X., Vullikanti, A. and Marathe, M. 2009a. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of 23rd ACM International Conference on Supercomputing (ICS'09). ACM Press, New York.
8. Salathé, M., et al., 2012. Digital epidemiology. PLoS Comput. Biol. 8 (7), e1002616
9. Nsoesie, E.O., Brownstein, J.S., Ramakrishnan, N., Marathe, M., 2013. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza Other Respir. Viruses 8 (3), 309–316.
10. Nishiura, H., 2011. Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (H1N1-2009). BioMed. Eng. Online 10 (1), 15
11. Ohkusa, Y., Sugawara, T., Taniguchi, K., Okabe, N., 2011. Real-time estimation and prediction for pandemic A/H1N1(2009) in Japan. J. Infect. Chemother. 17 (4), 468–472.
12. Hall, I.M., Gani, R., Hughes, H.E., Leach, S., 2007. Real-time epidemic forecasting for pandemic influenza. Epidemiol. Infect. 135 (3), 372–385.
13. Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J., Balcan, D., Goncalves, B., Perra, N., Colizza, V., Vespignani, A., 2012. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. BMC Med. 10 (1), 165. ISSN 1741-7015.
14. Shaman, J., Karspeck, A., 2012. Forecasting seasonal outbreaks of influenza. Proc. Natl. Acad.Sci. U.S.A. 109 (50), 20425–20430.
15. Shaman, J., Goldstein, E., Lipsitch, M., 2010a. Absolute humidity and pandemic versus epidemic influenza. Am. J. Epidemiol. 173 (2), 127–135.
16. Shaman, J., Pitzer, V.E., Viboud, C., Grenfell, B.T., Lipsitch, M., 2010b. Absolute humidity and the seasonal onset of influenza in the continental United States.. PLoS Biol. 8 (2), e1000316.
17. Chakraborty, P., Khadivi, P., Lewis, B., Mahendiran, A., Chen, J., Butler, P., Nsoesie, E.O., Mekar, S.R., Brownstein, J.S., Marathe, M.V., Ramakrishnan, N., 2014b. Forecasting a moving target: ensemble models for ILI case count predictions. In: Proceedings of the 2014 SIAM International Conference on Data Mining, 28 April 2014, pp. 262–270
18. <https://data.humdata.org/dataset/ebola-cases-2014>
19. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10-18, 2009
20. S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy: Improvements to the SMO Algorithm for SVM Regression. In: IEEE Transactions on Neural Networks, 1999

21. A.J. Smola, B. Schoelkopf (1998). A tutorial on support vector regression
22. Caruana, Rich, Niculescu, Alex, Crew, Geoff, and Ksikes, Alex, Ensemble Selection from Libraries of Models, The International Conference on Machine Learning (ICML'04), 2004.
23. Leo Breiman (1996). Bagging predictors. *Machine Learning*. 24(2):123-140
24. Leo Breiman (2001). Random Forests. *Machine Learning*. 45(1):5-32.

Hidden Markov Models for monitoring Circadian Rhythmicity in Telemetric Activity Data

Qi Huang, Dwayne Cohen, Sandra Koarzynski, Xiao-Mei Li,
Pasquale Innominato, Francis Lévi, and
Bärbel Finkenstädt*

Department of Statistics, University of Warwick,
Medical School, University of Warwick,
Coventry CV4 7AL, United Kingdom
<http://www2.warwick.ac.uk/fac/sci/statistics/>

Abstract. Wearable computing devices allow collection of densely sampled real-time information on movement enabling researchers and medical experts to obtain objective and non-obtrusive records of actual activity of a subject in the real world over many days. Our interest here is motivated by the use of activity data for evaluating and monitoring the endogenous circadian rhythmicity of subjects for research in chronobiology and chronotherapeutic healthcare. In order to translate the information from such high-volume data arising we propose the use of a Markov modeling approach which (a) naturally captures the notable square wave form observed in activity data along with heterogeneous ultradian variances over the circadian cycle of human activity, (b) solves the problem of thresholding activity into different states in a probabilistic way while respecting time dependence and (c) delivers parameter estimates, in particular probabilities of transitions between rest and activity, that are interpretable, irrespective of the model of measuring device, and important to circadian research.

Keywords: Hidden Markov models, Accelerometer data, Circadian Rhythm, Sleep-Wake cycle

1 Introduction

Questions of interest regarding the research of sleep-wake cycles in humans and mammals are commonly studied by measuring activity through gross motor movement where accelerometers have become a feasible and affordable way to obtain objective non-obtrusive recordings of rest-activity rhythms of free living individuals over many days [1–3]. Accelerometers measure the acceleration of the part of the body to which they are attached, often as part of a small communicative wearable device. The signal is preprocessed by the device to obtain physical activity (PA) counts accumulated over a specified time interval, called *epoch*. Time series PA data from such monitoring devices are subject to circadian rhythms and are of interest to the circadian research community. Ac-

* Corresponding author

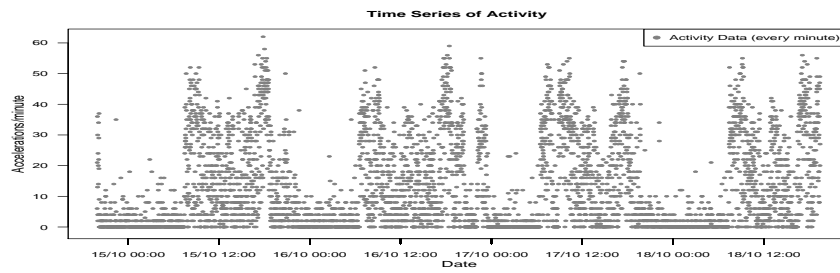


Fig. 1. Example of raw accelerometer data: Activity counts recorded per minute over 4 days with Move3 (Movisens GmbH, Germany) sensor with inbuilt accelerometer ADXL345 (Analog Devices, MA, USA) fixed to the chest of a healthy individual

tivity data can now be collected at short epoch lengths, such as every minute or every 15 seconds, over many days. The sensor used in our study (Move3, Movisens GmbH, Germany) is fixed to the chest and contains a triaxial accelerator model (ADXL345, Analog Devices, MA, USA). The device produces activity counts defined as the number of times an accelerometer waveform computed by the device, according to specifications of the frequency and filters that are specific to the manufacturer, crosses zero over the specified epoch length of 1 minute. Figure 1 gives an example of PA counts recorded every minute for a healthy individual over 4 days. Translating information from such high volume and complex data into interpretable and useful statistics is a challenging task, in particular if the aim is to perform long term monitoring of an individual. Apart from visually inspecting time plots the data are generally analyzed by deriving statistics, termed 'nonparametric variables' [4, 5], to quantify characteristics of interest to clinicians, sleep researchers and chronobiologists. These are generally focused around the *intradaily variability*, which measures the fragmentation of the rhythm, and *interdaily variability* which quantifies the entrainment to the 24 h light/dark cycle. An R-package to compute these alongside other statistics such as relative amplitude of activity, average activity values of the ten hours with maximal activity and the 5 hours with least activity, is provided by [6]. Evidence exists in the literature [7, 8] that the *intradaily variability* is a particularly important variable that is correlated with decreased sleep quality and cognitive functions as patients with Alzheimer's disease were found to have higher *intradaily variability* values [9]. Furthermore, in a clinical context, a series of studies [10–13] found that the *dichotomy index* $I < O$, which reports the percentage of activity epochs when in-bed, whose values were lower than the median level of activity when out-of-bed, was the most relevant statistic in predicting survival rates in cancer patients.

While there exist a number of nonparametric statistics to quantify the (mis)-timing of sleep-wake rhythms, and novel ones continue to be proposed [14], it remains an open task to quantify their variability and compute confidence in-

tervals. This will in particular be important if they are used in assisting with the decision making process of a health expert about an individual's therapy. Furthermore, most nonparametric statistics discussed above rely on being able to mark the beginning and end of prolonged rest periods. While different algorithms may be devised that identify individual-specific threshold values to classify between the states of rest and activity, this cannot always be determined unequivocally and requires hand-tuning, in particular if the subject's circadian rhythm is misaligned such as for shift workers.

More complex time series analysis approaches have been proposed, including spectral analysis and harmonic regression [15] or functional smoothing based on splines [16] applied to hourly PA recordings. Fourier methods are used to extract further parameters, namely acrophase, amplitude and period, that are typically of interest to studies of circadian rhythmicity. Spectral estimation using the methods proposed in [17] confirms that the activity data for healthy individuals usually exhibit a strong 24h periodicity as can be expected due the entrained endogenous circadian rhythmicity endorsed by the timing of the work and social environment. Although spectral analysis is well able to extract the circadian period, smooth functional forms, such as harmonic functions or splines, are not ideal for modelling the abrupt appearance of the transitions between the active and inactive states and will not detect short bouts of transitions caused, for example, by daytime naps or active interruptions at night. The data also show time changing variances in that PA values during the day show a markedly larger variability than over the prolonged rest period. Here we propose the use of hidden Markov models (HMMs) which provide the necessary tools to model the features observed in the data, and deliver estimated parameters that can be used to quantify the individual's sleep-wake behavior.

2 Model and Inference

Let $Y^{(T)} = \{Y_1, \dots, Y_{t-1}, Y_t, \dots, Y_T\}$ denote the observations on activity where $t \in \{1, \dots, T\}$ and T is the sample size. Let $S_t \in \{1, \dots, m\}$ denote the unobserved activity state at time t . The notation $P(\cdot)$ stands for the probability mass function or density function, whichever appropriate. We shall use the short notation for arbitrary X : $X^{(t)} = \{X_1, \dots, X_t\}$. The probabilistic structure of a HMM is represented a conditional independence graph which is a special case of a directed acyclic graph (DAG), and is based on the following two assumptions:

- (A1) The sequence of states S_t is a Markov chain satisfying the Markov property:
 $P(S_t | S^{(t-1)}) = P(S_t | S_{t-1})$,
- (A2) Conditionally on S_t , the Y_t 's are independent and Y_t depends on S_t only:
 $P(Y_t | S^{(t)}, Y^{(t-1)}) = P(Y_t | S_t)$.

It is straightforward to see that the joint distribution of the observations and the hidden states of the DAG is

$$P(Y^{(T)}, S^{(T)}) = P(S_1) \prod_{t=2}^T P(S_t | S_{t-1}) \prod_{t=1}^T P(Y_t | S_t), \quad (1)$$

from which the data likelihood can be obtained by summing over the possible combination of states (see, for example, [18])

$$\begin{aligned} P(Y^{(T)}) &= \sum_{s_1, \dots, s_T=1}^m P(S_1) \prod_{t=2}^T P(S_t|S_{t-1}) \prod_{t=1}^T P(Y_t|S_t) \\ &= \boldsymbol{\delta} \mathbf{P}(Y_1|S_1) \mathbf{\Gamma} \mathbf{P}(Y_2|S_2) \mathbf{\Gamma} \dots \mathbf{\Gamma} \mathbf{P}(Y_T|S_T) \mathbf{1}', \end{aligned} \quad (2)$$

where $\mathbf{P}(Y_t|S_t) \in \mathbb{R}^{m \times m}$ is the conditional probability matrix with j -th diagonal entries $\mathbf{P}(Y_t|S_t)_{j,j} = P(Y_t|S_t = j)$, $\mathbf{\Gamma} \in \mathbb{R}^{m \times m}$ is the Markov chain transition matrix with elements $\mathbf{\Gamma}_{j,k} = P(S = k|S = j)$, $\boldsymbol{\delta} \in \mathbb{R}^{1 \times m}$ is the initial state distribution and $\mathbf{1}' \in \mathbb{R}^{m \times 1}$ is a vector of ones. The HMM is hence parametrized by the non-zero entries in $\boldsymbol{\delta}$, $\mathbf{P}(Y_t|S_t)$ and $\mathbf{\Gamma}$. We shall use $\boldsymbol{\vartheta}$ to denote the vector of those unknown parameters. Given the output observation sequence $Y^{(T)}$, the maximum likelihood estimator of $\boldsymbol{\vartheta}$ can be efficiently found, either through direct maximization or based on an expectation maximization (EM) algorithm, called the Baum-Welch Algorithm [18], for which closed form expressions and computationally fast steps exist when the observational distribution $\mathbf{P}(Y_t|S_t)$ is Gaussian. However, caution is necessary as the likelihood may have local maxima and it is advised to test different starting values for the parameters.

3 Application to Activity Data

Data and Data Pre-processing We shall show results of fitting HMMs to PA count data recorded by the Move 3 sensor for 28 healthy individuals over the time of 4 – 5 days. In addition to activity, this device also provides minute recordings of the 3D position, via the angles with respect to three orthogonal axes set by the device and 5-min recordings of the skin temperature. Missing values occur as the individuals remove the device to avoid contact with water - typically this happens once a day for around 20 minutes. The missing values can be marked retrospectively by noting that the contemporaneous temperature records show a sudden decrease towards room temperature. Generally the missing data ratio is around 1 – 3% for the healthy individuals in this data set. The estimation algorithm can be altered in a straightforward way by propagating the transition matrix corresponding to the last time point preceding the missing values [19].

Collected over many days the data are of considerable size and it is desirable to be able to apply computationally efficient methodologies. In this study, we will assume Gaussianity of the observational densities of the square root transformed 5-min mean aggregated PA count data which corresponds to the assumption of a mixture of non-central Chi-square distributions at the original scale of the PA counts which can account for the non-negative domain and positive skew.

Number of States An obvious question is how many states m the model should have. We have estimated 2 – 6 state models for all 28 individuals, using information criteria such as Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) for model comparison [18]. The AIC tended to prefer models with 3 – 6 states while the BIC suggested more parsimonious models with

2–4 states. The estimation algorithm is sensitive to increasing model complexity and inconsistencies in interpretation between different individuals arise when 4 or more states are considered. For example, when $m = 4$, the algorithm may identify one inactive state during the night and three active states during the day for some individuals while for others it identifies two states at night and two for the prolonged active period. We concluded that models with $m = 3$ states are preferred for our purposes in that we did not encounter any convergence issues with the algorithm and the resulting model is simple yet detailed enough for the purpose of long term monitoring and consistently interpretable across many individuals.

Parameter Estimates The parameter estimates contain useful interpretable information about the individual's sleep-wake behaviour. We shall discuss typical results of fitting a HMM for two example subjects, A and B, say, with 3 states that can be interpreted as inactive (IA) for $S_t = 1$, moderately active (MA) for $S_t = 2$ and highly active (HA) for $S_t = 3$, where $S_t = j; j = 1, 2, 3$ also denotes the entry number of the corresponding state in all vectors and matrices. The estimated model parameters for subject A are as follows: the transition probabilities are

$$\hat{\Gamma} = \begin{pmatrix} 0.980 & 0.007 & 0.013 \\ 0.025 & 0.907 & 0.069 \\ 0.000 & 0.116 & 0.884 \end{pmatrix}$$

with conditional observation densities for IA state: $Y_t|(S_t = 1) \sim N(0.92, 0.68^2)$, for MA state: $Y_t|(S_t = 2) \sim N(3.1, 1.11^2)$ and $Y_t|(S_t = 3) \sim N(5.37, 0.74^2)$ for HA state. The initial state distribution is $\hat{\delta} = (0, 1, 0)$, i. e. the initial state is estimated to be MA.

For subject B the estimated model parameters are:

$$\hat{\Gamma} = \begin{pmatrix} 0.945 & 0.055 & 0.000 \\ 0.065 & 0.859 & 0.076 \\ 0.000 & 0.140 & 0.860 \end{pmatrix}$$

with conditional observation densities for IA state: $Y_t|(S_t = 1) \sim N(1.25, 1.26^2)$, for MA state $Y_t|(S_t = 2) \sim N(6.98, 2.23^2)$ and for HA state $Y_t|(S_t = 3) \sim N(12.06, 0.92^2)$. The initial state distribution is $\hat{\delta} = (0, 0, 1)$, i. e. the initial state is estimated to be HA.

The results for the transition probabilities in $\hat{\Gamma}$ for all 28 individuals are plotted in Figure 2.

As can be expected the diagonal elements of the transition matrix suggest a high chance of staying in the current state and this is highest for the IA state, as estimated by $\hat{\Gamma}_{1,1}$, due to the prolonged period of rest at night. The slightly lower values for $\hat{\Gamma}_{2,2}$ and $\hat{\Gamma}_{3,3}$, together with the elevated off-diagonal estimated probabilities $\hat{\Gamma}_{2,3}$ and $\hat{\Gamma}_{3,2}$, indicate that there is a higher chance of switching between the two active states. In fact, it is these transitions that

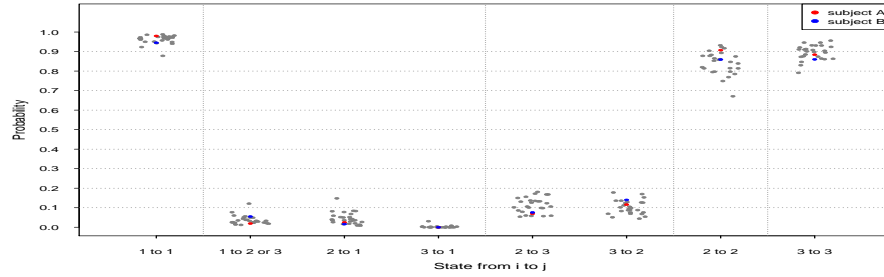


Fig. 2. Estimated transition probabilities for 28 healthy individuals. The integers 1, 2 and 3 represent the inactive (IA), medium active (MA) and highly active (HA) states, respectively.

account for the high variability observed in the data during the day due to the fact that people undertake a variety of physical actions. The transition from activity to rest is found to happen almost exclusively via the MA state as $\hat{\Gamma}_{3,1}$ is estimated to be zero, or very close to zero, for all individuals in the sample. A transition of interest is from IA to either of the active states with probability $(\hat{\Gamma}_{1,2} + \hat{\Gamma}_{1,3})$, which is equal to $(1 - \hat{\Gamma}_{1,1})$ where high values indicates many activity episodes where the person is likely to have interrupted sleep. For example, it is estimated to be 0.02 for subject A and 0.055 for subject B, which seems to indicate that subject B has experienced about twice as many sleep interruptions as A during the study time. The estimated transition probability from IA to any active state hence provides an alternative estimator of the nonparametric *intra-daily variability* statistic which is suggested and estimated for hourly PA count data in [4, 7–9]. We note that there is a positive correlation (Figure 3, top left) between $(\hat{\Gamma}_{1,2} + \hat{\Gamma}_{1,3})$ and $\hat{\Gamma}_{2,1}$, i.e. the transitions into and out of the IA state, as subjects who often interrupt their IA state will also more often need to get back to rest. Also, both transition probabilities are positively correlated with the estimated average probability of being active at night (Figure 3, top right and bottom right) which was computed for our sample by taking the average of the estimated $P(S_t = j | Y^{(T)})$ for $j = 2, 3$ during the prolonged IA periods. As can be expected there is a negative correlation between the estimated dichotomy index $I < O$ and the average probability of being active at night (Figure 3, bottom left). These three nonparametric statistics require that the PA data be partitioned into prolonged IA and active periods. Such a classification is not ambiguous and may require substantial hand-tuning, in particular for PA data from less rhythmic individuals, while the estimation of our HMM parameters is unequivocal and does not require the use of such classification algorithms.

The estimated conditional observation densities confirm our visual impression that the IA state has a lower variance. The MA state usually is characterized by a higher variance also in comparison to the HA state. The latter may be due to the dampening effect of the square root transformation. Figure 4 hence plots

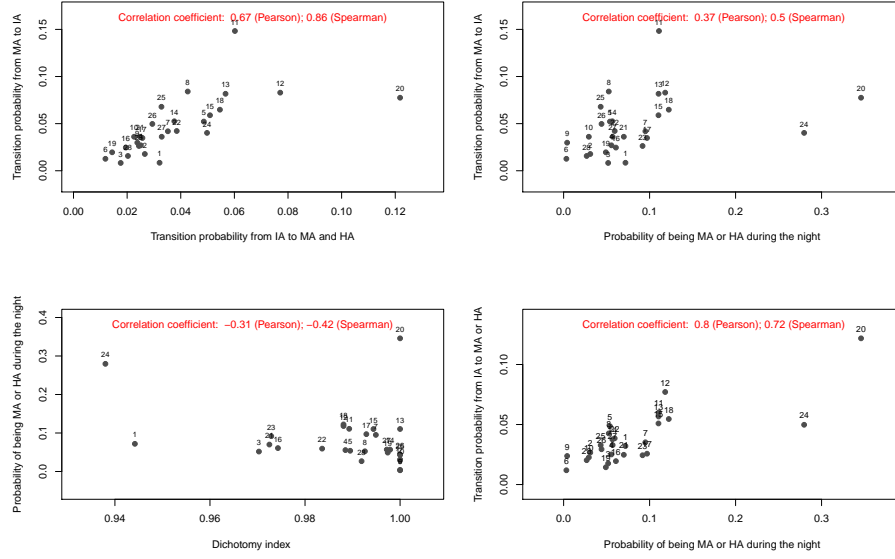


Fig. 3. Correlations and scatterplots between estimated transition probabilities, dichotomy index $I < O$ and probability of being in either of the two active states (MA or HA) during the night for 28 individuals. Subject A and B correspond to the dots marked by 16 and 18, respectively.

the mean and central 90% range of the three estimated observational densities for all 28 individuals, where the results are transformed back to the non-central Chi-square distribution at the scale of the PA counts. The mean of the highest state provides an alternative estimator of the *amplitude* without having to rely on Fourier methods. The three states identified by the HMM are specific to each subject and we can see a large variability in the intensity of the two active states between the individuals which is presumably due to their varying lifestyles. Naturally, there is small variability between subjects in the mean of the IA state.

Local decoding [18] can be used to estimate the predicted state at time t by

$$\hat{S}_t = \operatorname{argmax}_{j=1,\dots,m} P(S_t = j | Y^{(T)}),$$

where the estimated conditional state probabilities $P(S_t = j | Y^{(T)})$ are conveniently available as part of the inference algorithm. The predicted sequence of the most likely states for the two example individuals can be seen in the top panels of Figure 5 and 6, which show that for both individuals the IA state is predominant at night and that during the day there are many transitions between the MA and HA states. For an informative visualization we propose to plot $P(S_t = j | Y^{(T)})$ for $j = 1, 2, 3$ (which add to 1) cumulatively for each t , and

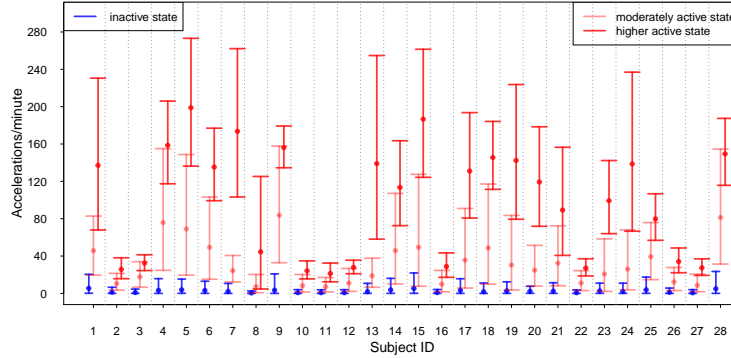


Fig. 4. Mean and central 90% range (on original scale of PA counts) of the estimated conditional observation densities for all 28 subjects. Subject A and B correspond to ID 16 and 18, respectively.

associate each state with a different colour (blue for IA, light red for MA, dark red for HA). We shall refer to the resulting plot as *cumulative state probability* (CSP) plot. The bottom panels of Figure 5 and 6 show the CSP plots for subjects A and B. These diagrams allow us to quickly assess how probable the most likely state is and what other states have noticeable probability and give us visual information on how well a person has rested. In particular if they have solid blue areas, i.e. rarely move into the active states during night, then we might deduce that the person has obtained a good night's rest, as the example subject A seems to have done. In contrast, subject B (Figure 6) has experienced many interruptions at night which may be indicative of relatively poor quality of sleep. Figure 7 compares the CSP plots of the average daily state probabilities for all 28 individuals where one can identify individuals who have experienced more or less interrupted rest at night, as well as differences between individuals who tend to start their night rest earlier or later.

4 Circadian periodicity and Time-varying transition probabilities

Spectral analysis confirms that the circadian cycle is the most dominant component in the spectrum of the PA counts. Although the HMM introduced above will reconstruct retrospectively the circadian rhythm in the states without any *a-priori* modelling assumption of periodicity, it will not reproduce the circadian peak in the spectrum of the PA counts. We can explicitly assume that the transition probabilities follow a circadian rhythm through adding a periodic parametric form as covariate. Banachewicz in [20] suggest the following logistic

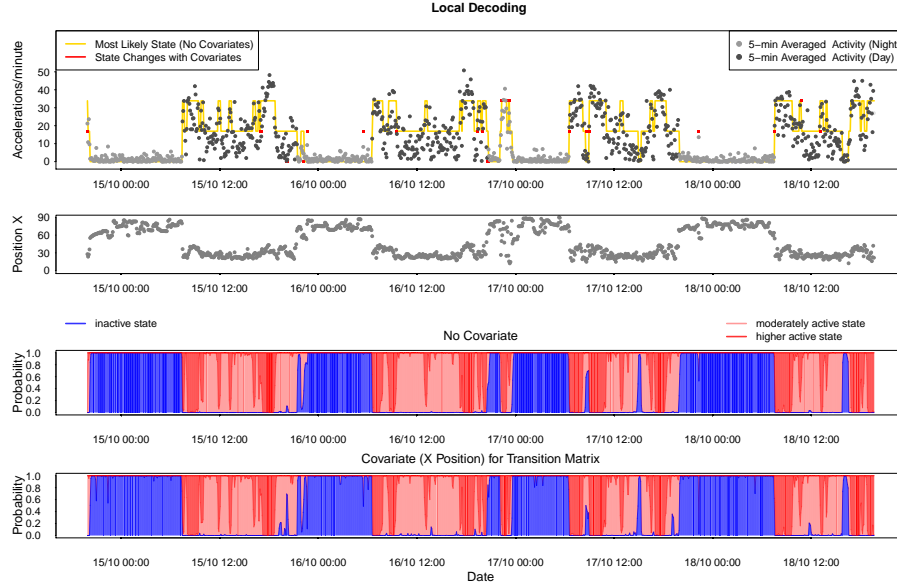


Fig. 5. State estimation for example subject A. (a) Top row: time series of activity with yellow line indicating the mostly likely state using local decoding. (b) Second row: time series of position (angle with respect to a device-specific X-axis). (c) Third row: CSP plot, i.e. cumulative plot of $P(S_t = j|Y^{(T)})$ for $j = 1$ (IA, blue), 2 (MA, light red), 3 (HA, dark red). (d) Bottom: CSP plot from HMM estimated with position covariate.

link functional approach

$$P(S_t = k | S_{t-1} = j, X_t) = \frac{\exp(\beta_{0,j,k} + \beta_{1,j,k} X_t)}{\sum_{k=1}^m \exp(\beta_{0,j,k} + \beta_{1,j,k} X_t)} \quad (3)$$

where X_t is a time varying covariate, here the 24-hour harmonic. Hence, we can estimate and analyze the circadian variation of each of the transition probabilities in the matrix $\mathbf{\Gamma}$ for each individual. Figure 8 compares for all 28 individuals the CSP plots with the estimated state probabilities for the fitted HMM with a circadian harmonic. These give smooth representations of the individual-specific typical circadian cycle in the transitions between the three states, and one can clearly distinguish between the various types of circadian rhythmicity such as early risers (for example individual 19), late types (individual 21) and individuals who experience a lot of interruptions at night (for example subjects 2, 20 and 24). We note that the estimated periodic HMM models provide synthetic data simulations that appear realistic and indistinguishable from real data. Such simulations are of practical importance as they can further be used to quantify the variability of nonparametric statistic of interest such as the $I < 0$ dichotomy index for each individual.

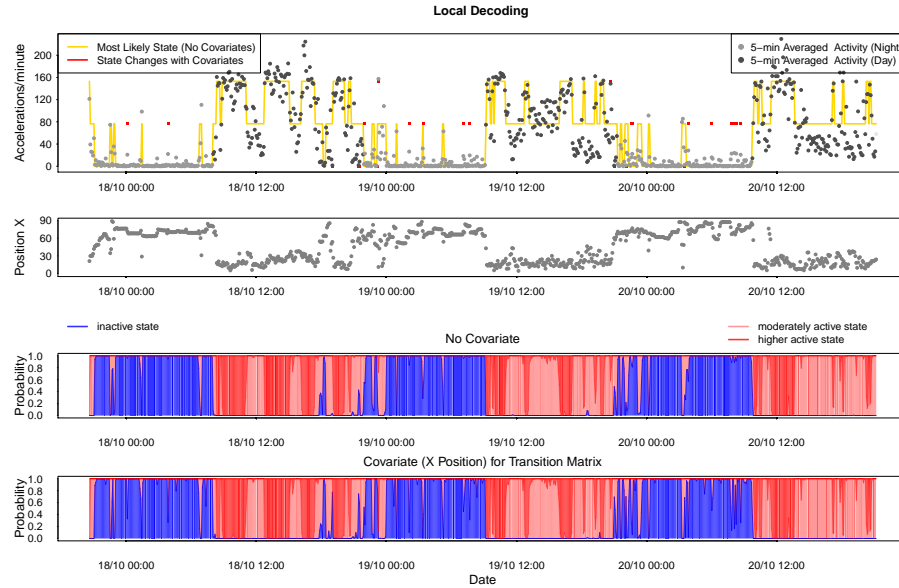


Fig. 6. State estimation for example subject B. (a) Top: time series of activity with yellow line indicating the mostly likely state using local decoding. (b) Second: time series of position (angle with respect to a device-specific X-axis). (c) Third: CSP plot, i.e. cumulative plot of $P(S_t = j|Y^{(T)})$ for $j = 1$ (IA, blue), 2 (MA, light red), 3 (HA, dark red). (d) Bottom: CSP plot from HMM estimated with position covariate.

5 Summary and Discussion

In this paper we propose the use of a hidden Markov modeling approach which can address the challenges of modelling activity data and provides a natural modelling framework for extracting information from them. The model can capture the characteristic features discernible in time series of activity measured over days, such as the notable square wave form with heterogeneous ultradian variances over the circadian cycle of human activity as well as the polyphasic nature of the sleep-wake cycle in rodents. The estimated parameters can be used to characterize the individual pattern of sleep-wake behaviour and to study the between-individual variability. We introduce the cumulative state probability (CSP) plot as a visual tool for inspecting the dynamic pattern of state transitions and their associated uncertainties. The possibility of assuming that the state transition probabilities may change over time according to covariate information and/or periodic functions allows for a wide range of modelling approaches that has the potential to deal with the multivariate complex and large physiological data sets that may in the near future be acquired regularly and cheaply due to the rapidly developing technology of wearable devices [21]. Parameter inference via maximum likelihood requires the use of optimization procedures for

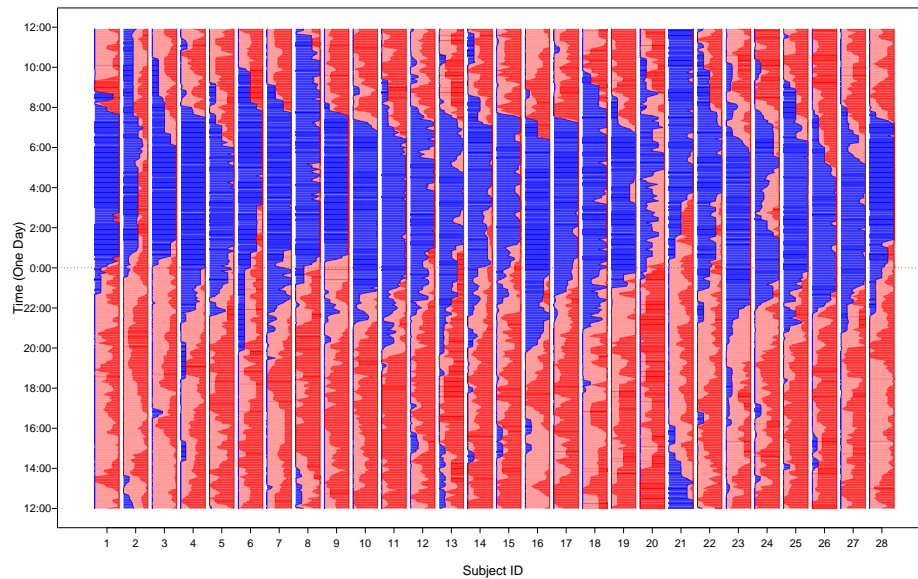


Fig. 7. CSP plots resulting from average daily state probabilities for all 28 individuals in sample. Subject A and B correspond to ID 16 and 18, respectively.

which computationally accessible methodologies exist at least for some standard distributional choices. We note that we have assumed Gaussianity for a suitable transformation of the data and hence our HMM models were relatively easy to implement in particular since some R packages such as *HiddenMarkov* and *dep-mixS4* are already available. The results from the analyses of pseudo-residuals indicate a reasonable model fit and the interpretability of the results are encouraging. Care must be taken as convergence of the inference algorithms is affected by increasing model complexity. Activity counts taken at very short lengths of epochs display a large proportion of zero and low integer counts during the prolonged IA states. Hence the development of estimation algorithms for mixtures of zero-inflated discrete distributions and Gaussian distributions for the active states may provide an interesting avenue to pursue in order to deal with shorter epoch lengths. However, Bai et al. [22] point out that there are significant differences in the computation of physical activity counts between manufacturers and even for new devices from the same manufacturer while wearable devices are developing rapidly gaining increasing market attention via smart watches, mobile phones and bracelets where there is currently no consensus about their quality in assessing activity duration and sleep quality [23]. Activity recordings mark the beginning of sleep periods by immobility of the subject and therefore tend to overestimate sleep and underestimate wake time [2, 24] in comparison to polysomnography (PSG), the current gold standard for measuring sleep, which

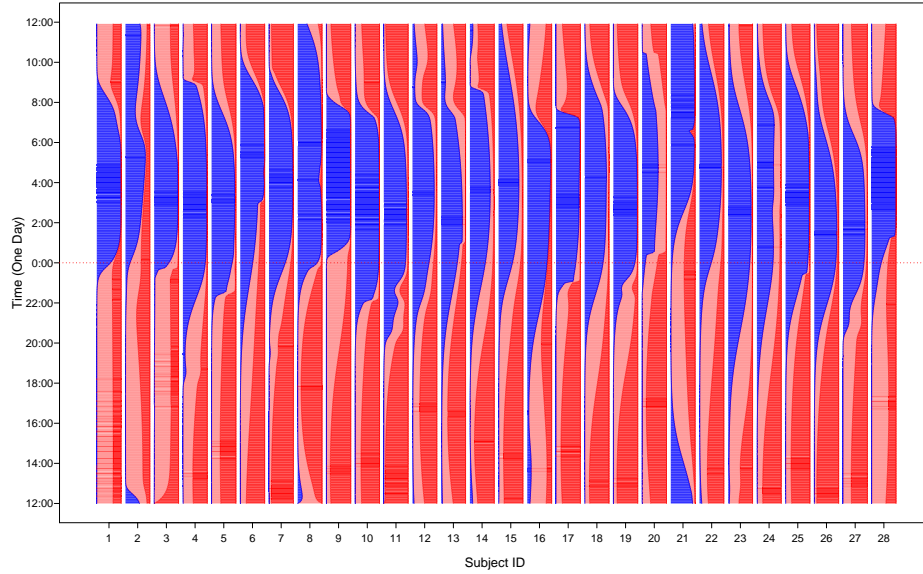


Fig. 8. CSP day profile plots resulting from HMM with circadian harmonic fitted to all 28 individuals in sample. Subject A and B correspond to ID 16 and 18, respectively.

will mark the onset of sleep through changes of electrical activity patterns in the brain. Hence, the accuracy of activity recordings obtained by accelerometers in measuring actual sleep continues to be investigated [25, ?]. In a recent study Migueles et al. [26] review data collection and processing criteria where they uncover significant effects on data comparability with respect to placement, epoch length, sampling frequency, frequency setting of the filtering process that selects the acceleration measured and treatment of missing data (usually due to removal to avoid contact with water) for different generations of accelerometers devices. Although it cannot address such differences in the quality of data resulting from different types of measuring device, an advantage of the HMM approach lies in that fact that it translates the information from the observed data into probabilities of activity states and thus enables a comparison between studies although they may be based on fundamentally different ways of measuring activity. Furthermore, it solves the problem of "thresholding" activity into different states in an appropriate way through a probabilistic model respecting dependencies in time and with the possibility of taking into account further data and additional information. We may in future research wish to include weekend effects and information about the state of the patient's disease and therapy. Finally, the HMM provides a model on the basis of which realistic artificial data can be simulated to quantify the individual-specific variability of nonparametric

statistics, such as the $I > O$, and thus will aid in evaluating the risk associated with the use of such statistics for therapeutic treatment decisions in clinic.

References

1. Sadeh, A.: The role and validity of actigraphy in sleep medicine: an update. *Sleep Med. Rev.* **15**(4), 259–267 (2011)
2. Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W., Pollak, C.: The role of actigraphy in the study of sleep and circadian rhythms. *American academy of sleep medicine review paper. Sleep* **26**(3), 342–392 (2003)
3. Ancoli-Israel, S., Martin, J.L., Blackwell, T., Buenaiver, L., Liu, L., Meltzer, L.J., Sadeh, A., Spira, A.P., Taylor, D.J.: The sbsm guide to actigraphy monitoring: Clinical and research applications. *Behavioral Sleep Medicine* **13**(sup1), 4–38 (2015)
4. Goncalves, B.S.B., Cavalcanti, P.R.A., Tavares, G.R., Campos, T.F., Araujo, J.F.: Nonparametric methods in actigraphy: An update. *Sleep Science* **7**(3), 158–164 (2014)
5. Goncalves, B.S.B., Adamowicz, T., Mazzilli Louzadac, F., Roberta Moreno, C., Araujo, J.F.: A fresh look at the use of nonparametric analysis in actimetry. *Sleep Medicine Reviews* **20**, 84–91 (2015)
6. Blume, C., Santhi, N., Schabus, M.: nparact package for r: A free software tool for the non-parametric analysis of actigraphy data. *MethodsX* **3**, 430–435 (2016)
7. Bromundt, V., Köster, M., Georgiev-Kill, A., Opwis, K., Wirz-Justice, A., Stoppe, G., Cajochen, C.: Sleep/awake cycles and cognitive functioning in schizophrenia. *Br J Psychiatry* **198**(4), 269–276 (2011)
8. Oosterman, J., van Someren, E., Vogels, R., Van Harten, B., Scherder, E.: Fragmentation of the rest/activity rhythm correlates with age-related cognitive deficits. *J Sleep Res* **18**(1), 129–135 (2009)
9. Hatfield, C., Herbert, J., van Someren, E., Hodges, J., Hastings, M.: Disrupted daily activity-rest cycles in relation to daily cortisol rhythms of home-dwelling patients with early alzheimer's dementia. *Brain* **127**(Pt 5), 1061–74 (2004)
10. Minors, D., Akerstedt, T., Atkinson, G., Dahlitz, M., Folkard, S., Levi, F., Mormont, C., Parkes, D., Waterhouse, J.: The difference between activity when in bed and out of bed. i. healthy subjects and selected patients. *Chronobiology international* **13**(1), 27–34 (1996)
11. Natale, V., Innominato, P.F., Boreggiani, M., Tonetti, L., Filardi, M., Parganiha, A., Fabbri, M., Martoni, M., Lévi, F.: The difference between in bed and out of bed activity as a behavioral marker of cancer patients: A comparative actigraphic study. *Chronobiology international* **32**(7), 925–933 (2015)
12. Lévi, F., Dugué, P.-A., Innominato, P., Karaboué, A., Dispersyn, G., Parganiha, A., Giacchetti, S., Moreau, T., Focan, C., Waterhouse, J., *et al.*: Wrist actimetry circadian rhythm as a robust predictor of colorectal cancer patients survival. *Chronobiology international* **31**(8), 891–900 (2014)
13. Chang, W.-P., Lin, C.-C.: Correlation between rest-activity rhythm and survival in cancer patients experiencing pain. *Chronobiology international* **31**(8), 926–934 (2014)
14. Fischer, D., Vetter, C., Roenneberg, T.: A novel method to visualise and quantify circadian misalignment. *Scientific Reports* **6** (2016)

15. Wang, J., Xian, H., Lici, A., Deych, E., Ding, J., McLeland, J., Toedebusch, C., Li, T., Duntley, S., Shannon, W.: Measuring the impact of apnea and obesity on circadian activity patterns using functional linear modeling of actigraphy data. *Journal of Circadian Rhythms* **9:11** (2011)
16. Fuster-Garcia, E., Juan-Albarracin, J., Bresó, A., Garcia-Gomez, J.M.: Monitoring changes in daily actigraphy patterns of free-living patients. In: Conference Proceedings IWBBIO: 18-20 March 2013; Granada, pp. 685–693 (2013)
17. Costa, M.J., Finkenstädt, B., Roche, V., Lévi, F., Gould, P.D., Foreman, J., Halliday, K., Hall, A., Rand, D.A.: Inference on periodicity of circadian time series. *Biostatistics* **14**(4), 792–806 (2013)
18. Zucchini, W., MacDonald, I.L.: *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall, United States (2009)
19. Peursum, P., Bui, H.H., Venkatesh, S., West, G.: Human action segmentation via controlled use of missing data in hmms. In: *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference On*, vol. 4, pp. 440–445 (2004). IEEE
20. Banachewicz, K., Lucas, A., van der Vaart, A.: Modelling portfolio defaults using hidden markov models with covariates. *The Econometrics Journal* **11**(1), 155–171 (2008)
21. Sanders, J.P., Loveday, A., Pearson, N., Edwardson, C., Yates, T., Biddle, S.J., Esliger, D.W.: Devices for self-monitoring sedentary time or physical activity: A scoping review. *Journal of medical Internet research* **18**(5) (2016)
22. Bai, J., He, B., Shou, H., Zipunnikov, V., Glass, T.A., Crainiceanu, C.M.: Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics* **15**(1), 102 (2013). doi:10.1093/biostatistics/kxt029
23. Wen, D., Zhang, X., Liu, X., Lei, J.: Evaluating the consistency of current mainstream wearable devices in health monitoring: A comparison under free-living conditions. *J Med Internet Res* **19**(3), 68 (2017). doi:10.2196/jmir.6874
24. Tryon, W.W.: Issues of validity in actigraphic sleep assessment. *SLEEP-NEW YORK THEN WESTCHESTER-* **27**(1), 158–165 (2004)
25. Marino, M., Li, Y., Rueschman, M.N., Winkelman, J., Ellenbogen, J.M., Solet, J., Dulin, H., Berkman, L.F., Buxton, O.M.: Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* **36**(11), 1747 (2013)
26. Migueles, J.H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J.R., Ortega, F.B.: Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Medicine*, 1–25 (2017). doi:10.1007/s40279-017-0716-0

Acknowledgements: QH, BFF, FL are supported by the UK Medical Research Council Council, Grant reference: MR/M013170/1.

Forecasting via Fokker-Planck using conditional probabilities

Abstract

Using a closed solution to a Fokker-Planck equation model of a time series, a probability distribution for the next observation is developed. This pdf has one free parameter, b . Various approaches to selecting this parameter have been explored: most recent value, weighted moving average, etc. Here we explore using a conditional probability distribution for this parameter b , based upon the most recent observation. These methods are tested against some real world product sales for both a one step ahead and a two step ahead forecast. Significant reduction in safety stock levels are found versus an ARMA approach, without a significant increase in out-of-stocks.

Chris Montagnon

Department of Mathematics, Imperial College, London SW7 2AZ, UK. May 2017.

1. Introduction

When forecasting a time series $\{X_t, t = 1, \dots, N\}$, rather than the 'best' (e.g. minimum squared error) prediction of a single value \hat{X}_{N+1} the expected value of the next point in the time series, one often requires a probability distribution of the possible values of \hat{X}_{N+1} . Kantz and Schreiber [1] proposed tackling this through a Fokker-Planck equation [2] but did not take this further because of difficulty estimating the parameters. Several more recent papers (eg Refs [3-5]) have sought to use forecasting methods based upon a diffusion model leading to a Fokker-Planck equation but the solutions have been numerical. References 6,7 and 8 also report difficulties in estimating the parameters in a Fokker-Planck model. In this paper we use a conditional probability approach to estimate these parameters.

In Ref [9] we modeled a time series using a drift coefficient $D^{(1)} = -\gamma x$ and diffusion coefficient $D^{(2)} = c - bx^2$ in a Fokker-Planck equation:

$$\frac{\partial W}{\partial t} = -\frac{\partial}{\partial x}(D^{(1)}W) + \frac{\partial^2}{\partial x^2}(D^{(2)}W)$$

where $W(x, t|x_N = X_N)dx$ is the probability of finding the actual X_{N+t} in $(x, x + dx)$ when the value X_N has been observed for x_N .

This lead to a differential equation in W :

$$(c - bx^2)W_{xx} + (\gamma - 4b)xW_x + (\gamma - 2b)W = W_t \dots (1.1)$$

The diffusion coefficient should be positive, so $c \geq bX_{\max}^2$, and one can show that the variance of $W(x, t)$ increases with increasing c , so c should be as small

as possible, giving:

$$c = bX_{max}^2, \text{ where } X_{max} = \max(|X_t|, t = 1, \dots, N) \dots (1.2)$$

In order to reach a closed solution in Ref [9] we needed the constraint:

$$\gamma = 3b \dots (1.3)$$

This lead to the solution:

$$W_b(x, t | x_n = X_N) =$$

$$\frac{e^{bt}}{2\sqrt{\pi tb}(X_{max}^2 - X_N^2)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{4tb} \left(\sin^{-1}\left(\frac{x}{X_{max}}\right) - \sin^{-1}\left(\frac{X_N}{X_{max}}\right) \right)^2 \right\} \dots (2)$$

The two constraints (1.2 and 1.3) on the parameters γ, c , and b mean that this distribution is dependent only on one free parameter, which in this paper we choose to be b . This paper explores how our knowledge of previous values of the time series $\{X_t, t = 1, \dots, N\}$ helps select a value or a distribution for b .

2. Definition of the past values of parameter b

We define $b_{\tau-j}$ to be the smallest b that makes the observed point $X_{\tau-j+1}$ just less than the 75% point of the distribution $W_b(x, 1 | x_{\tau-j} = X_{\tau-j})$; ie. $b_{\tau-j}$ is the solution to:

$$\int^{X_{\tau-j+1}} W_{b_{\tau-j}}(x, 1 | x_{\tau-j} = X_{\tau-j}) dx = 0.75 \text{ solved for } b_{\tau-j} \dots (3)$$

For a given time series, up to point τ , this generates a set of values $(b_1, b_2, \dots, b_{\tau-1})$. If we consider a discrete set of possible values for b , say $b^{(j)} : j = 1 \text{ to } 31$, we obtain a distribution of these values similar to that shown in figure 1.

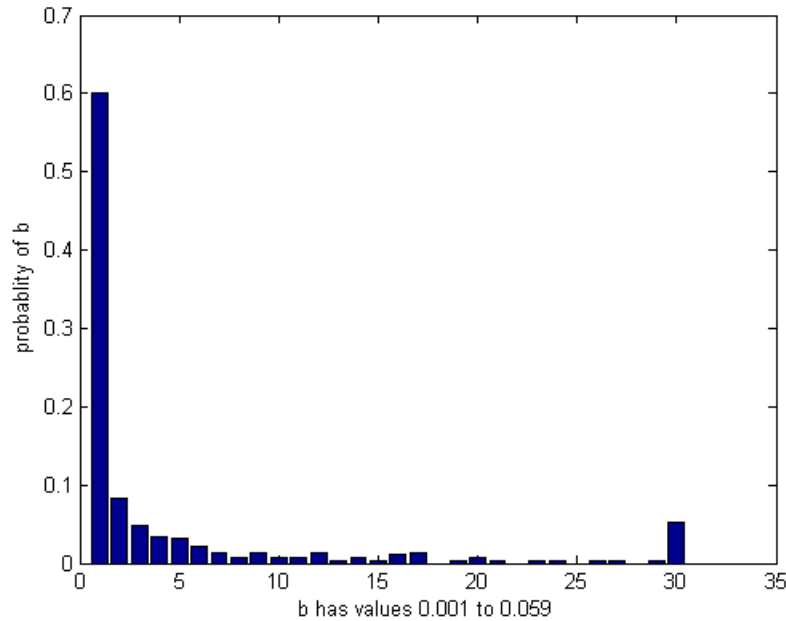
In this paper we explore ways of choosing b_τ based upon what has just happened ie the observed $b_{\tau-1}$ as defined in (3).

3. Method for one step ahead forecasts

We consider first the situation where we want a forecast for tomorrow (ie $D+1$) when we are at the end of today ($D+0$): so we might be placing an order to meet sales for tomorrow ($D+1$). Given a time series of sufficient length (eg. $N > 100$) and considering a discrete set of possible values for b , as above, we can form not only the overall probability distribution for b as in figure 1, but also set of discrete conditional probability distributions.eg.

$$p(b_\tau | b_{\tau-1} = b^{(k)}) \dots (4)$$

Figure 1
Probability Distribution for parameter b across all 300 points of the time series



Note:

1. Whenever $X_{\tau+1}$ is $<$ mean of $W_b(x, 1|x_\tau = X_\tau)$, then b is given the smallest value (in this case 0.001) in solving (4.2). Thus for at least 50% of the points this should be the value selected.
2. The final b value is the default value used when (4.2) does not solve, so creating the final probability shown which is really $\text{prob}(b \geq 0.059)$

Thus for example in an extreme situation we might find that every time the value $b^{(1)}$ occurred it was always followed by the b value $b^{(31)}$, in which case we would have :

$$\text{prob}(b_\tau | b_{\tau-1} = b^{(1)}) = \delta(b - b^{(31)})$$

For the test series we used, each of which had some 300 points, we found that in practice only $b^{(1)}, b^{(2)}, b^{(31)}$ occurred sufficiently often to build a conditional probability distribution as per (4): ie only for $k=1,2$ or 31 did we have enough ($>$ ca. 20) points to generate any meaningful distribution.

With this information on the likely values of b_τ that follow a particular observed value for $b_{\tau-1}$, at least for $b_{\tau-1} = b^{(1)}, b^{(2)}$ or $b^{(31)}$, the method for generating $W_b(x, t|x_n = X_N)$ became:

(i) If $b_{\tau-1} = b^{(1)}, b^{(2)}$ or $b^{(31)}$ then

$$W(x, 1|x_{\tau} = X_{\tau}) = \frac{1}{\mathcal{M}} \sum_{j=1}^{31} \text{prob}(b^{(j)}|b_{\tau-1}) W_{b^{(j)}}(x, 1|x_{\tau} = X_{\tau}) \dots (6)$$

where $b_{\tau-1} = b^{(k)}$ and \mathcal{M} is a normalising factor.

(ii) If $b_{\tau-1}$ not $= b^{(1)}, b^{(2)}$ or $b^{(31)}$ then

$$W(x, 1|x_{\tau} = X_{\tau}) = \frac{1}{28} \sum_{j=1}^7 (8-j) W_{b_{\tau-j}}(x, 1|x_{\tau} = X_{\tau}) \dots (7)$$

ie. as defined in refs 9 and 10 where this was shown to be one of the best methods of defining $W_{b_{\tau}}(x, t|x_n = X_N)$. ($b_{\tau-j}$ is as defined in (3)).

(iii) In the situation of (ii) above: ie. $b_{\tau-1}$ not $= b^{(1)}, b^{(2)}$ or $b^{(31)}$, we also tested using using a probability distribution

$$W(x, 1|x_{\tau} = X_{\tau}) = \frac{1}{\mathcal{M}} \sum_{k=1}^{31} \text{prob}(b^{(k)}) W_{b^{(k)}}(x, t|x_{\tau} = X_{\tau}) \dots (8)$$

where $\text{prob}(b^{(k)})$ is the probability distribution of $b^{(k)}$ unconstrained by the value of $b_{\tau-1}$, ie similar to the discrete pdf in figure 1.

3. Results for one step ahead forecasts

We applied the above method to the ten test series a defined in ref 11. We used 200 points for each series. This gave the results in table 1.

Table 1
Results for one step ahead forecasts
from various conditional probability approaches to b_{τ} in $W_{b_{\tau}}(x, 1|x_{\tau} = X_{\tau})$

	AR	Method 1a	Method 1b	Method 1c
Av Stock	112	101	106	104
stock out %	4.6	9.6	6.5	6.6

Key:

AR: Stock level is 95% pt of a Normal with mean = AR forecast, variance computed from past errors)

Method 1a: Stock level is 95% pt. of $W_{b_{\tau}}(x, 1)$, b chosen as (7) above for all b

Method 1b: Stock level is 95% pt. of $W_{b_{\tau}}(x, 1)$, b chosen as (6) and (7) above

Method 1c: Stock level is 95% pt. of $W_{b_{\tau}}(x, 1)$, b chosen as (6) and (8) above

The performance of all three versions of method 1 are compared to results using Normal distribution with mean equal to AR(7) forecast and with the variance calculated from the forecast errors from past data.

Using method 1a, ie. when stock level is set at 95% point of $W_b(x, 1|x_\tau = X_\tau)$ which is chosen as (7) above - but for all values of $b_{\tau-1}$, gave average stock level of 101 which is lower than the AR reference of 112 but stock outs are higher at 9.6%.

Taking note of the value of $b_{\tau-1}$ that has just occurred and using the conditional probabilities as per (i) and (ii) above (ie. method 1b) average stock level increased slightly over method 1a, at 106 but was still lower than the reference AR solution and stock outs were the lowest at 6.5%.

Applying a probability distribution for all $b_{\tau-1}$ using (6) and (8) above (ie method 1c) gave the lowest average stock level at 104 but stockouts rose slightly as as compared to method 1b: 6.6% vs. 6.5%.

Thus from these results we can conclude:

- using method 1a (computing $W_b(x, 1|x_\tau = X_\tau)$ from a weighted average of recent values) reduces stock by 10% versus a conventional AR method, but doubles the number of stockouts.
- introducing a 'conditional probability' approach (methods 1b and 1c) still reduces stock versus the AR method by some 6% but now stockouts are only slightly over the 5% target.

So the conditional probability method of calculating b_τ is worth pursuing.

4. Method for two step ahead forecasts

In many real world situations the reordering of stock to meet customer demand has to allow time for delivery from the supplying warehouse. Thus at the end of day 0 one might calculate the stock that would need to be delivered at the end of day 1 in order to meet demand in day 2. We will call this situation a two step ahead forecast.

In addition to making an estimate of demand in day 2 , one needs to take a view as to what might have happened in day 1 so as to compute what stock might be available at the end of day1/start of day 2 before taking into account how much be added to this to meet demand in day 2. To do this we need two pdfs:

$W_{b'}(x, 1|x_\tau = X_\tau)$: the pdf made at the end of day τ for sales in day $\tau + 1$.

$W_{b''}(x, 2|x_\tau = X_\tau)$: the pdf made at the end of day τ for sales in day $\tau + 2$.

Thus if S is the stock available at the end of day τ (day 0) after the delivery has been made that night: ie S is the total stock available for demand in day 1 (day $\tau + 1$). and if $W_{b''}^{95}(2)$ is the stock required at the start of day 2 (ie the level that if achieved would meet 95% of demand in day 2, after the delivery

that is to be made end day1/start day 2), then the order to be delivered at end of day 1 is expected to be:

$$\int_{-|X|_{max}}^{+|X|_{max}} \max \{ [W_{b''}^{(95)}(2) - \max((S-x), 0)], 0 \} \cdot W_{b'}(x, 1|x_{\tau} = X_{\tau}) \cdot dx \dots (9)$$

5..Results for two step ahead forecasts

In table 2 we see the results of this two step ordering process under various methods of choosing the values for b' and b'' .

These are two different AR solutions which are used as a reference. For the first solution (AR(a)) the pdf for sales in day 1 is taken as Normal with a mean ($m1$) = the AR one step forecast and a variance (s_1^2) calculated from past forecast errors. Also the pdf for sales in day 2 is taken as Normal with mean ($m2$) = the AR forecast from regression of X_t on $X_{t-2}, X_{t-3}, \dots, X_{t-8}$ and variance (s_2^2) from the past errors in this 2 step forecast. That is in (9) above:

$$W_{b'}(x, 1|x_{\tau} = X_{\tau}) \text{ is replaced by } \text{Normal}(m1, s_1^2) \dots (10)$$

and to find $W_{b''}^{(95)}(2)$

$$W_{b''}(x, 2|x_{\tau} = X_{\tau}) \text{ is made } = \text{Normal}(m2, s_2^2) \dots (11)$$

This reference AR(a) solution generates an average stock level of 112 with a stockouts at 8.5%.

The second AR solution, (AR(b)), takes the pdf of sales in day 1 as

$$W_{b'}(x, 1|x_{\tau} = X_{\tau}) = \delta(W^{(95)}(1) - x).$$

where $W^{(95)}(1)$ is the 95% point of (10). ie. we use a single value for our estimate of day 1 sales in calculating this order for delivery end day1/ start day 2. $W_{b''}^{(95)}(2)$ is again the 95% point of the Normal distribution (11). This method gives an average stock level of 160 and stockouts at 3.5%

In method 2a, the first application of our methods to this two step ahead problem, we take:

- the pdf for day 1, $W_{b'}(x, 1|x_{\tau} = X_{\tau})$, where b' is calculated as in method 1b above
- the pfd for day 2, $W_{b''}(x, 2|x_{\tau} = X_{\tau})$ where the b'' are as defined in (3) but of course the $b_{\tau-j}$ are redefined to reflect the 'best' $b_{\tau-j}$ such that $X_{(\tau-j)+2}$ is at the 75% point of $W_{b_{\tau-j}}(x, 2|x_{\tau-j} = X_{\tau})$

This method, method 2a, gives an average stock level of 114 but stock outs of 11.4%.

Table 2 Results for 2 step ahead forecasts from various conditional probability approaches to b in $W_{b_\tau}(x, 2 x_\tau = X_\tau)$					
	ARa	ARb	Method 2a	Method 2b	Method 2c
Av Stock	112	160	114	112	109
stock out %	8.5	3.5	11.4	8.4	8.2

Key:

ARa: pdf day 1, Normal (AR forecast day 1, variance from past),
pdf day 2, Normal (AR forecast day 2, variance from past)
ARb: pdf day 1, Delta (95% point of day1),
pdf day 2 Normal (AR forecast day 2, variance from past)
Method 2a: $W_{b_\tau}(x, 1)$, b as method 1b, (2) $W_{b_\tau}(x, 2)$, b as method 1a,
Method 2b: $W_{b_\tau}(x, 1)$, b as method 1b, (2) $W_{b_\tau}(x, 2)$, b as method 1b
Method 1c: $W_{b_\tau}(x, 1)$, b as method 1b, (2) $W_{b_\tau}(x, 2)$, b as method 1c

In method 2b, we introduce the conditional probabilities $p(b_j|b_k)$ for $k=1,2$, or 31, in order to calculate the pdf for day 2: ie.

(i) If $b_{\tau-1} = b^{(1)}, b^{(2)}$ or $b^{(31)}$ then

$$W(x, 2|x_\tau = X_\tau) = \frac{1}{\mathcal{M}} \sum_{j=1}^{31} \text{prob}(b^{(j)}|b^{(k)}) W_{b_j}(x, 2|x_\tau = X_\tau) \dots (12)$$

where $b_{\tau-1} = b^{(k)}$ and \mathcal{M} is a normalising factor.

(ii) If $b_{\tau-1}$ not = $b^{(1)}, b^{(2)}$ or $b^{(31)}$ then

$$W(x, 2|x_\tau = X_\tau) = \frac{1}{28} \sum_{j=1}^7 (8-j) W_{b_{\tau-j}}(x, 2|x_\tau = X_\tau) \dots (13)$$

With this we get the results shown in column 5 of table 2 : average stock level of 112 and stockouts at 8.4%.

Finally, in method 2c, we introduce a discrete probability distribution for all the b_j : ie not only $p(b_j|b_k)$ for $k=1,2$, or 31, but $p(b_j)$ for all other k . As shown in column 6 this reduces the average stock level slightly further: now 109, but stockouts stay much the same at 8.2%

6. Conclusion

Instead of setting reorder stock levels through a conventional approach with a forecast sales pdf Normal with mean equal to the AR forecast, we have used a solution to an appropriate Fokker-Planck equation to generate a pdf for sales $W_b(x, t|x_\tau = X_\tau)$ which has a free parameter b . Various methods (eg. see ref 9 and 10) have been tried to generate a b that gives a forecast pdf resulting in a low stock level and a low number of stockouts. In this paper we have used a conditional pdf for the value of b which depends on the value of the most recent

b observed. Applying this method to over 2,000 points in a set of test series, first to a one step ahead reordering system, reduces (when compared to an AR method) stock levels in these test series by some 7% although stockouts are still 1.6% above the target of 5%. When the reordering system requires orders to be placed at the end of day 0 for delivery at end of day 1 (and thus for use in day 2), using a conditional probability distribution to select the parameter b in the probability distribution for sales in day 2, $W_b(x, 2|x_\tau = X_\tau)$, gives an improvement of 3% in average stock level and also an improvement 0.3% points in stockouts , both compared to to a conventional AR forecasting approach.

Thus one may conclude that selecting b in $W_b(x, t|x_\tau = X_\tau)$ by a conditional probability approach is worth while.

Bibliography

- [1] Nonlinear Time Series Analysis. H Kantz, T Schreiber. *Cambridge University Press* 2003
- [2] The Fokker Planck Equation. H Risken. *Springer* 1996
- [3] Modeling the Sunspot Number Distribution with a Fokker Planck Equation. P L Noble, M S Wheatland. *Astrophysical Journal* 25 Feb 2011
- [4] Generalized Langevin equation driven by Levy processes: A probabilistic, numerical and time series based approach. A V Medino et alia. *Physica A* 391 2012
- [5] Construction of a Langevin model from time series with a periodical correlation function: Application to wind speed data. Z Czechowski, L Telesca. *Physica A* 392 2013
- [6] A note on Drift and Diffusion parameters from time series. P Sura, J Barsugli. *Physics :Letters A* 305 2002
- [7] Deterministic and probabilistic forecasting in reconstructed spaces. H Kantz and E Olbrich.
- [8] On the definition and handling of different drift and diffusion estimates. J Gottschall, J Peinke. *New Journal of Physics* 10 2008
- [9] A closed solution to the Fokker-Planck equation applied to forecasting. C.Montagnon *Physica A* 2014
- [10] Forecasting with the Fokker-Planck equation: Bayesian setting of parameters. C.Montagnon. *Physica A* 2017
- [11] Singular Value Decomposition and Time Series Forecasting. C.Montagnon . *Phd. Imperial College London* 2011
- [12] Time Series: Theory and Methods. P.J.Brockwell, R.A.Davis. *Springer* 1991
- [13] Handbook of Economic Forecasting Vol 1 Chapter 1: Bayesian Forecasting. I Geweke, C. Whiteman. *Elsevier BV* 2006

Forecasting of CO₂ emissions based on Preprocessing Techniques

Lida Barba*, Guillermo Machado, Lorena Molina,
Ana Congacha, Jorge Delgado, and Lady Espinoza

Facultad de Ingeniería, Universidad Nacional de Chimborazo,
Av. Antonio José de Sucre, 060150 Riobamba, Ecuador
{lbarba,gmachado,lmolina,
acongacha,jdelgado,lespinoza}@unach.edu.ec
<http://www.unach.edu.ec>

Abstract. During the last decades it has been observed an increasing trend of greenhouse gases emissions in the countries. Looking for the environment care, the United Nations declare the incorporation of principles of sustainability development into the policies and programs of the nations. In this context, the forecast of carbon dioxide (CO₂) emissions plays an important role to support the decision making of government and society. The aim of this work is to improve the accuracy of the linear model applied for multi-step ahead forecasting of CO₂ emissions in the Andean Community. The autoregressive model (AR) is improved through data preprocessing techniques. One is based on simple smoothing by means of moving average (MA), the second is based on Singular Spectrum Analysis (SSA), and the third is based on Multilevel Singular Value Decomposition (MSVD). The effectiveness of the combined models MA-AR, SSA-AR, and MSVD-AR are evaluated through the time series of CO₂ per capita emissions of the Andean Community (CAN) countries through historical data from 1960 to 2013. The empirical results provide significant evidence about the effectiveness of the preprocessing data in forecasting. The best combined model MSVD-AR is extended for multi-step ahead forecasting. Projections are presented for supporting the environmental management of government institutions of countries with similar geographical features and cultural diversity.

Keywords: carbon dioxide· Forecasting· Autoregressive Model· Singular Spectrum Analysis· Multilevel Singular Value Decomposition· CO₂.

1 Introduction

In recent years there has been interest of citizens, governments and organizations on environmental degradation. One of the Millennium Development Goals of the United Nations declares the incorporation of principles of sustainability development into the policies and programs of the nations. Unfortunately, according to the data located in

* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

the repositories of the World Bank Group [1]; the carbon dioxide emissions present an upward trend. In 2011, 32,3 billions of metric tons of CO₂ emissions were observed, which were increased to 48,9 in comparison with the emissions in year 1990.

Several investigations have determined that high CO₂ emissions increase the plant photosynthesis and reduce the transpiration [2]. The studies of Tao et al. [3] shown that the effects of CO₂ vary with the temperature, water availability, and solar radiation. The simulations shown that in 2020 as effect of the wheat productivity in China the CO₂ emissions will increase significantly, while it will decrease by the increase of O₃.

In front of the interactive effects of carbon dioxide emissions, an alternative for obtaining knowledge about this phenomenon is the forecasting. Some researchers provide useful projections to support the decision making. For example, Pérez-Suárez and López-Menéndez [5] present the CO₂ forecast of 150 countries based on the Kuznets environmental curve. The study shows an explained variance over 80% for 78 countries, including the CAN members (Ecuador, Colombia, Peru and Bolivia), and an Absolute Average Percent Error near of 7%. On the other hand, Pao And Tsai [4] applied the Gray model in comparison with the ARIMA model to predict the total CO₂ emissions in Brazil. The study presents MAPEs (Average Absolute Percentage Error) among 2.46% and 4.22%. Wu et al. [6] presented the forecast of CO₂ emissions for BRICS countries (Brazil, Russia, India, China and South Africa) by means of the Gray model, the study showed the relationship among the GDP and the energy with respect to the CO₂ emissions. The prediction shown an average MAPE of 2.36%.

The current work presents the forecasting of CO₂ emissions of the Andean Community, which is integrated of Ecuador, Colombia, Bolivia, and Perú. The Autoregressive (AR) model is implemented in conjunction of preprocessing techniques. A simple Moving Average (MA), Singular Spectrum Analysis, and Multilevel Singular Spectrum Analysis are used to improve de accuracy of the AR model.

Simple Moving Average (MA) based on three points obtains a smoothed time series. Singular Spectrum Analysis (SSA) is a technique relatively new for time series analysis, this technique was introduced by Broomhead and King [7]. SSA is commonly used for extracting components from a time series, a trend component, a seasonality component, a cyclic component, and/or a noisy component. SSA is implemented in four steps, embedding, decomposition, grouping, and diagonal averaging. The nature of the components depends of the processes of embedding and grouping, which are not standard processes. In this work the processes are implemented for obtaining two kind of components, one of low frequency and the other of high frequency. On the other hand Multilevel Singular Value Decomposition is a new method that has demonstrated effectiveness in forecasting of traffic accidents, it implements hierarchical decomposition for extracting components of low and high frequency[8].

The document is organized as follows. In Section 2 are described the forecasting methodology. Section 3 specifies the efficiency metrics. Section 4 describes the case studies. Section 5 presents the results and discussion. Finally Section 6 closes the work with the conclusions.

2 Forecasting Methodology

The forecasting methodology is presented in two stages, data preprocessing and prediction. Data preprocessing is implemented by means of Moving Average, Singular Spectrum Analysis, and Multilevel Singular Value Decomposition. Whereas prediction is developed through the Autoregressive model.

2.1 Data Preprocessing

Smoothing based on Moving Average (MA) Moving average is a smoothing strategy used in linear filtering to extract the noise of a time series. MA is a mean of a constant number of observations that can be used to describe a series that does not exhibit a trend [22]. MA typically is computed through the average of each 3 observed points observed. A time series of length N is smoothed with next equation:

$$\tilde{s}_k = \frac{1}{3} \sum_{i=k-1}^{k+1} x_i \quad (1)$$

where \tilde{s}_k is the k -th smoothed signal element, for $k = 2, \dots, N-1$, x_i is each observed element of original time series, terms \tilde{s}_1 and \tilde{s}_n has the same value of x_1 and x_N respectively.

The smoothed values given by the average of each 3-points will be used by the prediction model.

Singular Spectrum Analysis The aim of SSA is to decompose a time series in components, which could represent trend, a seasonality component, a cyclic component, and/or noise. SSA consists of four steps: embedding, decomposition, grouping and diagonal averaging [23].

The *embedding* step, maps the time series x of length N , to a sequence of multidimensional lagged vectors. A real matrix H of $L \times K$ dimension contains the lagged vectors, it is shown as follows:

$$H = \begin{pmatrix} x_1 & x_2 & \dots & x_K \\ x_2 & x_3 & \dots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \dots & x_N \end{pmatrix} \quad (2)$$

where x_1, \dots, x_N are the elements of the observed time series.

The identification of the effective window length L is developed after the execution of the second step (decomposition). Therefore, initially the window length for embedding is set in $L = 2$.

The *Decomposition* step consists in the Singular Value Decomposition (SVD) of the Hankel matrix obtained in the previous embedding step. The SVD of the trajectory matrix H has the form

$$H = \sum_{i=1}^L \sqrt{\lambda_i} U_i V_i^T, \quad (3)$$

where λ_i is the i th eigenvalue of the matrix $S = HH^T$ and $\sqrt{\lambda_i}$ is the i th singular value of the decomposition of H . The singular values are arranged in decreasing order of magnitudes. U_1, \dots, U_L is the corresponding orthonormal system of the eigenvectors of the matrix S , and $V_i = X^T U_i \sqrt{\lambda_i}$. Standard SVD terminology calls $\sqrt{\lambda_i}$ the singular values of the decomposition of H , and U_i and V_i the left and right singular vectors of H , respectively. The collection $(\sqrt{\lambda_i} U_i V_i)$ is called i th eigentriple of H . Elementary matrices can be represented with

$$H_i = \sqrt{\lambda_i} U_i V_i^T, \quad (4)$$

The *Grouping* step arranges the matrix terms H_i . Assume that the two components are required as result of the grouping step; then $I_1 = I = i_1, \dots, i_r$ and $I_2 = 1, \dots, d \setminus I$, where $1 \leq i_1 < \dots < i_r \leq d$. The time series will be separable by decomposition if there exist a collection of indices $I \subset i, \dots, d$ such that

$$X^{(1)} = \sum_{i \in I} X_i, \quad (5a)$$

$$X^{(2)} = \sum_{i \notin I} X_i \quad (5b)$$

The purpose of the grouping step is separation of the additive components of the time series. The set of indices I_1 are considered to obtain the matrix X_{I_1} , therefore $X_{I_2} = H - X_{I_1}$.

The matrices X_{I_1} and X_{I_2} are trajectory matrices, then there exist series that can be called components, c_{I_1} and c_{I_2} such that $x = c_{I_1} + c_{I_2}$.

This step of *Diagonal Averaging* is applied for transforming the grouped matrices $X^{(1)}$ and $X^{(2)}$ into new series of length N .

Let Y_j be an $L \times K$ matrix with elements $y_{i,j}$, $1 \leq i \leq L, 1 \leq j \leq K$, with $L < K, N = L + K - 1$, and $k = i + j$.

Diagonal averaging transfers the elements of the matrix Y_i to the component of low frequency C_L and high frequency C_H as it is shown below:

$$c_{i,j} = \begin{cases} \frac{1}{k-1} \sum_{m=1}^k y_{m,k-m} & \text{for } 2 \leq k \leq L \\ \frac{1}{L} \sum_{m=1}^L y_{m,k-m} & \text{for } L < k \leq K+1 \\ \frac{1}{K+L-k+1} \sum_{m=k-K}^L y_{m,k-m} & \text{for } K+2 \leq k \leq K+L \end{cases} \quad (6)$$

Multilevel Singular Value Decomposition The Multilevel Singular Value Decomposition method was proposed by Barba & Rodriguez [8] to decompose a non-stationary time series into a component of low frequency, and a component of high frequency. MSVD is implemented through multiple levels of decomposition where the steps of embedding, decomposition, and unembedding are executed, as it is illustrated in Fig. 1.

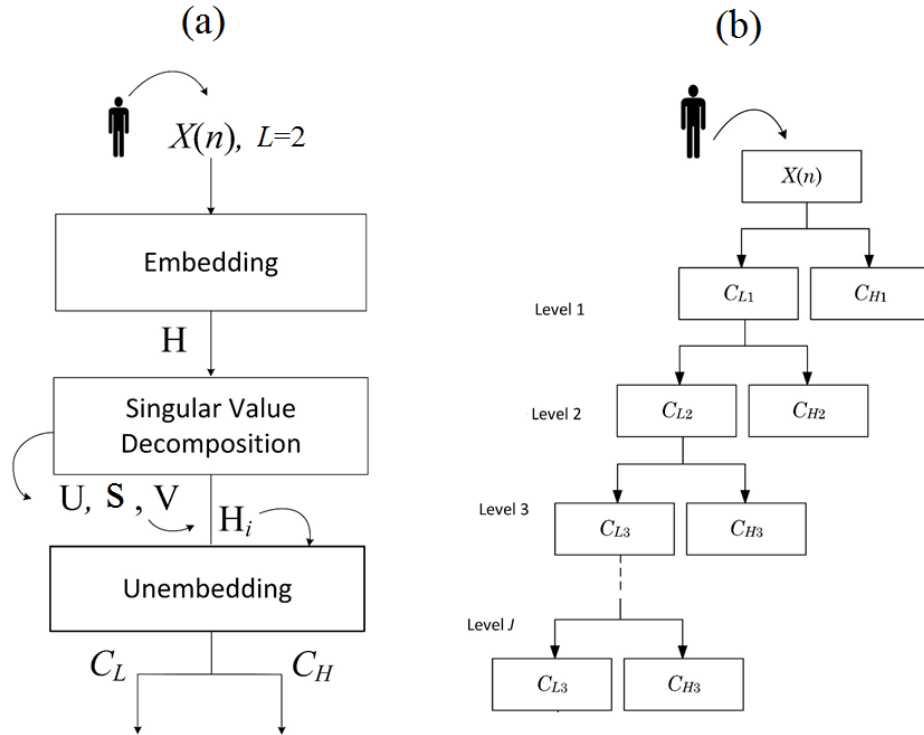


Fig. 1: Multilevel Singular Value Decomposition (a) Detailed processes at each level (b) General process

Embedding consists of mapping the original time series $X = (x_1, x_2, \dots, x_N)$ into a Hankel matrix, of $L \times (N - 1)$ dimensions, where $L = 2$ (rows). The embedding process is illustrated below:

$$H = \begin{pmatrix} x_1 & x_2 & \dots & x_{N-1} \\ x_2 & x_3 & \dots & x_N \end{pmatrix} \quad (7)$$

The *Decomposition* consists of obtaining the subspaces of H , which can be expressed as follows:

$$H = \sum_{i=1}^2 \sqrt{\lambda_i} U_i V_i^T, \quad (8)$$

where U is a square matrix of left singular vectors of dimension L , V is a square matrix of right singular vectors of dimension $N - 1$, and T is used for transposed matrix. While λ_i is the i th singular value of the decomposition of H_i .

Elementary matrices H_1 and H_2 can be obtained with:

$$X^{(1)} = \sum_{i \in I} X_i, \quad (9a)$$

$$X^{(2)} = \sum_{i \notin I} X_i \quad (9b)$$

$$H_1 = U_1 \times \sqrt{\lambda_1} \times V_1^T, \quad (10a)$$

$$H_2 = U_2 \times \sqrt{\lambda_2} \times V_2^T, \quad (10b)$$

The *Unembedding* step is developed for extracting the components from Matrices H_1 and H_2 . The elements are located in the first row and the last columns of each matrix, consequently the components are obtained with

$$C_L = (h_{11}^1, h_{12}^1, \dots, h_{1(N-1)}^1, h_{2(N-1)}^1), \quad (11a)$$

$$C_H = (h_{11}^2, h_{12}^2, \dots, h_{1(N-1)}^2, h_{2(N-1)}^2), \quad (11b)$$

The process is implemented in multiple levels through the component C_L , as it is illustrated in Fig. 1b. The process ends when the computation of the Singular Spectrum Rate (δR) reaches its maximum asymptotic value. The calculation of δR is shown below:

$$\delta R = \frac{R_{j-1}}{R_j}, \quad (12)$$

where R_j is the relative energy of the singular values obtained at each decomposition level j , this computation is developed from the second decomposition level, with respect to the previous decomposition level. The computation of R_j is given by

$$R_k = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \quad (13)$$

2.2 Prediction

The AR model is implemented to predict the time series. The best combination among MA-AR, SSA-AR and MSVD-AR are used for extending the model to multi-step ahead forecasting. The MIMO strategy calculates the multiple horizon forecast in a single step, and preserves the random relationships between historical values that are being used as predictors (Wang et al., 2016). The ARMIMO model is expressed as follows:

$$\tilde{x}(n+1), \dots, \tilde{x}(n+h) = f[z(n), z(n-1), z(n-P+1)], \quad (14)$$

where n is the current time instant, h is the size of the forecast horizon, z is the regressor vector, and P is the size of the regressor vector. In matrix form the expression is as follows:

$$X = \beta \times Z^T, \quad (15)$$

where X is the matrix of estimated values, β is the coefficients matrix of the regression of dimension $h \times 2P$, and Z is the matrix of autoregressive values of the low frequency component and the autoregressive values of the high frequency component. The matrix Z has dimension $N_t 2P$, where N_t is the number of training samples. The regression coefficients are estimated using the Least Squares (LS) method:

$$\beta = X \times Z^\dagger, \quad (16)$$

where Z^\dagger is the Moore-Penrose Pseudoinverse matrix (Serre, 2002).

3 Forecasting accuracy metrics

The accuracy of the prediction is computed with the metrics: Mean Absolute Percentage Error (MAPE), and the modified Nash-Sutcliffe Efficiency (mNSE).

$$MAPE = \left[\frac{1}{N_v} \sum_{i=1}^{N_v} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \right] \times 100 \quad (17)$$

$$RMSE = \sqrt{\frac{1}{N_v} \sum_{i=1}^{N_v} (x_i - \hat{x}_i)^2} \quad (18)$$

where N_v is the testing sample size, x_i is the i -th observed value, and \hat{x}_i is the i -th estimated value.

$$mNSE = 1 - \frac{SAE}{SAD}. \quad (19)$$

where SAE and SAD are defined with

$$SAE = \sum_{i=1}^N |x_i - \hat{x}_i|, \quad (20a)$$

$$SAD = \sum_{i=1}^N |x_i - \bar{x}|, \quad (20b)$$

4 Case Studies

The open repositories of the World Bank Group contains development data of several countries and a variety of topics. Among the time series are those related to carbon dioxide emissions in metric tons per capita of the countries.

The CO₂ emissions per capita of the four countries members of the Andean Community: Ecuador, Colombia, Bolivia, and Perú are presented in Fig. 2. The presented values are calculated by means of the ratio between the total CO₂ emissions and the population of each country. All samples have an annual collection interval, with records from year 1960 to 2013.

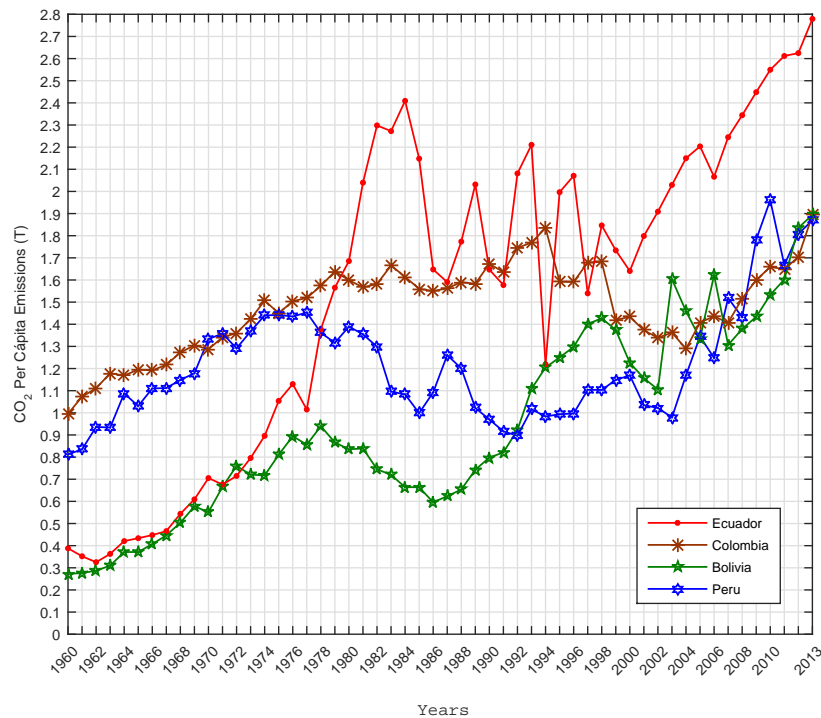


Fig. 2: CO₂ per capita emissions of CAN countries

The emissions in the last decade show an upward trend in the four CAN countries. In the case of Ecuador there is a considerable growth from 1977 with several peaks until 1998. From the year 2000 a more linear behavior, similar to 1960-1976, is observed. CO₂ emissions from Colombia, Peru and Bolivia show similar behavior in terms of variability, which is most evident in recent decades. Table 1 shows statistical and dispersion measurements of the observed data. The highest arithmetic mean of emissions is observed for Ecuador, followed by Colombia, Peru and Bolivia. The maximum value is reached by Ecuador with 2,779 metric tons, followed by Peru, Bolivia and Colombia with 1,961, 1,895 and 1,893 metric tons, respectively. In terms of dispersion measures, it is observed that Ecuador has a historical behavior of greater variability, with a standard deviation of 0.737 and a variance of 0.533, followed by Bolivia with a standard deviation of 0.429 and a variance of 0.181, while Colombia and Peru show a minimum variance of 0.039 and 0.068, respectively.

5 Results and discussion

The data preprocessing by means of simple smoothing based on MA-3 is presented in Fig. 3.

Table 1: Statistical Analysis of Data

	Min	Max	Mean	σ	σ^2
Ecuador	0,325	2,779	1,546	0,737	0,533
Colombia	0,996	1,893	1,479	0,200	0,039
Bolivia	0,272	1,895	0,940	0,429	0,181
Perú	0,812	1,961	1,221	0,262	0,068

Table 2: One-step ahead forecasting Results

	MA-AR		SSA-AR		MSVD-AR	
	MAPE	mNSE	MAPE	mNSE	MAPE	mNSE
	%	%	%	%	%	%
Ecuador	7.9260	23.19	1.0582	90.02	5.7e-05	99.9
Colombia	1.7508	78.87	0.3147	96.17	2.97e-02	99.9
Bolivia	2.9110	72.82	0.9017	92.16	3.45e-05	99.9
Perú	1.2544	93.65	0.2872	98.51	1.57e-05	99.9
Min	1.2544	23.19	0.2872	90.02	1.57e-05	99.9
Max	7.9260	93.65	1.0582	98.51	2.97e-02	99.9
Mean	3.4606	67.13	0.6405	94.22	7.45e-03	99.9

The data preprocessing based on SSA is presented in Fig. 4, the effective window length was identified through trial and error tests in $L = 7$ for all series. The first elementary matrix $H1$ was used to obtain the component of low frequency, whereas the rest of elementary matrices was grouped to obtain the component of high frequency.

The data preprocessing based on MSVD is presented in Fig. 5. The optimal number of decomposition levels was set in $J = 13$, which was due to the curve obtained by plotting the ΔR parameters, they shown the asymptotic value at $J = 13$, as shown in Fig. 6. The low frequency components show long duration fluctuations, while the high frequency components show short duration fluctuations.

The settings of the order for the AR models in all cases was established in $P = 12$, which is due to the information delivered by the Fast Fourier Transform Algorithm (Hahn and Valentine, 2013), showing relevant periods of 12 years at 5% significance level. The inputs of the AR model are regressor matrices formed with the values of the low and high frequency components. The coefficients were computed with the training sample composed by the 70% of data. The results of the one-step ahead forecast of CO₂ per capita emissions by means of all combined models MA-AR, SSA-AR, and MSVD-AR are presented in Table 2. Fig. 7 presents the curves of the observed data versus the predicted data via MA-AR and testing sample. Fig. 8 presents the curves of the observed data versus the predicted data via SSA-AR and testing sample.

From Table 2, the best results were reached through the combined model MSVD-AR. The MSVD-AR model is effective for all the analyzed time series. The average MAPE is of 0.00745 and the average efficiency (mNSE) is of 99.9%. The second best model is SSA-AR with an average MAPE of 0.6405% and a mNSE of 94.22%, the

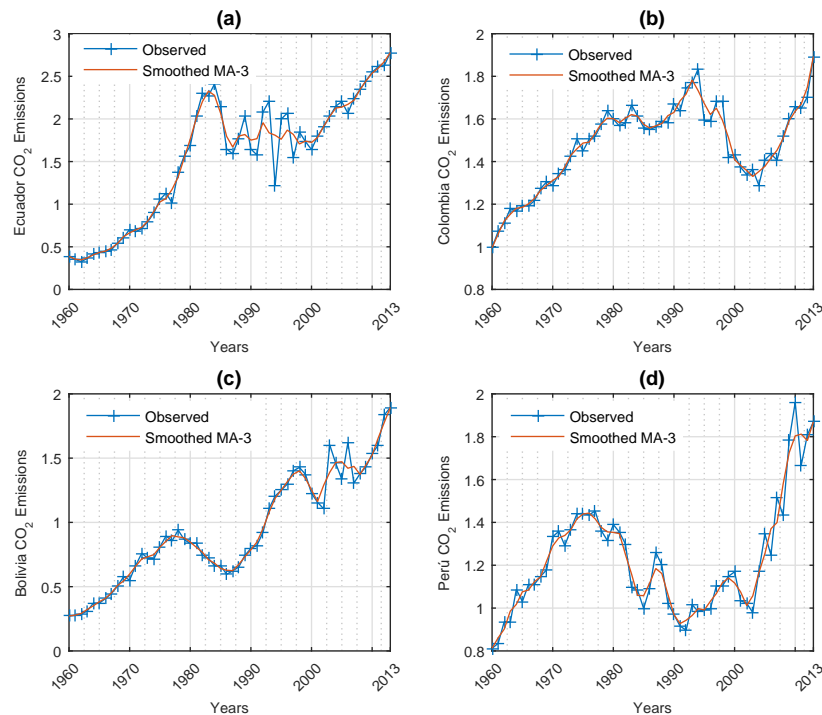


Fig. 3: Smoothing via Moving Average (3)

lowest accuracy was observed through MA-AR with an average MAPE of 3.4606 and a mNSE of 67.13%.

The best model MSVD-AR was used for multistep-ahead forecasting. The MIMO strategy was implemented and the results are presented in Table 3.

MSVD-MIMO presents good performance for all series (Table 3). The highest accuracy by means of MSVD-ARMIMO was obtained in the multi-step ahead forecasting of Perú emissions, with a mean MAPE of 0.1089%, and a mNSE of 99.3%. The lowest accuracy was obtained for the time series of Ecuador with a mean MAPE of 0.4082% and a mean mNSE of 94.3%.

The observed and predicted values via MSVD-ARMIMO combined model and the projections for 8-years ahead forecasting are presented in Figs. 9, 10, 11, and 12. From Figures, a good fit was achieved among the observed data and the predicted data during the data collection period for all countries.

6 Conclusions

In this work it was presented the forecast of CO₂ per capita emissions of four countries with similar conditions in terms of geographic conditions and cultural diversity

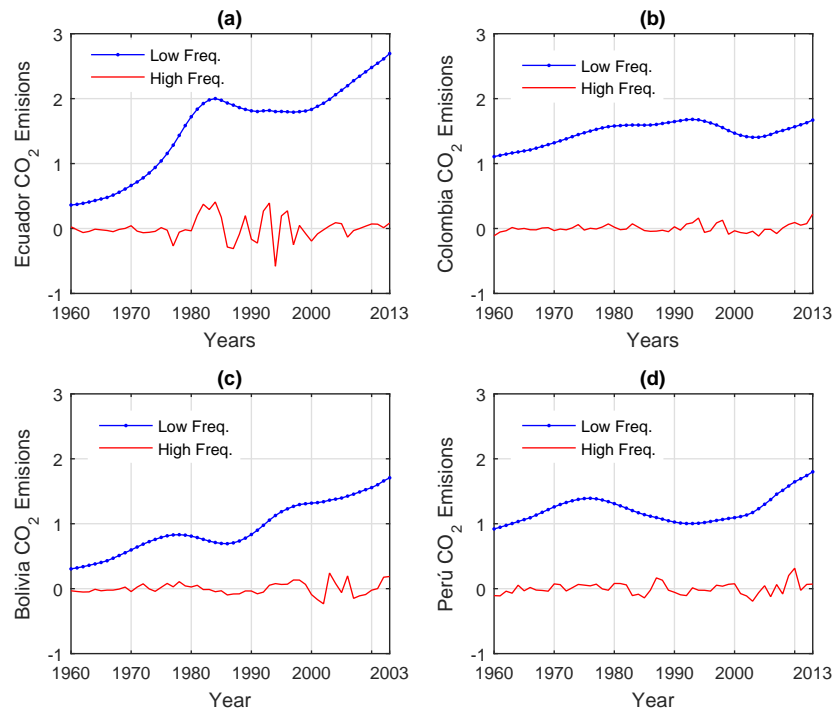


Fig. 4: Singular Spectrum Analysis Decomposition

Table 3: Multi-step ahead forecasting Results via MSVD-ARMIMO

Horizonte	Ecuador		Colombia		Bolivia		Perú	
	MAPE %	mNSE %	MAPE %	mNSE %	MAPE %	mNSE %	MAPE %	mNSE %
1	5.7e-05	99.9	2.972e-05	99.9	3.45e-05	99.9	1.57e-05	99.9
2	0.0007	99.9	0.0003	99.9	0.0004	99.9	0.0002	99.9
3	0.0048	99.9	0.0018	99.9	0.0029	99.9	0.0013	99.9
4	0.0233	99.7	0.0091	99.7	0.0139	99.9	0.0064	99.9
5	0.0894	98.9	0.0360	99.0	0.0544	99.5	0.0243	99.8
6	0.2813	95.9	0.1195	97.5	0.1781	98.2	0.0795	99.5
7	0.7972	87.6	0.3507	94.6	0.5159	94.5	0.2115	98.8
8	2.0687	72.5	0.9067	88.2	1.2243	83.6	0.5483	96.8
Min	5.7e-05	72.5	2.972e-05	82.2	3.45e-05	83.6	1.57e-05	96.8
Max	2.0687	99.9	0.9067	99.9	1.2243	99.9	0.5483	99.9
Mean	0.4082	94.3	0.1780	97.3	0.2487	96.9	0.1089	99.3

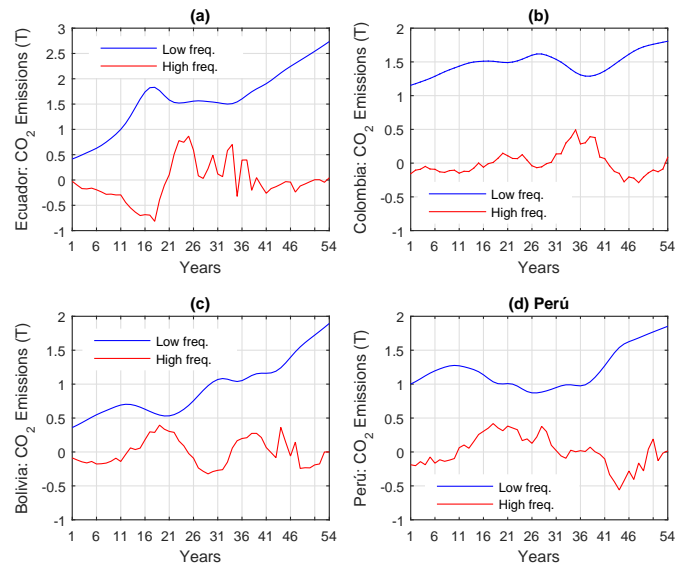


Fig. 5: Multilevel Singular Value Decomposition

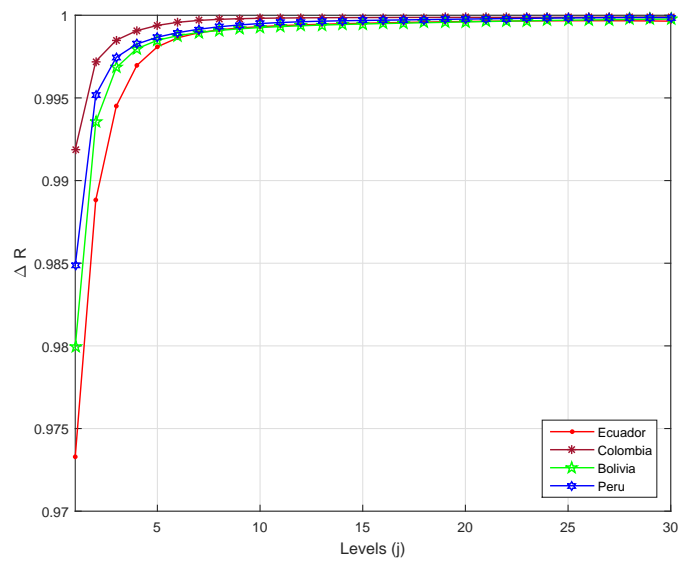


Fig. 6: Singular Spectrum Rate

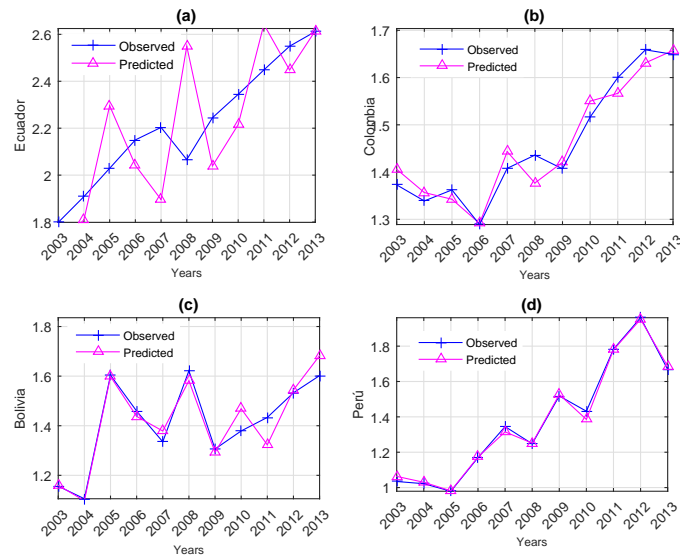


Fig. 7: One-step ahead prediction curves based on MA-AR, for testing sample

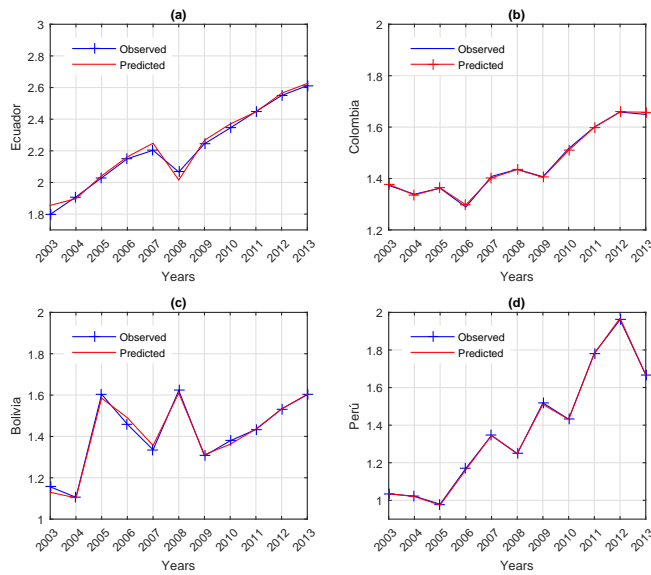


Fig. 8: One-step ahead prediction curves based on SSA-AR, for testing sample

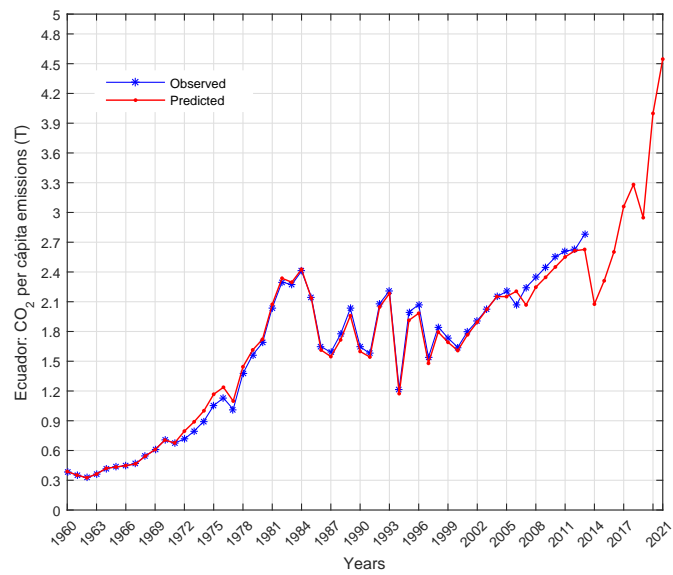


Fig. 9: Ecuador CO₂ per capita emissions (T)

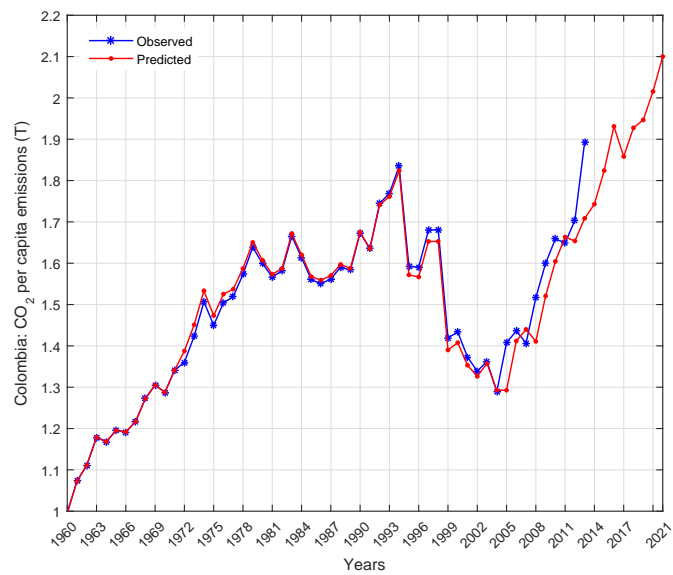


Fig. 10: Colombia CO₂ per capita emissions (T)

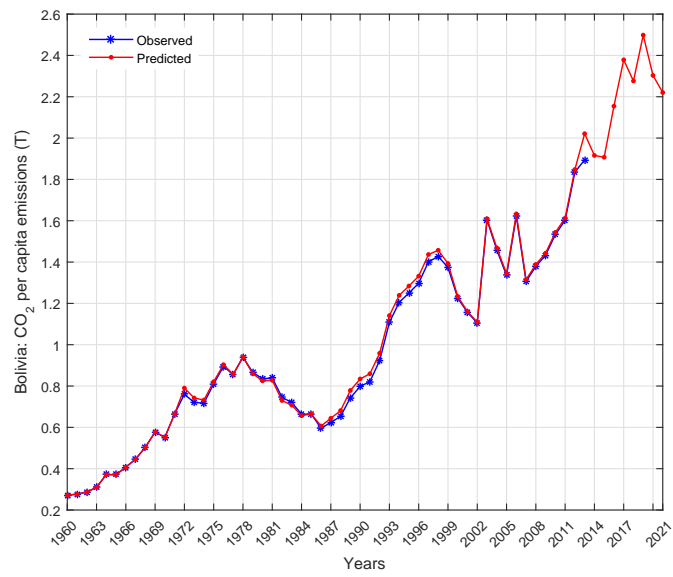


Fig. 11: Bolivia CO₂ per capita emissions (T)

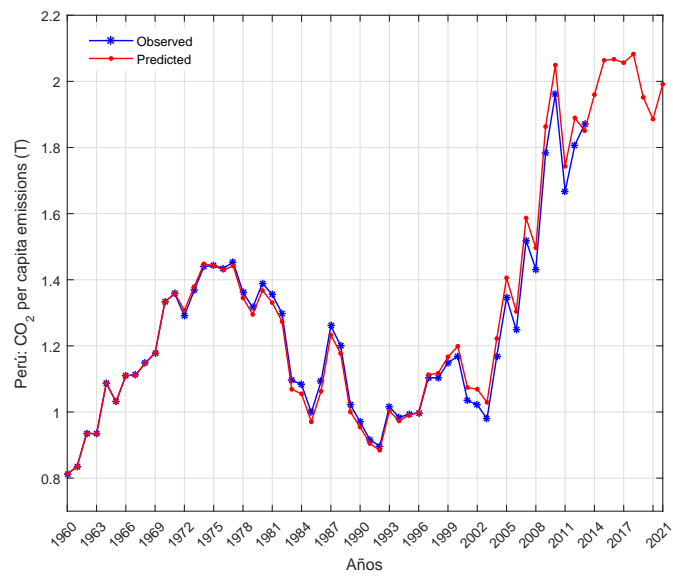


Fig. 12: Perú CO₂ per capita emissions (T)

(Ecuador, Colombia, Bolivia, and Perú). The forecasting methodology combines data preprocessing techniques and the Autoregressive model. The method was evaluated through 54-year observations from 1960 to 2013, the sample was separated into two groups, one experimental (training) with 70% of samples and the other group of testing with the remaining 30% of samples.

The results obtained with the testing sample demonstrated that the MSVD-AR model improves significantly the accuracy of the simple AR model in comparison with smoothing by Moving Average (MA-AR) and the combined model SSA-AR. Furthermore, the best model MSVD-AR extended to multiple horizon forecast via MIMO strategy (MSVA-ARMIMO) has shown improved accuracy level with respect to other approaches observed in the literature review. The average accuracy achieved for 8-years ahead forecasting via testing sample was of 0.1089% for MAPE, and 99.3% for mNSE. The projections for nine years from 2014 to 2021 shown that Ecuador, Colombia, Bolivia and Peru, will increase the level of CO₂ per capita emissions. According to these projections, Ecuador will be the country that would present the largest amount of emissions in comparison with the rest of CAN countries.

The results obtained could be a reference for public and private institutions to be observed and incorporated into their work plans for the care and preservation of the environment.

Given the effectiveness of the method, new forecasting simulations will be performed with time series coming from other countries and other areas of knowledge.

7 Acknowledgments

Thanks to Animal Production and Industrialization (PROANIN) Research Group of the Universidad Nacional de Chimborazo for supporting this work through the project Artificial Neural Networks to predict the carcass tissue composition of guinea pigs.

References

1. World Bank Group repository, <http://databank.worldbank.org/data/home.aspx>, (2017)
2. Kimball, B. A., Pinter Jr., P.J., Garcia, R.L., LaMorte, R.L., Wall, G.W., Hunsaker, D.J., Wechsung, G., Wechsung, F., Kartschall, T.: Productivity and water use of wheat under free-air CO₂ enrichment. *Global Change Biology*. 1(6): pp. 429–442 (1995)
3. Tao, F., Feng, Z., Tang, H., Chen, Y., Kobayashi, Z.: Effects of climate change, CO₂ and O₃ on wheat productivity in Eastern China, singly and in combination. *Atmospheric Environment*. 153: pp. 182–193 (2017)
4. Pao, H-T., Tsai C-M.: Modeling and forecasting the CO₂ emissions, energy consumption, and economic growth in Brazil. *Energy*. 36: pp. 2450–2458
5. Pérez-Suárez, R., López-Menéndez, A.: Growing green? Forecasting CO₂ emissions with Environmental Kuznets Curves and Logistic Growth Models. *Environmental Science & Policy*. 54: pp. 428–437 (2015)
6. Wu L., Liu S., Liu, D., Fang, Z., Xu, H.: Modelling and forecasting CO₂ emissions in the BRICS (Brazil, Russia, India, China, and South Africa) countries using a novel multi-variable grey model. *Energy*. 79: pp. 489–495 (2015)

7. Broomhead, D.S, King, G.: Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*. 20(2): 217 – 236 (1986)
8. Barba, L., Rodríguez, N.: A Novel Multilevel-SVD Method to Improve Multistep Ahead Forecasting in Traffic Accidents Domain. *Computational Intelligence and Neuroscience*. Volume 2017, Article ID 7951395, 12 pages (2017)
9. Grossmann, A., Morlet, J.: Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.* 15(4), pp. 723–736 (1984)
10. Shensa, M.: The Discrete Wavelet Transform: Wedding the A Trous and Mallat Algorithms. *IEEE Transactions on Signal Processing*. 40(10), pp. 2464–2482 (1992)
11. Nason, G., Silverman, B.: Wavelets and Statistics, The Stationary Wavelet Transform and some Statistical Applications. *Wavelets and Statistics*. Springer New York. pp. 281–299 (1995)
12. Mallat, S.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on pattern analysis and machine intelligence*. 11(7) pp. 674–693 (1989)
13. Hornik K., Stinchcombe X., White H.: Multilayer feedforward networks are universal approximators. *Neural Networks*. 2(5), pp. 359–366 (1989)
14. Svozil D., Kvasnicka V., Pospichal J.: Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*. 39(1), pp. 43–62 (1997)
15. Rojas I., Pomares H., Bernier J.L., Ortega J., Pino B., Pelayo F.J., Prieto A.: Time series analysis using normalized PG-RBF network with regression weights. *Neurocomputing*. 42(14), pp. 267–285 (2002)
16. Roh S.B., Oh S.K., Pedrycz W.: Design of fuzzy radial basis function-based polynomial neural networks. *Fuzzy Sets and Systems*. 185(1), pp. 15–37 (2011)
17. Chattopadhyay G., Chattopadhyay S.: Autoregressive forecast of monthly total ozone concentration: A neurocomputing approach. *Computers & Geosciences*. 35(9), pp. 1925–1932 (2009)
18. Maali Y., Al-Jumaily A.: Multi Neural Networks Investigation based Sleep Apnea Prediction. *Procedia Computer Science*. 24, pp. 97–102 (2013)
19. Liu F., Ng G.S., Quek C.: RLDD: A novel reinforcement learning-based dimension and delay estimator for neural networks in time series prediction. *Neurocomputing*. 70(79), pp. 1331–1341 (2007)
20. Scarselli F., Chung A.: Universal Approximation Using Feedforward Neural Networks: A Survey of Some Existing Methods, and Some New Results. *Neural Networks*. 11(1), pp. 15–37 (1998)
21. Gheyas I.A., Smith L.S.: A novel neural network ensemble architecture for time series forecasting. *Neurocomputing*. 74(18), pp. 3855–3864 (2011)
22. Yafee R., McGee, M.: *An Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS*. Academic Press, pp. 528 (2000)
23. Golyandina, N., Nekrutkin, V., Zhigljavsky, A.: *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC. pp. 15 - 44 (2001)
24. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathematics*. 2(2) pp. 164–168 (1944)
25. Marquardt, D.: An algorithm for least - squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*. 11(2) pp. 431–441 (1963)
26. Hahn, B., Valentine, D.: *Essential MATLAB for Engineers and Scientists*, Sixth Edition. Academic Press, Elsevier. pp. 333–339 (2013)

Analysis of Buildings Energy Losses Using Smart Monitoring

ATTOUE Nivine ^{a, b}, SHAHROUR Isam^a, YOUNES Rafic^b, ALJER Ammar^a,
LORIOT Marine^a

^a Laboratory of civil engineering and geo-environment, Lille University,
59650 Villeneuve d'Ascq, France

^b Modeling Center, Lebanese University, Lebanon

nivine.attoue@gmail.com, isam.shahrou@univ-lille1.fr, ryounes @
ul.edu.lb, ammar.aljer@univ-lille1.fr, loriot.marine@gmail.com

1 Introduction

With urban growth, the city's energy consumption continues to increase. In France, the building sector is the largest consumer of energy with a share of about 45% of which 73% for heating and air conditioning. This large consumption contributes to the increase in the greenhouse gases emission. Hence, France has committed to reduce the energy consumption by 20% and the greenhouse gases emissions by 40% by 2030. In order to reach these goals, we need a better understanding of the building thermal behavior through exploring sources of energy gain and losses. [1] and [2]. The main objective of this paper is to study these factors in the buildings sector using smart monitoring. The paper presents analysis of series of thermal tests conducted in an occupied office in different users' condition. Analysis shows the influence of the user behavior on energy consumption.

2 Methodology

The study is conducted in an occupied office room in the school of engineering Polytech'Lille in the North of France. One-month measurement series were recorded with intensive monitoring of both the temperature and humidity. Analysis of these tests allowed us to understand the temperature repartition in the room and to reduce the monitoring system. Then, tests were conducted with different usage condition to explore the influence of these condition on the energy consumption.

3 Monitoring System

Understanding the real conditions inside the building requires monitoring the indoor environment. Hence, a new monitoring system was designed to reduce the energy consumption and the monitoring system cost. It is composed of a central unit, wireless sensors and friendly users' interface.

The central unit with a free and open software communicates with sensors using radio frequency (RF) protocol ensuring the management of the monitoring system. It is formed of a small computer without screen or keyboard, a 'Raspberry Pi', which hosts the free and open source Linux operating system for data storage, analysis and display.

A local Wi-Fi network is created by this unit enabling the access to data and information stored.

Several parameters were tracked at a chosen time intervals using the wireless sensors that are connected to the central unit. The main function of these sensors is the pursuit of indoor comfort parameters (temperature, humidity and lighting) and the control of doors/windows (open or closed). These parameters are monitored in a multi-parameters smart card and sent using one communication system. Sensors used in our experimentation are associated with PanStamp and Inodesign programmable modules.

A web friendly interface was designed to enable users to access easily to all the information concerning the indoor environment [3] and [4].

4 Experimental program

The experimental study was conducted in an occupied room in building D at the first floor in Polytech'Lille. At first, around 90 sensors were installed to follow the thermal condition inside the room. Some were placed at the same location to explore the reliability of the monitoring system, others were installed at the three walls, façade, at the center (air) and outside (exterior parameters). The façade is formed of well insulated, two double glazing windows. The left wall, adjacent to the facade, was equipped by three levels of sensors (top, middle and bottom) and by three other spots, for each level, each one with a certain distance from the facade (nine sensors at this wall in total). Fig.1 illustrate this monitoring system.

After one month of monitoring, a database was built with a time series measurements having an interval of five minutes. Analyzing these preliminary allowed us to understand the building thermal behavior. Then, two sets of scenarios were executed to study the influence of usage conditions on energy consumption. An air conditioner with an adjustable temperature and a power meter was used during these experimentations. The first set of testing was executed on a temperature of 17°C, the other one on 20°C. Each set of scenario consists of one day of cooling the room at a certain temperature, one day with opened windows and cooling and another one with closed stores and cooling.

Height : 238 cm	THL_left_Top_01 00060B4F14FFFFFF		THL_Left_Top_02 00060A1514FFFFFF		THL_Left_Top_03 0006153B14FFFFFF	
Height : 144 cm	THL_Left_Middle_01 0006150C14FFFFFF		THL_left_Middle_02_2 000616E014FFFFFF	THL_left_Middle_02_1 000616E514FFFFFF	THL_left_Middle_02_3 000616E814FFFFFF	THL_left_Middle_03 000609C414FFFFFF
Height : 50 cm	THL_left_Bottom_01 0006153914FFFFFF		THL_left_Bottom_02 00060A1614FFFFFF		THL_left_Bottom_03 000609C714FFFFFF	

Fig. 1. Monitoring plan for the left wall.

5 Analysis

5.1 Preliminary tests

The reliability of the monitoring system was checked by the comparison of data recorded by sensors located at the same location. These tests showed that data recorded by sensors located at the same location are very closed (Figure). These results confirmed the reliability of the monitoring system. Analyses were then conducted to study the variation of the temperature and humidity in the room in normal operating conditions. They showed that the facade was the most influenced by the outside with a difference of 2°C in average compared to the wall temperature which has the least impact from exterior as we can see in the figure2. The average difference between the wall and the air temperatures was 1°C.

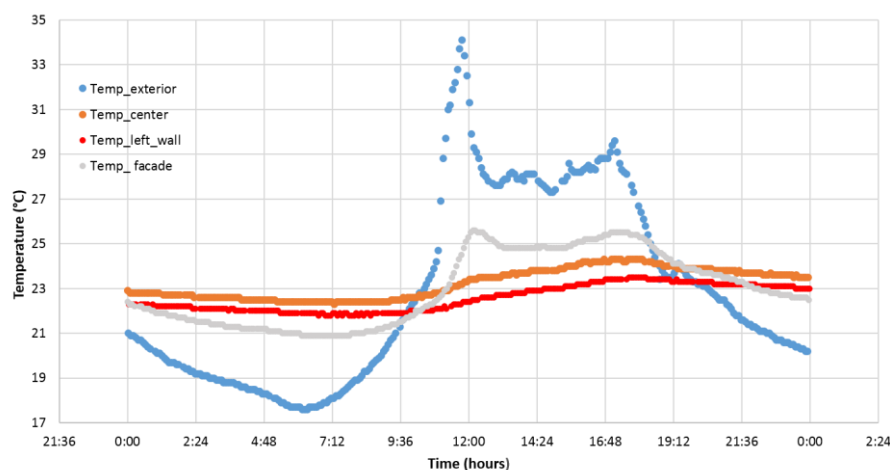


Fig. 2. Temperature variation for wall, facade and air.

Comparison of the different measurements data with varying the distance of the sensors from the facade points to, as before, by approaching the façade the measurements are more influenced by the outside conditions.

Later, comparing the temperature of different levels indicates that the temperature increases by 0.4°C in average from bottom to top at three different positions. Then by comparing humidity measurements, a difference of 4% in average between top and bottom was found.

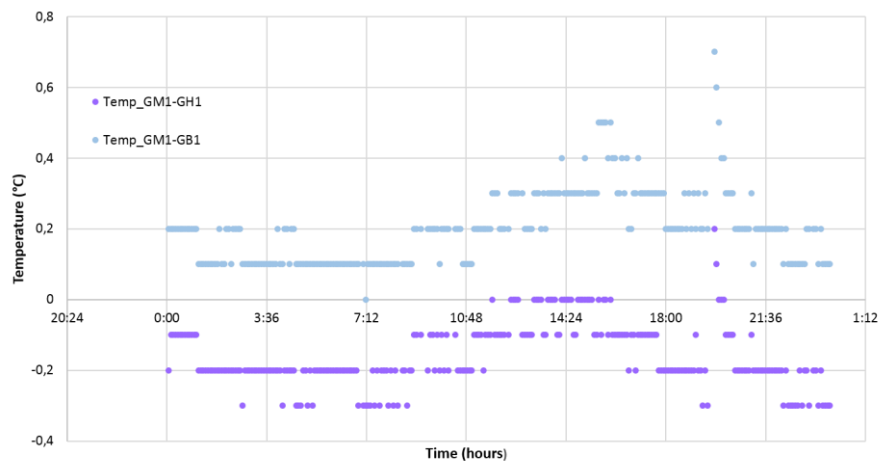


Fig. 3. Temperature variation for different levels.

Due to the variation of these parameters from spot to another in the room, our study will focus on the facade where the heating exchange occurs the most. All the thermal analysis done here after are deliberated in function of the facade parameters.

5.2 Analysis of experimental scenarios

Two sets of scenarios were executed to study the impact of occupants' behavior. For the first one, an air conditioner was launched at 17°C for four days. We opened the windows for 24h, then analyzed the consumption needed. By comparing the energy consumption for closed and opened windows (Figure 4), we noticed an increase by 33% in average. This is illustrated by figure 5 where the energy for the two scenarios were represented with the sum of the exterior temperature.

Afterwards, we closed the curtains for 24h and observed the evolution of the consumption. We noticed an increase by 28% in average when the curtains were closed. This is represented by figure 6.

We repeated the same experimentations with the air conditioner launched at 20°C . The energy consumed was 5 times less than the one consumed at 17°C . When opening the windows, we noticed that the consumption increased by 50% in average.

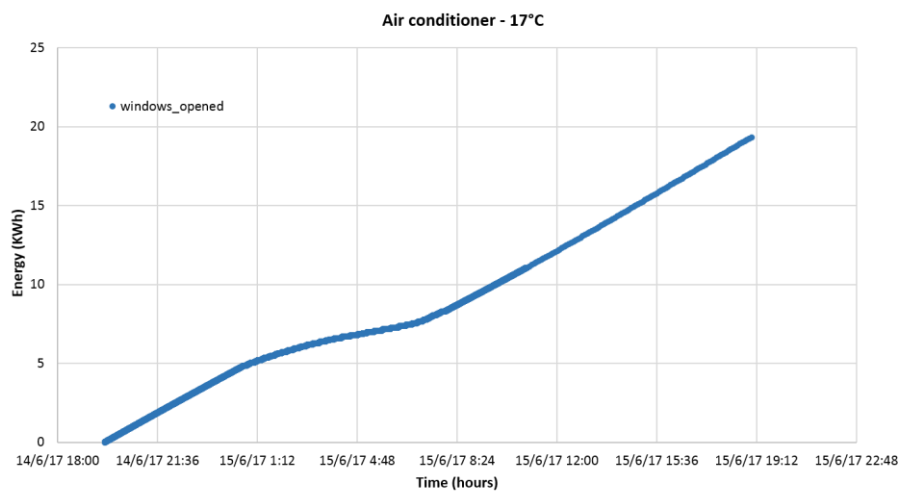
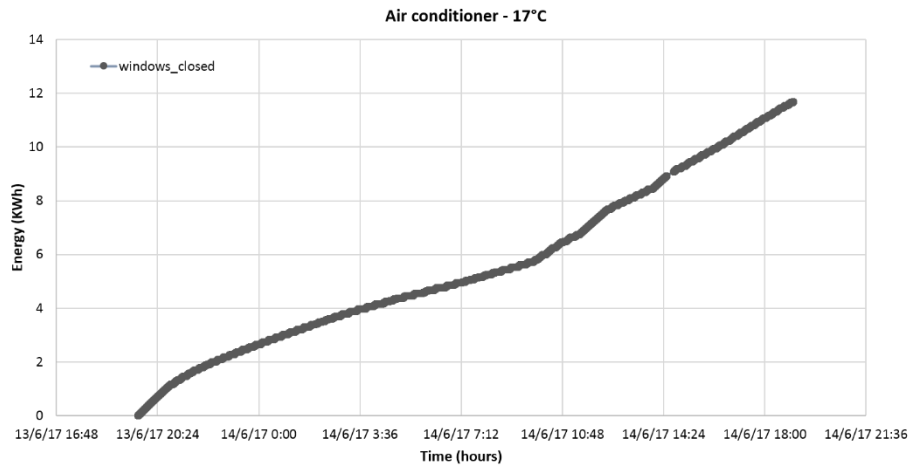


Fig. 4. Energy consumption for opened and closed windows with time.

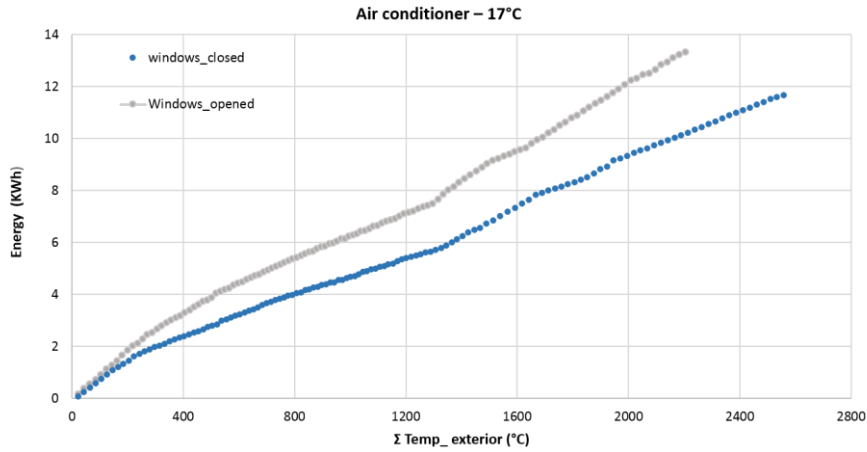


Fig. 5. Energy consumption in cases of opened and closed windows with exterior temperature.

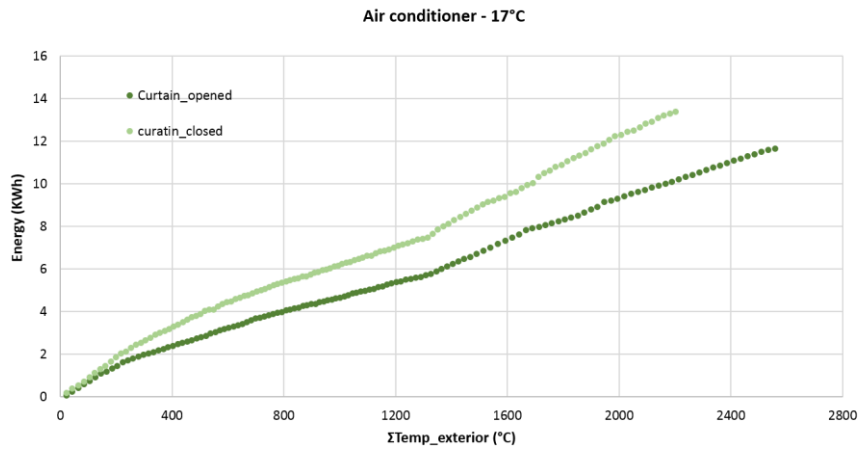


Fig. 6. Energy consumption in cases of opened and closed curtains with exterior temperature.

6 Conclusion

This paper included an experimental analysis of the influence of the usage condition on the building energy consumption using an advanced monitoring system. A preliminary study allowed the optimization of the monitoring system by focusing the monitoring system on the external wall.

Analysis of different usage conditions showed that the energy consumption is largely influenced by the window's opening, the interior operating temperature and the use of stores. The work continues to develop a prediction model that provided the energy consumption in function of the operating conditions.

Acknowledgement. This research received funding from French University Agency (AUF), the Lebanese National Council for Scientific Research CNRS-L and the University of Lille.

7 References.

1. Agence de l'environnement et de la maitrise d'énergie (ADEME), <http://www.ademe.fr/climat-air-energie-edition-2015>.
2. T. Hong, S. C. Taylor-Lange, S. D'Oca, D. Yan, S. P. Corgnati: Advances in research and applications of energy-related occupant behavior in buildings. *Energy and Buildings* 116 (2016) 694–702
3. SEMTECH Advanced communications and sensing, www.semtech.com
4. A. Janssens: Reliable building energy performance characterization based on full scale dynamic measurements. International Energy Agency, Technical report, EBC Annex 58

Forecasting UK House Prices During Turbulent Periods*

Alisa Yusupova[†] and Efthymios G. Pavlidis

Department of Economics, Lancaster University Management School, UK

Abstract

In this paper, we provide an extensive investigation of the ability of a battery of static and dynamic econometric models to forecast UK national and regional housing prices over the last two decades. Our results suggest that, due to changes in the set of predictive variables that drove UK house prices during the upturn and downturn in real estate markets in the 2000s, methods that allow both the underlying model specification and parameter estimates to vary over time produce more accurate out-of-sample forecasts than methods where the number of predictors is kept fixed. Furthermore, we find i) that parsimonious forecasting models perform better than models which include a large number of predictive variables, and ii) that there is no single variable that uniformly outperforms all others in terms of predictive power. A credit availability indicator, however, appears to be the main driver of UK house prices during the boom phase of the 2000s.

Keywords: Real House Price Growth; Regional UK Markets; Forecasting; Autoregressive Distributed Lag Models; Dynamic Models; Structural Housing Models

JEL Classification: C22; C53; E37

1. Introduction

The latest boom and bust episode in housing markets and its decisive role in the Great Recession has generated a vast interest in the dynamics of house prices, and has led many international organizations, central banks and research institutes to closely monitor developments in housing markets across the world. For instance, the International Monetary Fund recently established the Global Housing Watch, the Globalisation and Monetary Policy Institute of the Federal Reserve Bank of Dallas initiated a project on monitoring international property price dynamics, and the UK Housing Observatory¹ initiated a similar project for the UK national and regional housing markets. At the same time, a substantial empirical literature has emerged that deals with the forecastability of house price movements (for a comprehensive survey, see Ghysels et al., 2012). Surprisingly, this literature has concentrated almost entirely on the United States, leaving national and regional markets of other countries mostly unexplored.

A country where house price forecastability is of prime importance is the United Kingdom. Similarly to the United States, in the United Kingdom, housing activities account for a large fraction of GDP and of households expenditures; real estate property comprises the largest component of private wealth (excluding private pensions) and mortgage debt the main liability of households (Office for National Statistics, 2014). Thus, abrupt movements in house prices cause

*We are grateful to Mike Clements, Ivan Paya, David Peel, Mike Tsionas, and seminar and conference participants at the 36th International Symposium on Forecasting, the 3rd KoLa Workshop on Finance and Econometrics, and the 10th International Conference on Computational and Financial Econometrics for useful comments and suggestions.

[†]Correspondence to: Alisa Yusupova, Department of Economics, Lancaster University Management School, Lancaster, LA1 4YX, UK. a.yusupova@lancaster.ac.uk.

¹The UK Housing Observatory is a project of the Economics Department at Lancaster University Management School, available at <http://www.lancaster.ac.uk/lums/economics/research/housing/>.

large variations in households' wealth and adversely affect households' ability to borrow and spend, which ultimately impacts on financial stability and the real economy - as clearly demonstrated by the financial crisis of the 2000s.²

In this paper, we contribute to the existing literature by providing an extensive investigation of the ability of a battery of econometric models to forecast UK house prices over the turbulent period from 1995:Q1 to 2012:Q4. Due to the documented heterogeneity of house-price behaviour across time and space (see, e.g., Yusupova et al., 2017, Pavlidis et al., 2016), we examine both national and regional forecastability and consider both static and dynamic econometric models. The set of forecasting models considered is mainly motivated by the studies of Rapach and Strauss (2009) and Bork and Møller (2015) on the US market. The former study focuses on Autoregressive Distributed Lag models, while the latter employs Dynamic Model Averaging (DMA) and Dynamic Model Selection (DMS). In addition to these models, we also examine forecasts generated by simple OLS regressions, Bayesian Vector Autoregressions (BVAR), Time-Varying Parameter models, Bayesian Model Averaging, and Bayesian Model Selection. Two common characteristics of all of the above models is that, first, they are a-theoretical (in the sense that they do not build on micro-foundations) and, therefore, are not internally consistent, and second, that they are fitted to a relatively small number of predictors, which implies that they do not exploit the available information in a data-rich environment such as the one of the UK housing market. Following the recent macro and forecasting literature, we extend the set of models used by Rapach and Strauss and Bork and Møller to incorporate a popular housing model, the DSGE model of Iacoviello and Neri (2010), and a factor-augmented BVAR model estimated using a large macroeconomic dataset for the UK economy.

In summary, our findings suggest that models that allow both the underlying specification and parameter estimates to vary over time, i.e. DMA and DMS, produce more (and, in some cases, dramatically more) accurate forecasts than methods where the number of predictors is kept fixed. DMS, in particular, performs remarkably well. First, it uniformly outperforms the simple linear autoregressive benchmark for the national and all the regional housing markets and, second, it captures particularly well the housing boom up to 2004 and the price collapse of 2008. The superiority of dynamic over static models is consistent with recent evidence that suggests that the relationship between real estate valuations and conditioning macro and financial variables, such as domestic credit, displayed a complex of time-varying patterns over the last decades (Aizenman and Jinjark, 2014).

Our results also demonstrate that, out of the national- and regional-level predictors considered, there is no single variable that consistently leads to significant improvements in predictive accuracy relative to the benchmark. The key drivers of house price movements appear to vary considerably across regions, over time and across forecast horizons. For volatile regions, the index of credit availability and growth in industrial production appear to be the best house price predictors, particularly during the boom phase. While, for relatively stable property markets, the stock of dwellings is the main determinant of housing dynamics on the eve of the house price collapse.

The rest of the paper is structured as follows. A description of the housing data and the property price predictors is presented in Section 2. Section 3 compares the predictive accuracy of the alternative forecasting models, evaluates the performance of these models over time, and investigates their optimal dimension and the key determinants of future house price movements. Section 4 provides concluding remarks.

2. Data

2.1. House Prices

We use quarterly mix-adjusted national and regional house price indices for the period 1975:Q1 to 2012:Q4 reported by Nationwide.³ We follow the classification of UK regions adopted by Nationwide and consider 13 regional housing markets. To transform the data into real units, we divide nominal property price indices by the Consumer Price Index

²In line with this reasoning, the Bank of England has been assessing the resilience of the UK banking system to house price shocks over the last few years, and is currently considering a potential sharp downturn in commercial and residential property prices as one of the key elements of its 2017 stress testing scenario (Bank of England, 2017).

³Nationwide is the UK based world's largest building society and one of UK's largest mortgage providers. The Nationwide database, which stretches back as far as 1952, contains data on UK national and regional house prices and housing affordability estimates. Details of the methodology used to construct regional and national property price indices as well as information about the regional composition are available from the web page of Nationwide House Price Database: <http://www.nationwide.co.uk/media/MainSite/documents/about/house-price-index/nationwide-hpi-methodology.pdf>

(all items) obtained from the OECD Database of Main Economic Indicators. In our application we use the annualised log transformation of real property price inflation calculated as

$$\Delta \ln p_{r,t} = 400 \times \ln \left(\frac{P_{r,t}}{P_{r,t-1}} \right), \quad r = 1, \dots, 14, \quad (1)$$

where $P_{r,t}$ stands for the level of the real house price index of market r (either the national or one of the 13 regional) at time t .

Table 1 presents selected descriptive statistics for the annualised housing price growth rate series over the whole sample period, as well as over the latest boom (1995:Q1-2007:Q3) and bust (2007:Q4-2012:Q4) episodes. Looking at the full-sample statistics, we observe large differences in mean growth rates across regions. The highest mean growth rates have been recorded in the metropolitan and the southern areas, in particular Greater London, where real housing price inflation was about 3% between 1975:Q1 and 2012:Q4. The midland areas showed relatively moderate house price growth: East Midlands, West Midlands, Wales and East Anglia recorded an average real property price inflation of less than 2% over the entire sample period, while the northern regions, including Yorkshire and Humberside and Northern Ireland, experienced the lowest real house price growth: 1.58% and 1.25% respectively.

[INSERT TABLE 1]

Turning to the subsample statistics (columns 6-13 of Table 1), we observe substantial differences in regional house price behaviour during the recent boom and bust periods. During the upturn in residential and commercial property prices, average house price inflation across all regional markets was 8.2%, which is nearly four times larger than the figure for the entire sample. Northern Ireland was the region with the highest housing inflation (12.2%) and the highest maximum annualised real property price growth rate (47%) over the period. In the mainland, the five southern areas (Greater London, Outer Metropolitan, Outer South East, South West and East Anglia) experienced house price growth that was on average about 20% higher than in the remaining seven regions of the country.

During the recent downturn in real estate prices, all regional markets recorded negative mean growth rates that varied from -18.2% in Northern Ireland to -3.45% in Greater London. Furthermore, for the national-level data and for a number of regional markets the full-sample minimum growth rates occurred during the recent bust (e.g., Northern Ireland, Outer Metropolitan, and Wales). We note that the property markets of metropolitan and southern areas, which rose the most during the boom phase, experienced higher mean growth rates during the downturn relative to the rest of the country. Specifically, average housing inflation across the five southern areas was -4.6%, while the corresponding statistic for the remaining areas was -7.7%. Among all regions under consideration, Northern Ireland was the most volatile property market during the out-of-sample period, followed by the North and Wales, while the real estate markets of West Midlands, East Midlands and Outer Metropolitan were relatively stable. Overall, the above statistics highlight the heterogeneity of UK real estate markets.

2.2. House Price Predictors

For each region in our sample, we consider 10 economic variables as potential predictors of future house price movements: 4 regional-level and 6 national-level predictors.⁴ The variables measured at the regional level include the price-to-income ratio, income growth, the unemployment rate, and the growth in labour force. Whilst national-level predictors consist of the real mortgage rate, the spread between yields on long-term and short-term government securities, growth in industrial production, the number of housing starts, growth in real consumption, and the index of credit conditions proposed by Fernandez-Corugelo and Muellbauer (2006). The first 9 variables have been used by Bork and Møller (2015) to forecast house price movements in the US metropolitan states. The last variable has not been employed in a forecasting context before but has been shown to be an important determinant of UK regional property price behaviour in-sample (see Yusupova et al., 2017).⁵

For evaluating the performance of the factor-augmented Bayesian VAR model, in addition to the ten predictive variables introduced above, we exploit information from a large dataset of main economic indicators, which contains

⁴All 10 predictive variables used to forecast UK house price inflation are measured at the national level.

⁵The reader is referred to the online supplementary appendix to Yusupova et al. (2017) for a detailed description of the methodology, estimation results and sources of the data.

97 macroeconomic time-series. The composition of this dataset is motivated by the works of Stock and Watson (2009), Koop and Korobilis (2009) and Koop (2013) on forecasting macroeconomic series.⁶

3. Empirical Results

We begin our empirical analysis by evaluating the performance of the battery of forecasting models⁷ relative to the AR(1) benchmark over the entire out-of-sample period, from 1995:Q1 to 2012:Q4. For doing so, we employ the test for equal predictive accuracy of nested models of Clark and West (2007). The basic idea behind this test is that, under the null hypothesis that the benchmark model is the true data generating process, the forecasts of larger models are noisy due to the estimation of parameters whose population values are zero. As a consequence, the mean squared forecast error (MSFE) of the benchmark model is on expectation smaller than that of the larger model and must be adjusted. Clark and West (2007) propose the following adjusted MSFE difference

$$\hat{f}_{j,t+h} = (y_{t+h} - \hat{y}_{AR1,t+h})^2 - \left[(y_{t+h} - \hat{y}_{jt,t+h})^2 - (\hat{y}_{AR1,t+h} - \hat{y}_{jt,t+h})^2 \right], \quad (2)$$

where y_{t+h} denotes the realised value of property price inflation at time $t + h$, $\hat{y}_{AR1,t+h}$ and $\hat{y}_{jt,t+h}$ stand for the forecasts of y_{t+h} made at time t using the AR(1) benchmark model and the candidate forecasting model j respectively, and h denotes the forecast horizon. The authors show that the distribution of the t -statistic obtained by regressing $\hat{f}_{j,t+h}$ on a constant is approximately normal in large samples. Thus, the null hypothesis can be rejected in favour of the one-sided alternative (that the candidate model is able to generate more accurate forecasts) at the 5% significance level when the statistic is greater than 1.645.

3.1. Comparison of Forecast Accuracy

Table 2 presents ratios of MSFEs of the various forecasting models to the AR(1) benchmark together with the corresponding Clark and West (2007) test statistics for the national and the 13 regional real estate markets, and for a forecast horizon of one quarter ($h = 1$).⁸ The statistics highlighted in bold correspond to rejections of the null hypothesis of equal predictive accuracy at the 5% significance level.

[INSERT TABLE 2]

It is evident from the table that the DMS^{0.95}, which allows for relatively rapid variation in both underlying model specification and coefficient estimates, is by far the best model (column 4 of Table 2). This is the only model that outperforms the benchmark for the national market, and it is also the only model that outperforms the benchmark for all regional markets. The average improvement in predictive accuracy across regions is about 16%, which is similar to that for the national market. The smallest improvement is 8.4% for Greater London, and the largest is 22.5% for Scotland and Outer South East. Comparing these results with those for the DMS^{0.99} with slow forgetting (column 2 of Table 2), we observe that the MSFE ratios increase with the value of the forgetting factors, and the number of rejections of the null hypothesis decreases dramatically, from 13 to 4. This a particularly interesting finding because it suggests that the determinants of UK property prices and their marginal effects vary considerably over time. Put it differently, UK national and regional housing markets are characterised by substantial instability. Looking at the results for the other dynamic model that allows for parameter shifts and changes in the underlying specification, the DMA, we observe that this model fails to match the predictive accuracy of the DMS. For forgetting factors $\alpha = \lambda = 0.95$ (column 5 of Table 2), it offers significant forecast gains in only 4 out of the 13 regional property markets; and, for $\alpha = \lambda = 0.99$ (column 3 of Table 2), it does not manage to outperform the benchmark in any of the housing markets under consideration.

⁶For a detailed description of all data series, information on the sources of the data and the transformation undertaken to achieve stationarity of the variables please refer to the online appendix.

⁷For a detailed description of the alternative forecasting models considered in the paper please refer to the online appendix on the authors' webpage.

⁸The four-quarters-ahead results are qualitatively similar and are not reported here. These results are available from the authors upon request.

The second and third best forecasting methods in our list are the UBVAR and the mean combination of the individual ARDL forecasts, ARDL¹ (columns 11 and 15 of Table 2). The former method generates significantly more accurate forecasts than the AR(1) in 10 regions, with an average gain in forecast accuracy of about 11%. The latter outperforms the benchmark in 8 regional markets with an average improvement of 6%. The results for individual ARDL models, reported in Table 3, suggest that the performance of ARDL¹ is, to a large extent, due to the predictive content of last period's property price inflation in neighbouring areas. Specifically, the bottom panel of Table 3 (Contiguous Regions) shows that including house price growth in contiguous regions leads to significant MSFEs reductions for all regional real estate markets but one, Outer Metropolitan. On the other hand, the evidence for the remaining core house price predictors (Table 3, National and Regional Variables) is somewhat mixed. All of the predictors offer significant gains in forecast accuracy for some regional markets, but lead to significant losses for others. For instance, the inclusion of spread, which is the best performing variable in terms of rejections of the null, improves forecast accuracy by 11.4% for Wales, but worsens forecast accuracy by nearly 13.5% for Outer Metropolitan.⁹ Thus, similarly to the findings of Rapach and Strauss (2009) for the US, there is no single predictor that consistently outperforms the benchmark for the UK housing markets. Furthermore, by comparing the individual ARDLs to each other, we observe that none of them outperforms all others for all regions. It follows that none of the ten core predictors can qualify as 'best'. Given the lack of a best predictor and the importance of last period's house price growth in contiguous regions, it is not surprising that forecasts combinations yield more accurate predictions than individual ARDL models. This result is in line with the large empirical literature on forecast combination that has emerged over the years (see, e.g., Becker and Clements, 2008, Clemen 1989, Diebold and Lopez, 1996, Fang, 2003, Hendry and Clements, 2002, Hibon and Evgeniou, 2005, Makridakis and Winkler, 1983, Rapach and Strauss, 2009, Timmermann, 2008).

[INSERT TABLE 3]

Turning to the results for the remaining forecasting models in Table 2, we observe that in general these models perform poorly. The BMA, the UBFAVAR and the 'kitchen-sink' ALL are not able to generate significantly lower MSFEs than the benchmark for any of the property markets in our sample; and the DMA with time-invariant coefficients (DMA^λ = 1), the SBFAVAR and the TVP-AR-X fail to improve forecast accuracy for all regional markets but one. This outcome is consistent with Koop and Korobilis (2012) and Bork and Møller (2015), who argue that the use of a large number of explanatory variables can cause model over-fitting and, as a result, lead to inaccurate predictions. Finally, we note that the performance of the DSGE model of Iacoviello and Neri (2010) is extremely poor (column 19 of Table 2). Among the various forecasting models, the DSGE model is ranked last in forecasting UK national house price inflation with an MSFE ratio of 1.38. This finding complements those of Gupta et al. (2011) for the US housing market. It also complements the study of Edge and Gurkaynak (2011), which shows that standard medium-scale DSGE models forecast poorly macroeconomic variables.

3.2. Time Evolution of Out-of-Sample Performance

To gain insight into the evolution of the out-of-sample performance of alternative forecasting methods, we follow Bork and Møller (2015) and compute the cumulative difference between squared predictive errors

$$CDSFE_{j,r,t} = \sum_{i=1994:Q4+h}^t (e_{AR1,r,i}^2 - e_{j,r,i}^2), \quad (3)$$

where $e_{AR1,r,i}$ and $e_{j,r,i}$ stand for the prediction errors of the benchmark AR(1) model and the j th alternative model for market r , respectively. Figure 1 plots the sum of the $CDSFE_{j,r,i}$ across all property markets in our sample against time. This statistic constitutes an overall measure of out-of-sample performance. A positive (negative) value of the summary statistic at a specific time period implies that model j produces, on average, more (less) accurate predictions of future house price inflation than the benchmark up to that period; while a positive (negative) change in the value of the summary statistic unveils periods in which the forecast accuracy of model j is superior (inferior) to that of the AR(1).

⁹Interestingly, according to the results of the individual ARDL models, Wales is the housing market with the strongest forecastability: all of the house price predictors succeeded in significantly reducing the forecast error of the benchmark model in this region.

[INSERT FIGURE 1]

Examination of Figure 1 reveals the superiority of the DMS^{0.95} over the other forecasting models for $h = 1$. We observe that the DMS^{0.95} consistently produces more accurate forecasts than the other predictive techniques during the first decade of the recent house price boom, from the start of the evaluation period until around 2004. Furthermore, the DMS^{0.95} is doing remarkably well in capturing the property price downturn of 2008. On the other hand, the predictive power of the model, as that of all other models, drops in 2004:Q3. This is a period during which all regional property markets experienced a sharp reversal in house price growth rates. Annualised property price inflation figures dropped by more than 31% at the national level, by 63% in Yorkshire & Humberside, by 54% in East Midlands, and by 53% in Scotland. Similarly to 2004:Q4, all forecasting models perform poorly at the start of the bust phase and in the end of 2008 - beginning of 2009, when again all regional housing markets experienced a few consecutive quarters of negative property price inflation.

3.3. Best House Price Predictors Over Time and Across Regions

A conclusion that emerges from our empirical analysis so far is that models that include the entire set of predictive variables and do not allow for time variation in the number of predictors tend to perform poorly. Dynamic models, on the other hand, demonstrate superior predictive ability. In light of these findings two research questions that are interesting to examine are: 1) What is the optimal size of the forecasting model at each point in time? 2) Which are the most important predictors at each point in time? To answer these questions, we follow Koop and Korobilis (2012) and look at the estimated probability weights in the DMA.¹⁰

Let $Size_{k,t}$ denote the number of predictors (excluding the intercept and the lags of the dependent variable) of model k (with $k = 1, \dots, 1024$), and $\pi_{t|t-1,k}$ stand for the probability that model k should be used for forecasting at time t . Then the expected number of predictors used to construct the DMA forecast at time t is given by

$$E(Size_t) = \sum_{k=1}^K \pi_{t|t-1,k} Size_{k,t}. \quad (4)$$

Figure 2 plots the median $E(Size_t)$ across regional markets for the DMA with $\alpha = \lambda = 0.95$. It also plots the corresponding 16th and 84th percentiles to provide a measure of regional variation in $E(Size_t)$. By looking at Figure 2, we observe, first, that the median value of $E(Size_t)$ hovers around 4 predictors throughout the sample period and, second, that the 16th and 84th percentile band is narrow, which implies that there is small regional variation. Thus, the DMA results advocate parsimonious forecasting models for all regional markets when the forecast horizon is short.

[INSERT FIGURE 2]

Turning to the second question, the DMA probability weights, $\pi_{t|t-1,k}$, can also be used to cast light on which variables are important for predicting future property price movements, and examine how the best house predictors vary over time, across regional markets, and across forecast horizons. Following Koop and Korobilis (2012), for each predictor in our dataset, we scan through the 1024 models of the DMA and select those, which contain that specific predictor. The probability that the DMA assigns to these models is called the posterior inclusion probability, and reflects the importance of including the predictor in forecasting.

[INSERT FIGURE 3]

Figure 3 plots the median, the 16th percentile and 84th percentile of the estimated inclusion probabilities for one-quarter-ahead horizon. The figure illustrates in a clear manner why predictive methods that allow changes in the underlying model specification tend to produce more accurate forecasts than methods that keep the set of predictors fixed. With the exception of the price-to-income ratio, the mortgage rate and the spread (for which inclusion probabilities are virtually constant over time and across regions), it is not the same set of predictors that is driving house

¹⁰Alternatively, one could employ the DMS model size and predictors. The DMS approach yields qualitatively similar results, which are not reported here but are available from the authors upon request.

prices during periods of upswing and downturn, and across real estate markets. The prime example is the indicator of credit availability. According to the results displayed in Figure 3, the median posterior inclusion probability for this predictor increased from around 40% at the start of the sample period to above 80% during the boom phase of 2000s, and then collapsed back to their original levels. It is worth noting that, for some regional markets, the posterior inclusion probability for this variable reached almost 100% during the first part of 2000s.

3.4. Best House Predictors and the Volatility of Housing Markets

As a final exercise, we investigate how the set of house price predictors varies with the volatility of regional real estate markets. Figure 4 shows the posterior probabilities of including various house price predictors in the DMA forecasting exercise for each of the three most volatile and three most stable property markets in our sample. Diagrams on the left display the posterior inclusion probabilities for Northern Ireland, the North and Wales. Diagrams on the right show which predictors are important in the relatively stable property markets of West Midlands, Outer Metropolitan and East Midlands. For illustration purposes, we chose to report the three most important information variables for each regional real estate market, classifying a predictor with the highest posterior inclusion probability during the out-of-sample period as important.

[INSERT FIGURE 4]

Looking at the results, displayed in Figure 4, we cannot identify a single predictive variable that consistently has a higher inclusion probability. In two out of the three volatile regions (Wales and the North), credit availability is the key house price predictor during the recent boom. The probability of including the index of credit conditions increases from around 0.4 in 1995 to almost unity in 2004, reflecting the important role of credit liberalization in house price changes during this period. Growth in industrial production is another variable that plays an important role in predicting future house price movements in two volatile property markets, Northern Ireland and the North. Interestingly, the probability of including this variable in the forecasting model of Northern Ireland, which is the most volatile region in our sample, is around 0.8 for most of the evaluation period: from the mid-90s until the end of 2006. The predictive ability of industrial production starts to decline in the first quarter of 2007, when house price inflation started to slow down, and falls to 0.5 by 2009, following the sharp downturn in property prices in the area. Lastly, we observe that the mortgage rate is another important predictor of house price inflation in Northern Ireland and the North, however the probability of including this variable in the forecasting exercise is only marginally above 0.5.

Moving on to the graphs for the stable housing markets of West Midlands, Outer Metropolitan and East Midlands (Figure 4), we note that there is no single information variable that turns out to be equally important in all three markets. The gap between yields on long-term and short term gilts, the housing stock and the credit availability indicator stand out as main determinants of future house prices in two stable regional markets out of three. In West Midlands and Outer Metropolitan the spread is an important predictor during the bust phase. The two other predictors (the housing stock and the credit availability indicator), on the contrary, are important during the recent upswing in real estate prices. The index of credit availability is the key determinant of property price inflation in the midland areas at the start of the boom phase, from the first quarters of 2001 until the end of 2004. While, the stock of dwellings has the greatest predictive power in the second half of the upturn period, from the last quarters of 2004 until the collapse of property prices. Similarly to the volatile property markets, the remaining predictors show mixed predictive ability.

The main conclusion that emerges from the above empirical analysis is that there is considerable variation in the choice of house price predictors over time and across regions. This provides an explanation of why forecasting strategies that allow changes in the underlying model specification tend to produce more accurate forecasts of future house price movements.

4. Conclusion

In this paper, we provided an extensive evaluation of the forecastability of property price movements in the national and the 13 regional real estate markets of the UK over the past few decades. For doing so, we employed a rich macroeconomic dataset and a battery of static and dynamic econometric methods, including ARDL, BVAR, BFAVAR, TVP, BMA, BMS, DMA and DMS, as well as the structural DSGE model of Iacoviello and Neri (2010).

In summary, our results indicate that dynamic models that allow for changes in both the parameter estimates and the underlying model specification deliver more accurate forecasts than models in which the set of house predictors is kept fixed. Among the various models, the DMS with low values of the forgetting factors performs best in terms of forecast accuracy, outperforming the benchmark AR(1) for all real estate markets. By examining the performance of the methods over time, we found that dynamic models are doing remarkably well in capturing the upswing in real estate markets in the late 1990s - early 2000s as well as part of the latest price collapse of 2008:Q1-2008:Q3 and 2009:Q1-2009:Q3.

The estimation results of the DMA enabled us to shed light on the optimal number of predictors included in the forecasting model and on the best economic variables for predicting future house price movements. Two important conclusions with regard to model dimensionality are, first, that parsimonious models are preferred to models with a large number of predictive variables and, second, that the number of property price predictors is, generally, stable over time and across regions. With regard to the best predictors, the probabilities of including different economic variables in the forecasting exercise reveal that there does not exist a single predictor that is consistently chosen by the dynamic models as the key determinant of future property price movements. The credit availability indicator, however, was found to be a particularly important determinant of property price inflation for the majority of regional markets during the boom phase of the 2000s.

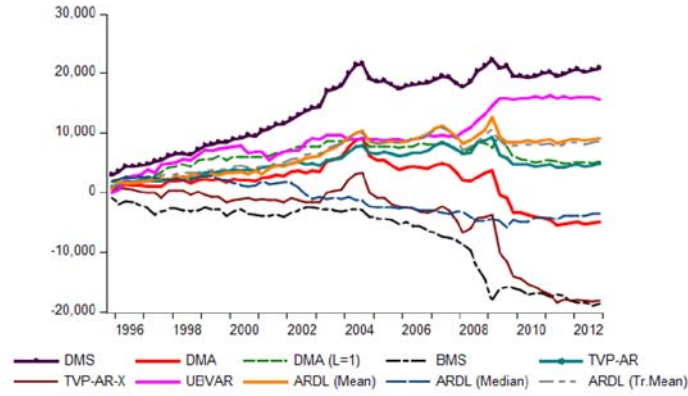
As a final exercise, we examined how the set of house price predictors varies with the volatility of regional real estate markets. By looking at the three most volatile and three most stable property markets in our sample, we found that the main predictors in volatile regions generally differ from those in stable housing markets. Our findings suggest that the index of credit availability and the growth in industrial production were the key drivers of house price inflation in volatile regions, particularly during the boom phase; while in relatively stable property markets, the stock of dwellings was the main determinant of housing dynamics on the eve of the house price collapse.

References

- Aizenman J. and Jinjark Y. (2014). Real Estate Valuation, Current Account and Credit Growth Patterns, Before and After 2008-9 Crisis. *Journal of International Money and Finance*, vol.48, pp.249-270.
- Aron J., Muellbauer J. and Murphy A. (2006). Housing Wealth, Credit Conditions and Consumption. *MPRA Paper No. 24485*, University Library of Munich, Germany.
- Bank of England (2017). *Stress Testing the UK Banking System: Key Elements of the 2017 Stress Test*. London: Bank of England.
- Bayoumi T. (1993). Financial Deregulation and Household Saving. *The Economic Journal*, vol.103(421), pp.1432-1443.
- Becker R. and Clements A.E. (2008). Are Combining Forecasts of S&P500 Volatility Statistically Superior? *International Journal of Forecasting*, vol.24, pp.122-133.
- Bork L. and Møller S.V. (2015). Forecasting House Prices in the 50 States Using Dynamic Model Averaging and Dynamics Model Selection. *International Journal of Forecasting*, vol.31, pp.63-78.
- Cameron G., Muellbauer J. and Murphy A. (2006). Was There a British House Price Bubble? Evidence from a Regional Panel. *Economics Series Working Papers 276*, University of Oxford, Department of Economics.
- Clark T.E. and West K.D. (2007). Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. *Journal of Econometrics*, vol. 138(1), pp.291-311.
- Clemen R.T. (1989). Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting*, vol.5, pp.559-583.
- Diebold F.X. and Lopez J.A. (1996). Forecast Evaluation and Combination. *National Bureau of Economic Research, Working Paper 192*.
- Fang Y. (2003). Forecasting Combination and Encompassing Tests. *International Journal of Forecasting*, vol.19, pp.87-97.
- Fernandez-Corugedo E. and Muellbauer J. (2006). Consumer Credit Conditions in the United Kingdom. *Bank of England Working Paper No. 314*.
- Edge, R. M. and Gurkaynak R. S. (2011). How Useful are Estimated DSGE Model Forecasts? *Federal Reserve Board: Finance and Economics Discussion Series 2011-11*.

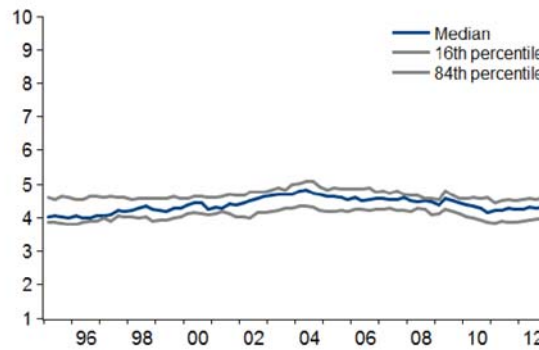
- Ghysels E., Plazzi A., Torous W. and Valkanov R. (2012). Forecasting Real Estate Prices. In G. Elliot and A. Timmermann, eds., *Handbook of Economic Forecasting, Vol II*. Elsevier.
- Gupta R., Kabundi A. and Miller S.M. (2011). Forecasting the US Real House Price Index: Structural and Non-Structural Models With and Without Fundamentals. *Economic Modelling*, vol.28, pp.2013-2021.
- Hendry D.F. and Clements M.P. (2002). Pooling of Forecasts. *Econometrics Journal*, vol.5, pp.1-26.
- Hibon M. and Evgeniou T. (2005). To Combine or Not to Combine: Selecting Among Forecasts and Their Combinations. *International Journal of Forecasting*, vol.21, pp.15-24.
- Iacoviello M. and Neri S. (2010). Housing Market Spillovers: Evidence from an Estimated DSGE Model. *American Economic Journal: Macroeconomics*, vol.2(2), pp.125-164.
- Koop G.M. (2013). Forecasting with Medium and Large Bayesian VARs. *Journal of Applied Econometrics*, vol.28(2), pp.177-203.
- Koop G.M. and Korobilis D. (2009). Bayesian Multivariate Time Series Methods for Empirical Macroeconomics. *Working Paper Series 4709*, The Rimini Centre for Economic Analysis.
- Koop G.M. and Korobilis D. (2012). Forecasting Inflation Using Dynamic Model Averaging. *International Economic Review*, vol.53(3), pp.867-886.
- Makridakis S. and Winkler R.L. (1983). Averages of Forecasts: Some Empirical Results. *Management Science*, vol.29(9), pp.987-996.
- Meen G. (1990). The Removal of Mortgage Market Constraints and the Implications for Econometric Modelling of UK House Prices. *Oxford Bulletin of Economics and Statistics*, vol.52(1), pp.1-23.
- Muellbauer J. (2002). Mortgage Credit Conditions in the UK. *Economic Outlook*, vol.26(3), pp.11-18.
- Office for National Statistics (2014). United Kingdom National Accounts, The Blue Book.
- Pavlidis E.G., Yusupova A., Paya I., Peel D.A., Martinez-Garcia E., Mack A., Grossman V. (2016). Episodes of Exuberance in Housing Markets: In Search of the Smoking Gun. *Journal of Real Estate Finance and Economics*, vol.53(4), pp.419-449.
- Rapach D.E. and Strauss J.K. (2009). Differences in Housing Price Forecastability across US States. *International Journal of Forecasting*, vol.25, pp.351-372.
- Sarno L. and Taylor M. (1998). Real Interest Rates, Liquidity Constraints and Financial Deregulation: Private Consumption Behaviour in the U.K. *Journal of Macroeconomics*, vol.20(2), pp.221-242.
- Stock J.H. and Watson M.W. (2009). Forecasting in Dynamic Factor Models Subject to Structural Instability. In: *The Methodology and Practice of Econometrics: Festschrift in Honour of D.F.Hendry*, edited by N. Shephard and J. Castle, Oxford: Oxford University Press, pp.1-57.
- Timmermann A. (2008). Elusive Return Predictability. *International Journal of Forecasting*, vol.24(1), pp.1-18.
- Yusupova A., Pavlidis E.G., Paya I., Peel D.A. (2017). Exuberance in the UK Housing Markets. *Lancaster University Management School, Working Paper, Paper No. 2017/012*.

Figure 1: Out-of-Sample Performance Over Time



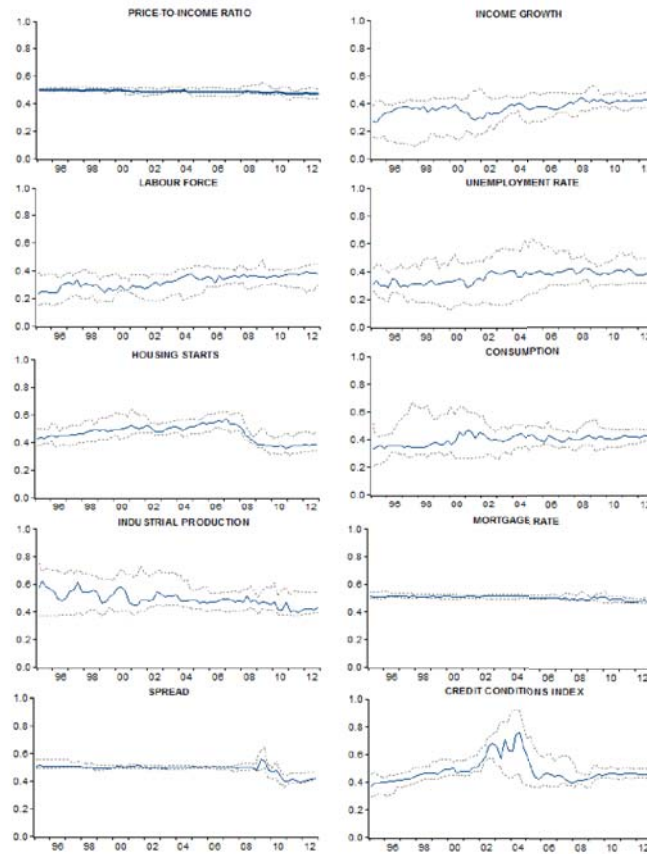
Notes: The figure plots the cumulative difference between the squared forecast errors of the AR(1) model and each of the rival models j , $\sum_{r=1}^{13} CDSFE_{j,r,t}$, where r is a regional index, for the forecast horizon $h = 1$. For illustration purposes, we only show $CDSFE$ statistics for the DMA and DMS models with $\alpha = \lambda = 0.95$, and we omit the results for models that fail to beat the benchmark in at least one regional real estate market (see Table 2).

Figure 2: Expected Model Dimension



Notes: The figure shows the time evolution of the expected number of house price predictors used to generate the $DMA^{0.95}$ h -step-ahead forecasts: $E(Size_t) = \sum_{k=1}^K \pi_{t|t-1,k} Size_{k,t}$, where $Size_{k,t} = 1, \dots, 10$ denotes the number of information variables included in model k with $k = 1, \dots, 1024$ (Koop and Korobilis, 2012). The diagram illustrates the median of the DMA expected dimension across the 13 UK regional property markets, together with the 16th and 84th percentiles.

Figure 3: Posterior Probabilities of Inclusion



Notes: The figure plots the time evolution of the posterior inclusion probability of each of the 10 core house price predictors in the DMA^{0.95} forecasting exercise. Blue solid lines show the median posterior inclusion probability across the 13 regional UK housing markets. Grey dashed lines display the corresponding 16th and 84th percentiles.

Figure 4: Posterior Probabilities of Inclusion for Volatile and Stable Property Markets

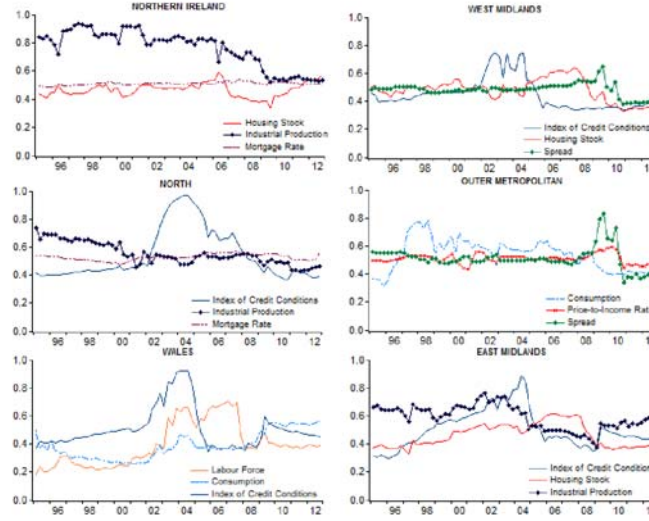


Table 1: Descriptive Statistics of Annualised Real Property Price Growth Rates

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	1975:Q1-2012:Q4				1995:Q1-2007:Q3				2007:Q4-2012:Q4			
	Mean	Std.dev	Min	Max	Mean	Std.dev	Min	Max	Mean	Std.dev	Min	Max
East Anglia	2.09	15.27	-45.78	52.74	7.99	11.59	-17.36	33.30	-5.41	12.46	-27.21	20.70
East Midlands	2.06	13.21	-28.32	59.23	7.51	10.57	-15.88	32.87	-5.69	9.72	-25.16	7.86
Greater London	3.05	14.11	-30.77	37.70	9.76	10.63	-21.00	29.96	-3.45	13.14	-24.79	17.67
N Ireland	1.25	17.33	-59.86	46.91	12.19	14.63	-17.22	46.91	-18.21	19.68	-59.86	29.67
North	1.85	13.76	-33.36	41.46	7.45	14.99	-24.32	40.99	-6.43	9.06	-26.17	10.56
North West	2.13	12.10	-27.11	37.69	7.05	10.77	-26.66	33.54	-6.69	9.68	-24.51	15.96
Outer Met	2.56	13.12	-32.12	41.72	8.25	8.86	-9.61	27.06	-4.23	12.06	-32.12	15.72
Outer S East	2.43	13.91	-30.25	40.47	8.68	9.71	-10.78	33.24	-4.92	12.30	-30.05	16.31
Scotland	1.69	11.23	-33.11	33.16	6.72	10.76	-19.31	33.16	-6.09	9.75	-26.68	14.45
South West	2.36	14.13	-37.38	57.19	8.46	9.94	-14.29	36.00	-5.26	11.57	-34.07	18.63
West Midlands	1.91	13.51	-47.44	63.29	7.27	9.29	-10.91	31.68	-5.79	9.45	-23.14	11.01
Wales	1.68	14.50	-35.91	52.31	7.39	13.52	-27.70	35.24	-6.26	13.82	-35.91	27.44
Yorks & Hside	1.58	14.19	-37.73	47.26	7.57	12.36	-13.19	39.66	-6.69	10.43	-23.24	8.93
UK	2.12	11.39	-26.98	39.08	8.19	7.99	-11.29	29.28	-5.65	10.29	-26.98	13.11

Notes: The table reports descriptive statistics of national and regional annualised log real house price growth rates over the whole sample (1975:Q1-2012:Q4), the housing boom (1995:Q1-2007:Q3) and the housing bust (2007:Q4-2012:Q4) periods. Annualised housing price inflation is computed as $\Delta \ln hp_{r,t} = 400 \times \ln \left(\frac{P_{r,t}}{P_{r,t-1}} \right)$, where $P_{r,t}$, $P_{r,t-1}$ stand for current and last period's level of Nationwide house price indices deflated by the Consumer Price Index (all items), $r = 1, \dots, 14$.

Table 2: Forecast Accuracy (forecast horizon $h = 1$)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
	DMS ^{0.99}	DMA ^{0.99}	DMS ^{0.99}	DMA ^{0.95}	DMA ^{0.95}	BMS	BMA	TVP-AR	TVP-AR-X	UBVAR	SBVAR	UBFAVAR	SBFAVAR	ARDL ¹	ARDL ²	ARDL ³	ALL	DSGE
EA	1.089 (1.379)	1.214 (0.351)	0.822 (3.586)	1.021 (1.329)	1.122 (0.812)	1.180 (1.416)	1.357 (0.657)	0.982 (1.195)	1.158 (0.649)	0.789 (3.592)	1.386 (0.78)	1.119 (-0.19)	1.230 (-0.38)	0.912 (2.377)	0.967 (1.412)	0.970 (1.428)	1.154 (-0.52)	
EM	1.042 (1.068)	1.087 (0.229)	0.800 (3.308)	1.014 (1.305)	1.031 (1.015)	1.144 (0.174)	1.169 (-0.34)	0.979 (1.294)	1.152 (0.802)	0.789 (2.879)	1.228 (0.308)	1.107 (-0.16)	1.194 (0.086)	0.929 (1.849)	1.005 (-0.06)	0.953 (1.458)	1.171 (-0.29)	
GL	1.049 (2.049)	1.087 (1.473)	0.916 (3.350)	1.123 (-0.11)	1.034 (1.433)	1.136 (1.949)	1.144 (1.916)	1.033 (0.179)	1.211 (0.562)	0.811 (3.459)	1.319 (-0.91)	1.163 (0.001)	1.212 (0.207)	0.947 (2.054)	0.971 (1.875)	1.207 (1.113)	1.105 (0.593)	
NI	1.135 (-0.19)	1.235 (-1.09)	0.853 (3.575)	0.969 (1.797)	0.989 (1.783)	1.313 (-0.86)	1.318 (-1.48)	0.841 (3.284)	0.998 (1.741)	0.989 (1.619)	1.160 (2.546)	1.008 (0.307)	1.091 (0.231)	1.018 (1.142)	1.014 (1.187)	1.061 (1.458)	1.129 (-0.43)	
NT	1.112 (-0.42)	1.132 (-1.92)	0.867 (3.311)	1.038 (1.415)	1.102 (-0.65)	1.176 (-2.03)	1.150 (-1.98)	0.973 (1.508)	1.001 (1.846)	0.813 (3.497)	1.485 (0.917)	1.103 (0.994)	1.073 (0.629)	0.896 (2.844)	0.932 (2.290)	0.927 (2.489)	1.041 (0.633)	
NW	1.232 (-0.22)	1.274 (-0.79)	0.848 (3.227)	1.162 (-0.17)	1.209 (-1.07)	1.310 (-0.15)	1.337 (-0.49)	0.989 (1.193)	1.238 (0.006)	0.959 (2.399)	1.213 (0.933)	1.064 (0.085)	1.302 (-1.09)	0.929 (2.028)	0.956 (1.679)	0.952 (1.932)	1.214 (-1.14)	
OM	1.019 (1.959)	1.329 (-1.07)	0.892 (2.525)	1.221 (0.157)	1.395 (-0.13)	1.329 (0.521)	1.571 (-1.07)	1.069 (-0.03)	1.368 (-1.09)	1.092 (1.861)	1.144 (0.384)	1.301 (-0.19)	1.362 (-0.46)	1.038 (0.016)	1.055 (-0.38)	1.132 (0.108)	1.269 (-0.57)	
OSE	1.032 (-0.22)	1.080 (-1.71)	0.775 (4.198)	0.953 (2.220)	0.999 (0.773)	1.024 (0.945)	1.059 (0.039)	0.961 (1.869)	1.051 (2.086)	1.094 (1.260)	1.216 (-0.53)	1.194 (-0.27)	1.379 (-1.47)	1.020 (0.261)	1.035 (-0.29)	1.082 (-0.49)	1.308 (-1.42)	
SC	0.945 (2.402)	0.989 (1.294)	0.775 (4.198)	0.953 (2.220)	0.999 (0.773)	1.024 (0.945)	1.059 (0.039)	0.961 (1.869)	1.051 (2.086)	0.894 (2.569)	1.544 (1.301)	1.017 (0.327)	0.964 (1.730)	0.989 (0.366)	1.021 (-1.11)	0.986 (0.497)	0.995 (1.293)	
SW	0.963 (1.909)	1.019 (1.125)	0.865 (3.185)	1.067 (1.204)	1.119 (-0.36)	0.982 (1.699)	1.099 (0.248)	1.012 (1.034)	1.119 (1.635)	0.939 (2.181)	1.024 (-1.39)	1.146 (0.057)	1.324 (-0.95)	0.947 (1.878)	0.967 (1.424)	0.974 (1.653)	1.258 (-1.17)	
WM	0.949 (1.745)	0.971 (1.339)	0.819 (3.098)	0.993 (1.798)	0.971 (1.371)	1.144 (-0.56)	1.056 (-0.05)	0.958 (1.809)	1.182 (1.298)	0.785 (3.328)	1.156 (1.769)	1.069 (0.761)	1.266 (0.194)	0.879 (2.459)	0.895 (2.271)	0.869 (2.460)	1.166 (-0.45)	
WW	0.836 (2.799)	0.945 (1.283)	0.804 (3.118)	0.999 (1.282)	0.957 (1.527)	0.895 (2.324)	0.960 (1.112)	0.982 (1.003)	1.133 (0.814)	0.751 (3.959)	1.495 (-0.76)	1.047 (0.082)	1.097 (0.196)	0.864 (3.227)	0.891 (2.944)	0.909 (2.693)	1.046 (0.615)	
YH	0.972 (1.326)	1.069 (-0.81)	0.827 (2.375)	1.039 (1.035)	1.041 (0.561)	1.025 (0.391)	1.047 (-0.29)	1.018 (0.297)	1.123 (1.288)	0.911 (2.552)	1.314 (-0.13)	1.047 (0.197)	1.054 (0.881)	0.966 (1.612)	1.009 (0.687)	1.018 (0.744)	0.978 (1.381)	
UK	1.012 (1.252)	1.059 (0.179)	0.832 (3.626)	1.084 (0.533)	1.094 (-0.62)	0.992 (1.359)	1.039 (0.608)	1.031 (-0.494)	1.203 (0.163)	1.094 (0.452)	1.116 (0.865)	1.206 (-0.53)	1.316 (-1.66)	1.013 (-1.39)	1.006 (-1.91)	1.091 (0.423)	1.206 (-1.52)	1.387 (0.160)

Notes: The table reports MSFE ratios of the various forecasting models relative to the AR(1) benchmark together with their respective Clark and West (2007) test statistics (shown in parentheses). Figures in bold indicate that the null hypothesis of equal forecast accuracy can be rejected at the 5% level of significance.

Table 3: Forecast Performance of ARDL Models (forecast horizon $h = 1$)

(1) Predictor	(2) EA	(3) EM	(4) GL	(5) NI	(6) NT	(7) NW	(8) OM	(9) OSE	(10) SC	(11) SW	(12) WM	(13) WW	(14) YH	(15) UK
<i>Regional Variables:</i>														
Income Growth	0.997 (0.924)	1.037 (-1.694)	1.042 (0.904)	1.012 (1.273)	0.923 (2.509)	0.908 (2.680)	1.101 (-0.648)	1.039 (0.466)	1.012 (0.089)	1.117 (0.504)	0.925 (1.784)	0.914 (2.569)	1.019 (0.517)	1.009 (-1.022)
Labour Force	0.974 (1.459)	1.044 (-0.224)	0.991 (1.825)	1.053 (1.033)	0.947 (2.029)	0.983 (1.433)	1.088 (-0.744)	1.021 (0.399)	1.077 (-0.329)	1.010 (0.378)	0.918 (2.001)	0.902 (2.779)	1.080 (-0.311)	1.016 (-0.548)
Unemployment Rate	1.035 (0.766)	1.105 (0.333)	0.946 (2.172)	1.075 (0.917)	1.058 (1.758)	1.129 (-0.088)	1.093 (-0.296)	1.345 (-1.882)	1.162 (-1.154)	1.004 (1.041)	1.007 (1.216)	0.918 (2.796)	1.005 (1.425)	1.042 (-1.236)
Price-to-Income	0.997 (0.924)	1.037 (-1.694)	1.053 (0.743)	1.056 (0.728)	0.939 (2.265)	0.908 (2.680)	1.098 (-1.003)	1.017 (0.156)	1.009 (-0.692)	1.112 (0.596)	0.974 (1.092)	0.914 (2.569)	1.060 (0.232)	1.009 (-1.019)
<i>National Variables:</i>														
CCI	0.999 (0.851)	1.029 (-1.119)	0.978 (1.856)	1.045 (1.080)	0.965 (1.799)	1.009 (1.024)	1.086 (-1.022)	1.047 (-0.742)	1.033 (-1.062)	1.025 (0.044)	0.954 (1.537)	0.911 (2.627)	1.031 (0.092)	1.009 (-0.927)
Consumption	0.979 (1.051)	1.025 (-1.119)	1.013 (1.413)	1.017 (1.167)	0.930 (2.397)	1.087 (0.252)	1.059 (-0.222)	1.060 (-0.261)	0.988 (0.783)	0.947 (1.622)	1.111 (1.224)	0.904 (2.641)	1.013 (0.704)	1.055 (-0.132)
Housing Stock	0.957 (2.241)	1.064 (-0.313)	0.986 (1.892)	1.030 (1.008)	0.966 (2.152)	1.026 (1.205)	1.020 (1.675)	1.116 (-1.878)	0.968 (1.215)	1.083 (-1.131)	1.109 (0.805)	0.911 (2.526)	1.012 (1.198)	1.054 (-0.603)
Industrial Production	0.997 (0.907)	1.045 (-1.628)	0.985 (1.799)	1.014 (1.221)	0.942 (2.277)	0.999 (1.092)	1.096 (-0.863)	1.059 (-1.185)	1.046 (-1.695)	1.005 (0.477)	0.956 (1.731)	0.919 (2.576)	1.039 (-0.241)	1.009 (-1.681)
Mortgage Rate	0.995 (0.918)	1.018 (-0.918)	0.941 (2.408)	1.014 (1.154)	0.942 (2.113)	1.013 (0.684)	1.088 (-0.984)	1.045 (-0.528)	1.039 (-1.427)	1.029 (1.387)	0.908 (2.033)	0.902 (2.857)	1.037 (-0.219)	1.009 (-1.106)
Spread	0.975 (1.881)	1.018 (0.868)	0.983 (2.007)	1.062 (0.696)	0.902 (2.623)	0.995 (1.497)	1.135 (-0.661)	1.044 (0.670)	1.036 (-1.431)	1.044 (1.387)	0.936 (2.238)	0.886 (2.922)	1.005 (0.659)	1.037 (-0.017)
<i>Contiguous Regions:</i>														
	1.034(EM)	0.897(YH)	0.971(EA)	0.995(SC)	0.802(YH)	1.128(NT)	1.070(OSE)	0.899(OM)	0.969(NT)	0.799(OSE)	0.961(NW)	0.857(NW)	1.032(NT)	
	(0.517)	(2.148)	(1.997)	(1.932)	(3.502)	(0.811)	(1.999)	(1.844)	(0.977)	(2.152)	(1.598)	(3.245)	(1.029)	
	0.749(OSE)	0.946(NW)	0.804(OM)		0.842(NW)	0.876(YH)	1.117(GL)	1.059(SW)	0.847(ND)	1.117(WW)	0.973(WW)	0.748(WM)	0.912(NW)	
	(3.375)	(2.358)	(3.056)		(2.949)	(2.810)	(-0.014)	(-0.904)	(3.276)	(1.191)	(2.168)	(3.443)	(2.254)	
	0.694(OM)	0.799(WM)				1.001(WW)	1.023(EA)	1.091(WM)		1.032(WM)	0.835(SW)	0.793(SW)	0.827(EM)	
	(4.145)	(2.515)				(1.669)	(0.937)	(0.370)		(0.784)	(2.868)	(3.159)	(2.697)	
	0.967(GL)	0.818(OSE)				0.926(WM)		1.089(EA)			0.855(OSE)			
	(1.757)	(2.649)				(2.538)		(-0.846)			(2.909)			
		1.019(EA)				0.909(EM)		1.069(EM)			0.787(EM)			
		(1.388)				(2.803)		(0.159)			(3.316)			

Notes: The table reports MSFE ratios of individual ARDL models relative to the AR(1) benchmark together with their respective Clark and West (2007) test statistics (shown in parentheses). Figures in bold indicate that the null hypothesis of equal forecast accuracy can be rejected at the 5% level of significance. Each ARDL model includes a constant, lags of the dependent variable and one of the predictors listed in the first column.

Impact of weather forecasting accuracy over the electric demand predictions quality

Paula Cernuda, Eduardo Caro¹, Jesús Juan
Universidad Politécnica de Madrid, Madrid, Spain

Abstract

Technical literature is rich in references pertaining to short-term electricity demand forecasting. These algorithms usually require the information of some variables such as: meteorological data, information about holidays, etc. Among all these variables, the temperature is the most relevant information from the perspective of electricity demand forecasting. This work analyzes the effect of the weather forecasting inaccuracies over the electric demand predictions. It has been performed a complete case study using real-world data from the Spanish mainland system, withdrawing statistically sound conclusions.

Keywords: electricity demand forecasting, weather predictions, forecasting accuracy.

Introduction

Technical literature is rich in references pertaining to short-term electricity demand forecasting. To compute the 24 next-day hourly forecasts, these algorithms usually require the information of some variables such as: meteorological data, information about “holidays” or “non-working days”, hour-changing days, etc. Among all the aforementioned variables, the temperature is the most relevant information from the perspective of electricity demand forecasting.

In this study, we analyze the effect of the weather forecasting accuracy over the quality of the electricity demand predictions. Specifically, the Spanish mainland system is analyzed.

Temperature forecasts

The weather forecasting database employed for this study provides 10 day-ahead temperature predictions for some geographical locations throughout Spain. The electricity demand forecasting algorithm makes use of the following locations: Madrid, Barcelona, Málaga, Bilbao, Zaragoza, Valencia, Cáceres, Murcia, Oviedo and Sevilla, as indicated in Figure 1.

¹ Corresponding author: eduardo.caro@upm.es



Figure 1. Geographical locations used by the electricity demand forecasting software.

To evaluate the numerical accuracy of the weather forecasts, the one day-ahead prediction is used as a reference, and the relative error e is computed as:

$$e = (t_{D,K} - t_{D,1}) / t_{D,1}$$

where t_{D+K} corresponds to the temperature forecasting for the day $D+K$, computed the day D . The MSE (mean squared error) is computed for the ten above-mentioned locations, from $K = 2$ to $K = 10$, and the obtained results are depicted in Figs. 2 and 3: the MSE for each horizon is plotted in Fig. 2, as well as its 95% confidence intervals; the MSE for each horizon and location is plotted in Fig. 3.

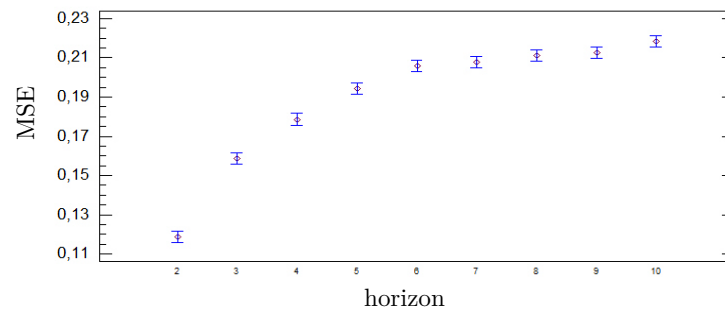


Figure 2. MSE for each horizon

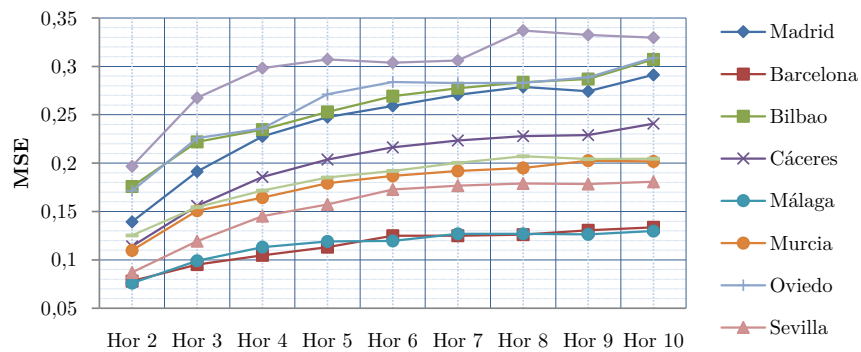


Figure 3. MSE for each horizon

From Figs. 2 and 3, it can be observed:

- Higher values of K produce higher values for MSE, i.e., the quality of the weather forecasting is impoverished for larger horizons, as expected.
- The increment of MSE is higher from lower horizons: for example, from $K = 2$ to $K = 3$, the weather forecast quality is impoverished more than from $K = 9$ to $K = 10$.
- Coastal cities (such as Barcelona or Málaga) usually exhibit higher weather forecasting accuracy.

Case Study

In this section, we have studied the influence of the weather forecasting accuracy over the demand predictions. The electricity demand forecasting software has been used, varying the weather database and observing the resulting forecasted demand results.

This forecasting software computes the 24 next-day hourly predictions at 10.00 am: i.e., ten values are computed using one-step forecast, and fourteen values are computed using two-steps forecasts.

The demand to be forecasted corresponds to the Spanish mainland electric power system, from January 1st, 2016, to December 31st, 2016. In this case study, this software has been employed using weather forecasts for the following using the following horizons: 1, 3, 5, 7 and 10. The demand forecasting quality is measured in terms of MSE, and

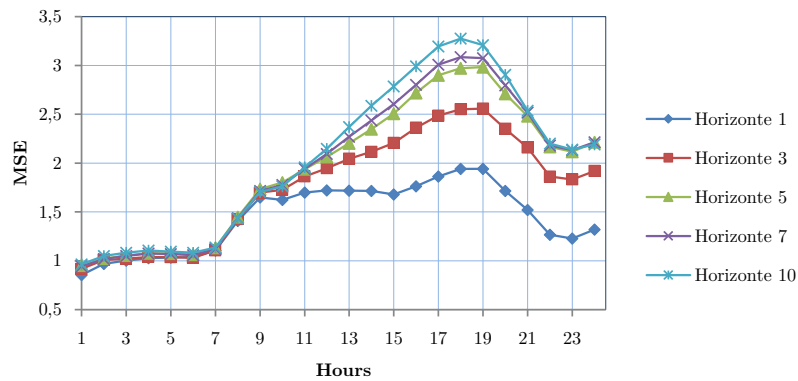


Figure 4. MSE of demand forecasts for each hour and horizon.

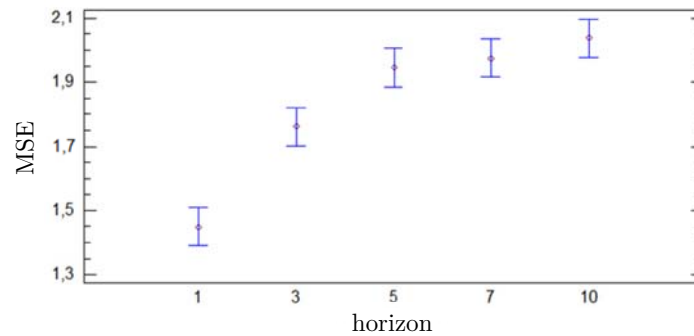


Figure 5. MSE of demand forecasts for each horizon.

From Figs. 4 and 5, it can be observed:

- A significant effect of weather forecasting accuracy is observed from 10h to 24h. The effect of weather forecasting accuracy from 1h to 9h is negligible.
- The higher effect of weather forecasting accuracy is observed at hours 17h to 19h, which provokes a MSE increment of 73% approximately.
- The MSE of demand accuracy is incremented around 25% from $K = 1$ to $K = 3$. There are no significant differences over demand quality for K values from 5 to 10.

Conclusions

This work analyzes the effect of the weather forecasting inaccuracies over the electric demand predictions. A complete case study using real-world data has been carried out, denoting that: the mean squared error of the weather forecasts can achieve a 0.21% (0.13% for coastal cities, 0.34% for interior locations) for larger prediction horizons; whereas the mean squared error of the electric demand forecasts vary from 1.4% to 2.0%.

Paula Cernuda is currently pursuing the B.S. degree in industrial engineering of the *Universidad Politécnica de Madrid* (Madrid, Spain). She has realized her final degree project collaborating with the Statistics Department.

Eduardo Caro works in the *Universidad Politécnica de Madrid* (Madrid, Spain), in the Statistics Department. His research interests include power system estimation, optimization, and electricity demand forecasting.

Jesús Juan works in the *Universidad Politécnica de Madrid* (Madrid, Spain), in the Statistics Department. His research interests include design of experiments, time series, multivariate analysis and reliability analysis.

A New Approach for Time Series Decomposition and Prediction

Yading Yue, Guangan Zhuang, Rong Zhang, Jianchun Zhao, Lichun Liu

Tencent Inc., Shenzhen, China

adenyue@tencent.com

Abstract. We proposed an approach to decompose a univariate time series into a set of sine functions whose frequencies are not necessarily integers as the discrete Fourier transform, and designed an efficient algorithm of linear time complexity to calculate the parameters of the sine functions. We also used such a decomposition to make predictions into the future, and showed the prediction accuracy is comparable with other approaches especially in cases of quasi-periodic time series.

Keywords: Time series, decomposition, prediction, forecasting, algorithm, Fourier.

1 Introduction

The decomposition of time series involves deconstructing a time series into several components, each representing one of the underlying categories of patterns. There are two principal types of decomposition^[1]: decomposition based on rates of change, typically into trend component, cyclical component, seasonal component and irregular component, and decomposition based on predictability. Typically, discrete Fourier transform (DFT), discrete wavelet transform (DWT), Hilbert–Huang transform (HHT)^[2], and Singular Spectrum Analysis (SSA)^[3] are often used.

DFT converts a sequence of N complex numbers in the time domain into an N -periodic sequence of complex numbers. DWT has extra advantage over Fourier transforms in temporal resolution: it captures both frequency and location information (location in time).

HHT decomposes a time series into intrinsic mode functions (IMF) along with a trend, and obtain instantaneous frequency data. It is designed to work well for non-stationary and nonlinear data. Specifically, HHT uses the empirical mode decomposition (EMD) to decompose a signal into a finite and often small number of components. In contrast to other common transforms like the Fourier transform and wavelet transform, HHT is more like an empirical approach applied to a data set, rather than a theoretical tool.

adfa, p. 1, 2017.

© Springer-Verlag Berlin Heidelberg 2017

DFT, DWT, and EMD are not able to predict the time series directly, unless their constituents are modeled to make predictions individually before the predicted values are integrated into a final prediction into the future.

SSA aims to make a decomposition of the original series into the sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and a structureless noise. SSA involves a singular value decomposition (SVD) operation which is quadratic in time complexity ^{[4] [5]}. SSA can be used directly to predict the future values with limited horizon.

In this paper we propose an alternative method, named quasi Fourier (QF), in salutation to the great mathematician Joseph Fourier, for decomposition of time series with linear time complexity, while the frequencies of the component sine functions need not be integers, and prediction can be naturally done by extrapolation of the constituent sine functions.

2 Our Approach

Assume the variation of a time series with values at equal intervals can be represented by the addition of a set of sine functions (cosine functions will do too, equivalently, but in the paper we use sine functions only) and a constant (intercept) which accounts for the non-zero average values. We need to determine the parameters (amplitude, frequency, and phase) of each of the sine functions.

Suppose there is a time series $\{r_j\}$, $j=1, 2, \dots, M$. We wanted to find a function

$$y_j = r_a + \sum_{i=1}^N a_i * \sin(b_i x_j + c_i) \quad (1)$$

to approximate the original time series such that a loss function (e.g., a metric in terms of MSE) can be as minimum as possible, where N is the order of the decomposition and also the number of sine functions, r_a is the average of historical data values, i.e., $r_a = \sum_j r_j / N$, x_j is the independent variable taking the values of 1, 2, ..., M , and a_i , b_i , c_i are the parameters to be determined. The above equation is also the formula for reconstruction and prediction of the time series.

For efficiency, we can get the parameters of sine functions incrementally, that is, after the intercept r_a is obtained, we first determine $\{a_1, b_1, c_1\}$ to minimize the residues $z_j^{(1)} = (r_j - r_a) - a_1 \sin(b_1 x_j + c_1)$ in terms of MSE, then determine $\{a_2, b_2, c_2\}$ to minimize the residues $z_j^{(2)} = z_j^{(1)} - (a_2 \sin(b_2 x_j + c_2))$ in terms of MSE, ..., and finally determine $\{a_N, b_N, c_N\}$ to minimize the residues $z_j^{(N)} = z_j^{(N-1)} - (a_N \sin(b_N x_j + c_N))$. The detailed algorithm is given next.

2.1 Algorithm

The pseudo code for the QF algorithm follows.

Input: original time series r_j , $j = 1, \dots, M$; the average of the historical values r_a .

Output: a_i, b_i, c_i , $i = 1, \dots, N$, where N is typically 10, 30, 100, or more.

1. for ($i = 1$ to N)
2. if ($i = 1$) then residues $z_j = r_j - r_a$; else $z_j = z_{j-a_{i-1}} \sin(b_{i-1} x_j + c_{i-1})$
3. assign $E =$ one big number (e.g., 10^{20})
4. for each $b \in [b_{\min}, b_{\max}]$
5. fix the value of a temporally;
6. find the best c , denoted as c^* ;
7. find the best a , denoted as a^* ;
8. if ($E'(a^*, b, c^*) < E$)
9. $E = E'$; $b^* = b$;
10. $\{a_i, b_i, c_i\} = \{a^*, b^*, c^*\}$.

where E' is the MSE of the sine curve approximating the current residues z_j , dependent on determined a^*, b, c^* . Let us give more explanations.

Line 4: for each $b \in [b_{\min}, b_{\max}]$. We divide the interval $[b_{\min}, b_{\max}]$ into multiple intervals of equal width, such that b takes one dimensional search interval by interval. Usually we assign $[b_{\min}, b_{\max}] = [\pi/(2 \cdot M), \pi/2]$, to cover a broad range of possible frequencies, and the number of intervals m can be thousands.

Line 5: fix the value of a temporally, which can be any value, usually 1.0. It can be verified that the magnitude of value a fixed here does not affect the calculation of other parameters later.

Line 6: find the best c , denoted as c^* .

We do this by trying each of the 3 values of c : $\{0, \pi/2, \pi\}$, to get the MSE of each of the 3 sine curves approximating the residues z_j , denoted as $e1, e2, e3$ respectively, as shown in figure 1.

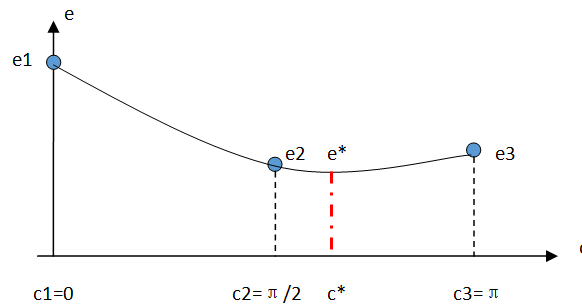


Fig. 1. Error Curve: e vs. c .

It is not difficult to prove that e is a sine function of c , so the 3 solid circles on the above figure are on one sine curve. Assume this sine curve is expressed as

$$e = p + q \sin(c - c^* - \frac{\pi}{2}) \quad (2)$$

where p , q , and c^* need to be determined, and we know that e^* corresponding to c^* is the lowest point on the sine curve. Substituting the three c values into the above equation, with necessary manipulations, we get

$$\begin{aligned} e1 &= p - q \cos(c^*) \\ e2 &= p - q \sin(c^*) \\ e3 &= p + q \sin(c^*) \end{aligned} \quad (3)$$

We can easily get: $p = (e2+e3)/2$, and $c^* = \arctg((e2-p)/(e1-p))$. Then q can also be obtained with any of the above three equations in (3) if needed. Thus c^* is obtained.

Line 7: find the best a , denoted as a^* . Once b and c^* are known, a^* can be obtained by least square method:

$$a^* = \operatorname{argmin}_a \left(L \equiv \sum_{j=1}^M (z_j - a \sin(b * x_j + c^*))^2 \right) \quad (4)$$

or more specifically, by setting $\partial L / \partial a = 0$ and solving the resulting linear equations.

2.2 Complexity Analysis

If we try to get the $\{a_i, b_i, c_i\}$ by brutal force, such as a grid search in 3 dimensions, each having m intervals, then the time complexity is $O(M*N*m^3)$, where M is the number of data points being modeled, N is the order of decomposition, and m is the number of intervals in each dimension. But our approach is only $O(M*N*m)$, since we only need to search for the best b one dimensionally and the other two parameters a and c are derived analytically, which is a remarkable reduction in time complexity.

3 Applications

The QF algorithm that we proposed can be used in all cases where previous approaches were applied. We give some experiments below.

3.1 Prediction of sunspots

To make comparisons with other time series prediction approaches, we take the sun spots data and build QF model. The monthly smoothed Sunspot time series has been obtained from the SIDC (World Data Center for the Sunspot Index) [6]. Sunspot series from November 1834 to June 2001 (2000 points) are selected and scaled between [0, 1]. The first 1000 samples of time series are selected to train and the remainder 1000 samples are kept to test the one-step prediction accuracy. The orders of the decomposition $N = 300$, and the number of intervals in $[b_{\min}, b_{\max}]$ is $m=3000$. The prediction error is measured in NMSE [14].

Comparison of the prediction errors reported in the literature and the proposed approach (1000 Sunspot time series test samples) is given in table 1. It can be seen that our approach is the third best among the all being compared.

Approaches	Prediction error (NMSE)
Wavelet Packet, Neural Networks [7]	1.25E-01
Approximation and Correction, McNish–Lincoln [8]	8.00E-02
Nonlinear Dynamical System, Sello [9]	3.40E-01
Interpolation of Waldmeier's Standard Curves [9]	5.60E-01
Geomagnetic Index as Precursor Model [10]	1.85E-00
Neural Networks, RBF-OLS [11]	4.60E-02
Neuro-Fuzzy, LLNF-LoLiMot [11]	3.20E-02
Evolving Recurrent Neural Networks [12]	2.80E-03
Multi-layer perceptron (MLP) [13]	9.79E-02
Elman–NARX Neural Networks [14]	5.90E-04
The proposed: QF	5.67E-03

Table 1. Comparison of Prediction Error for Sunspot Data

To check how prediction errors depend on the algorithm parameters, figure 2 shows that the NMSE decreases with the number of orders of decomposition N , and with the number of segments in b 's range m , but too big an m does not help further in reducing the prediction error.

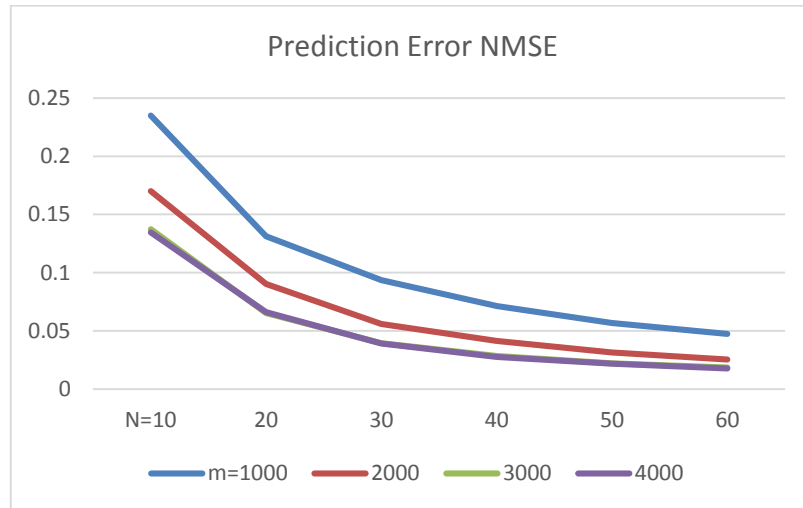


Fig. 2. Prediction Error Affected by N , m .

We note in figure 3 that bigger N leads to smaller NMSE, showing no over-fitting problem. However, too big an N does not help reduce prediction error further.

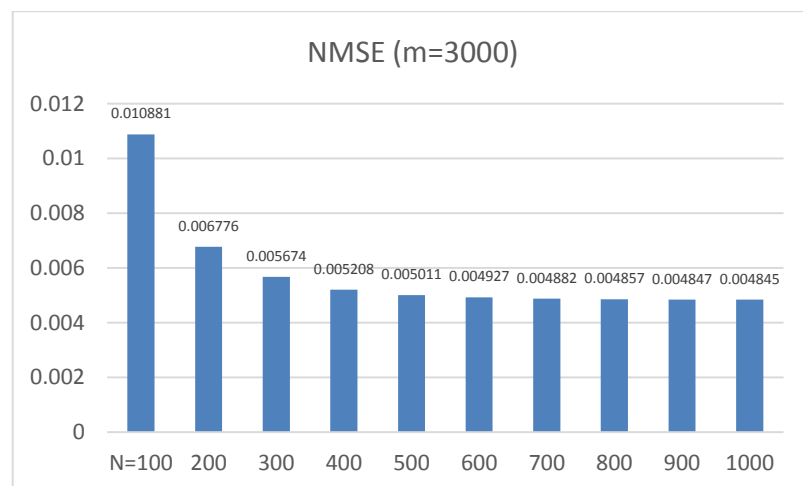


Fig. 3. Prediction Error for Bigger N .

3.2 Daily revenue prediction

Given the daily transaction volume data of an online digital product A in Tencent Inc., the lifetime of which spanned from very early years up to September 30 of one year, we were asked to predict the daily values in the next 31 days in October. These data usually demonstrated a weekly fluctuations (quasi-periodic) since users tend to buy more on weekends, and also high values on the first 7 days of October, which are the National Holidays in China.

At the end of October we compared the real values (blue) with the predicted values (orange), as in figure 4, and the predictions followed the real values closely, with MAE=9.1%, which impressed the business unit and gave them confidence on our model for more later applications.

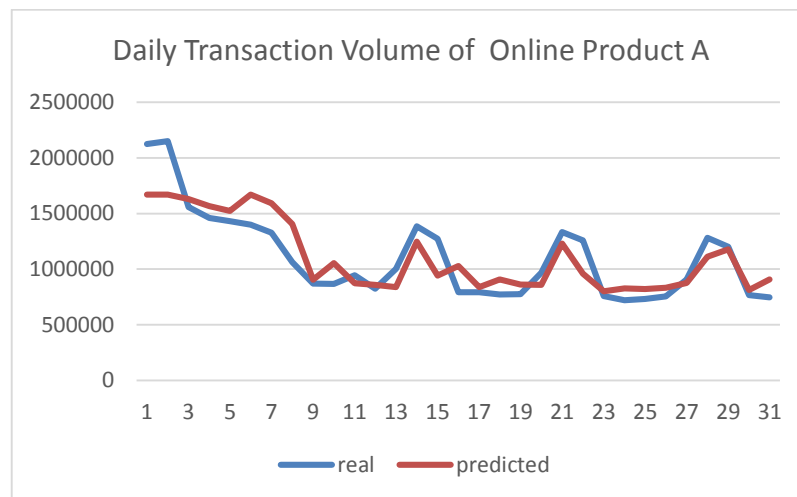


Fig. 4. Prediction of the Future 31 Days, Product A.

3.3 Estimation of revenue loss due to system failure

One online game B was affected when some of the servers were out of service during November 13 – 21 (9 days), we were asked to estimate the revenue loss.

First we built a QF model to check its backtracking prediction accuracy, and we got the results for the days just before November 13, that is, October 27 – November 12, in figure 5. The predicted values (orange) are very close to the real values (blue), with MAE as small as 1.43%, thus the model is deemed workable for the case.

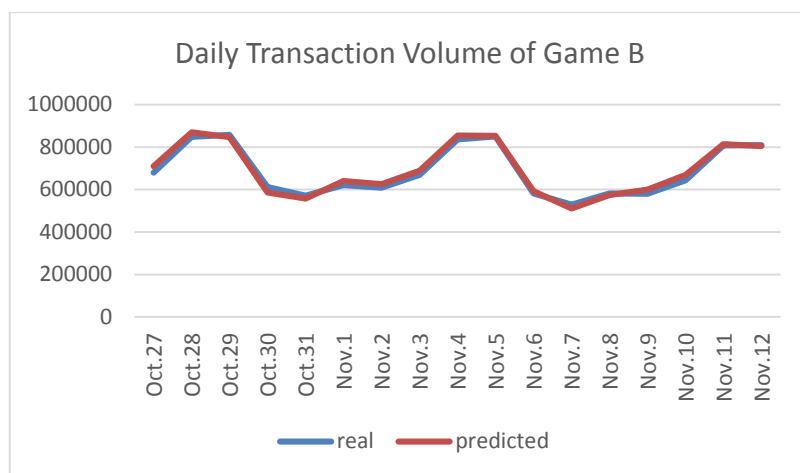


Fig. 5. Validation of the Prediction Model, Game B.

Then, we updated our model to predict the values during November 13 – 21 (9 days), in figure 6, and took the predicted as what should be if there were no system failure. By adding the differences of the “would-be real values” (orange) and the real values (blue), we got the total revenue loss in the 9 days as RMB 1,521,342 yuan.

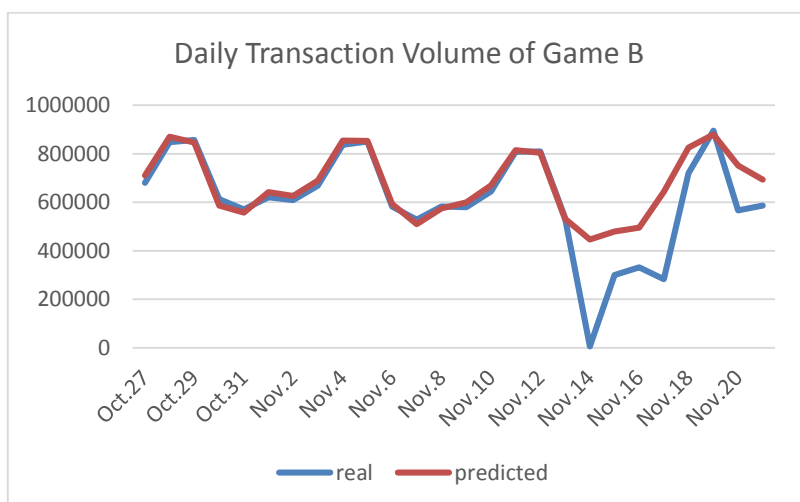


Fig. 6. Prediction Result for Calculating the Revenue Loss, Game B.

4 Conclusions

In this paper we proposed an efficient univariate time series decomposition and prediction approach similar to discrete Fourier transform in that it approximates the time series by a set of sine functions, but the frequencies of which do not have to be integers like the discrete Fourier transform. Its time complexity is linear rather than higher orders of polynomials if brutal force approaches is adopted, and it is simple to implement with an incremental learning algorithm.

The approach presented here is by no means a universal function approximator, but a speciation for the case of time series represented as a function of time t . The approach is apparently not applicable to time series which increases monotonously to possible infinity. Fortunately most times series in industry are bounded in a finite range.

The experiments showed its accurate predictions and how it was used in several real applications in industry. Further work could be extension to multivariate cases or incorporated with other approaches to form hybrid models for a better result.

References:

1. Dodge, Y. The Oxford Dictionary of Statistical Terms. New York: Oxford University Press. 2003. ISBN 0-19-920613-9
2. Huang, N. E.; Long, S. R.; Shen, Z. "The Mechanism for Frequency Downshift in Nonlinear Wave Evolution". *Advances in Applied Mechanics*. 32: 59–111. 1996. doi:10.1016/S0065-2156(08)70076-0
3. Ghil, M., R. M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, et al. "Advanced spectral methods for climatic time series", *Rev. Geophys.* 2002, 40(1), 3.1–3.41.
4. Hossein Hassani, Singular Spectrum Analysis: Methodology and Comparison, *Journal of Data Science* 5(2007), 239-257
5. Michael P. Holmes, Alexander G. Gray, Charles Lee Isbell, Jr., Fast SVD for Large-Scale Matrices, <http://sysrun.haifa.il.ibm.com/hrl/bigml/files/Holmes.pdf>
6. Sunspot Number graphics, <http://www.sidc.be/silso/ssngraphics>
7. K. Teo, L. Wang, Z. Lin, Wavelet packet multi-layer perceptron for chaotic time series prediction: effects of weight initialization, *Computational Science* 2074 (2001) 310–317.
8. A. G. McNish, J. V. Lincoln, Prediction of sunspot numbers, *Transactions American Geophysical Union* 30(1949) 673.
9. S. Sello, Solar cycle forecasting: a nonlinear dynamics approach, *Astronomy & Astrophysics* 377(2001) 312–320.
10. K. Denkmayr, P. Cugnon, About sunspot number medium-term predictions, in: G. Heckmanetal. (Ed.), *Solar-Terrestrial Prediction Workshop V*, Hiraiso Solar Terrestrial Research Center, Japan 1997, p.103.
11. A. Gholipour, B. N. Araabi, C. Lucas, Predicting chaotic time series using neural and neuro fuzzy models: a comparative study, *Neural Processing Letters* 24 (2006) 217–239.
12. Q. Ma, Q. Zheng, H. Peng, T. Zhong, L. Xu, Chaotic time series prediction based on evolving recurrent neural networks, in: *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 2007.

13. T. Koskela, M. Lehtokangas, J. Saarinen, K. Kaski, Time series prediction with multilayer perceptron, FIR and Elman neural networks, in: Proceedings of the World Congress on Neural Networks, 1996, pp.491–496.
14. Muhammad Ardalani-Farsa, Saeed Zolfaghari, Chaotic time series prediction with residual analysis method using hybrid Elman–NARX neural networks, *Neurocomputing* 73(2010) 2540–2553

Short-term time series forecasting based on internal smoothing of Padé interpolants

Minvydas Ragulskis*, Kristina Lukoseviciute, Tadas Telksnys, and Zenonas Navickas

Research Group for Mathematical and Numerical Analysis of Dynamical Systems,
Kaunas University of Technology
Studentu 50-147, Kaunas LT-51368, Lithuania
`minvydas.ragulskis@ktu.lt`, `kristina.lukoseviciute@ktu.lt`,
`tadas.telksnys@ktu.lt`, `zenonas.navickas@ktu.lt`
<http://www.minvydasragulskis.com/>

Abstract. Short-term time series forecasting technique based on Padé interpolants and adaptive internal smoothing is presented in this paper. Adaptive corrections of time series data in the window of observation allows to construct near-optimal Padé extrapolant. Computational experiments with real world time series are used to demonstrate the efficiency of the proposed approach.

Keywords: time series, forecasting, Padé interpolant, internal smoothing

1 Introduction

Time series forecasting is an important technique used in a large variety of applications in different areas of science, engineering, finance and economics in general. The basic idea of any time series prediction algorithm is to identify a mathematical model generating the analyzed series and project this model into the future. Many different time series forecasting models and techniques have been developed during the recent decades. Conditionally, these methods can be classified into long-term and short term time series forecasting algorithms [1].

Time prediction horizon correlates with this classification – usually only short predictions suffice for short-term time series. It is true that predictors with even one time step forward horizons are important in a variety of applications [1]. Such techniques are widely used in finance [2–4]; electricity demand and the associated price forecasting problem [5–7]; wind power; passenger demand [8] and many others.

One time step forward prediction algorithms are usually based on the extrapolation of the available data. It is well known that Padé interpolants can be used for generating mathematical models of complex nonlinear processes [9].

* Corresponding author.

Padé functions, defined as ratios of univariate polynomials of, in general, different orders, have classically been used to approximate smooth functions with known Taylor series [10]. Padé functions have also been applied in split-step approximations of the solutions to differential equations [11–13], approximation of elliptic-type functions [14], generalized Euler transforms [15] and cosmographic analysis [16].

Significant attention has been recently devoted to Padé interpolation schemes. Padé-type rational and barycentric interpolation is considered in [9]. A rational interpolation scheme with a superpolynomial rate of convergence that reduces the Runge effect and can be used on discontinuous functions has been developed in [17]. A robust and efficient implementation of the rational interpolation scheme can be found in [18].

Padé approximants have also been applied to time series analysis and forecasting, primarily as a tool for the construction of ARMA models. In [19], the Padé approximation is used to accomplish the LS identification of an unstable ARMA equations. A method to identify the order of an ARMA time series model and to compute its coefficient based on the Padé approximant is presented in [20]. These methods have been applied to real-life time series in the field of economics [21, 22].

The main objective of this article is to present a new application of Padé-type methods in short-term time series forecasting. This paper is organized as follows: internal smoothing of algebraic is discussed in Section 2; an overview of Padé interpolants is presented in Section 3; the pre-processing algorithm of time series data is given in Section 4; the fitness function construction is discussed in Section 5; computational experiments are discussed in Sections 6 and 7; concluding remarks are given in the last section.

2 Internal smoothing of algebraic interpolants – preliminaries

Internal smoothing of an algebraic interpolant has been introduced in [23]. The main idea of this smoothing procedure is based on a projection of the reconstructed algebraic model into the future. However, instead of trying to make a straightforward projection of the model, a conciliation between the variability of the algebraic interpolant and the smoothness of moving average time series estimates is considered. We will use the standard industrial moving average algorithm to smooth the time series:

$$MA_t = \frac{1}{s} \sum_{j=0}^{s-1} x_{t-j-1}, \quad (1)$$

where MA_t is a smooth value at time moment t ; s is the averaging window. In general, the width of the averaging window should be preselected for each time series is not related to the length of the observation window used for the algebraic interpolant.

Now, time series elements in the observation window are individually perturbed by corrections $\varepsilon_1, \dots, \varepsilon_M$. It is clear that additional constraints for the corrections are required in order to make this extrapolation problem well-posed. A fitness functions for the set of corrections $\varepsilon_1, \dots, \varepsilon_M$ can be maximized in order to reconstruct a near-optimal algebraic skeleton representing the underlying dynamics in the observation window [23]:

$$F(\varepsilon_1, \dots, \varepsilon_M) = \frac{1}{\alpha \sum_{j=1}^M |\varepsilon_j| + |\tilde{x}_t - MA_t|}, \quad (2)$$

where \tilde{x}_t is an exact algebraic extrapolant constructed over the perturbed elements of the time series in the observation window; the parameter $\alpha > 0$ determines the penalty proportion between the sum of corrections and the difference of forecast produced by algebraic extrapolant and moving average. It is clear that the target function would be unbounded at all zero corrections if algebraic extrapolant constructed over non-perturbed elements of the time series would coincide to moving average prediction.

The objective of this paper is to employ Padé interpolants for the algebraic prediction of the time series evolution and to enhance the intelligent perturbation of the analyzed time series.

3 Discrete Padé approximation scheme

Traditionally, Padé approximations are used to approximate a smooth function with a Taylor series expression by means of a rational function. In this paper, we apply the Padé approximant to time series data.

Suppose the time series $(t_1, x_1), \dots, (t_M, x_M)$ is given; t_k denotes the time variable and x_k denotes the measurement taken at time t_k . Note that for time series where the length of the time interval is unknown, it can be taken that $t_k = k, k = 1, \dots, M$ with no impact on the approximation.

The order (m, n) Padé function reads:

$$[m/n]_x(t) := \frac{\sum_{j=0}^m a_j t^j}{1 + \sum_{j=1}^n b_j t^j}; \quad a_j, b_j \in \mathbb{R}. \quad (3)$$

It is recommended to select $m \geq n$ [10]. The function (3) applied to the time series data yields the following system of linear equations with respect to the parameters $a_0, \dots, a_m, b_1, \dots, b_n$:

$$[m/n]_x(t_k) = x_k; \quad k = 1, \dots, M. \quad (4)$$

Inserting (3) into (4) and simplifying yields:

$$\sum_{j=0}^m a_j t_k^j - x_k \sum_{l=1}^n b_l t_k^l = x_k; \quad k = 1, \dots, M. \quad (5)$$

The system (5) can be rewritten in matrix form:

$$\mathbf{W}\mathbf{p} = \mathbf{x}, \quad (6)$$

where

$$\mathbf{W} = \begin{bmatrix} 1 & t_1 & \dots & t_1^m & -x_1 t_1 & \dots & -x_1 t_1^n \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_M & \dots & t_M^m & -x_M t_M & \dots & -x_M t_M^n \end{bmatrix}; \quad (7)$$

and

$$\mathbf{p} = [a_0 \dots a_m \ b_1 \dots b_n]^T; \quad \mathbf{x} = [x_1 \dots x_M]^T. \quad (8)$$

Let us denote $N = \dim \mathbf{p} = m + n + 1$ the total number of parameters in (3). Noting that \mathbf{W} is nonsingular if $t_k \neq t_l, x_k \neq x_l; k \neq l$ (which is always satisfied for time series, since $t_k < t_l$ for $k < l$) yields the unique least-squares solution to (6) for $N \leq M$:

$$\mathbf{p} = \left(\mathbf{W}^T \mathbf{W} \right)^{-1} \mathbf{W}^T \mathbf{x}. \quad (9)$$

Note that for $N = M$, the interpolant is obtained. However, this is impractical for time series analysis, because of the large number of parameters required and the negative impact of the Runge effect [24].

4 Pre-processing algorithm

The pre-processing algorithm of the time series data is given below. This algorithm normalizes the time series and selects optimal parameters for the Padé extrapolant using full sort.

Algorithm 1: Time series pre-processing

Input : x_1, \dots, x_M – time series;
 \overline{M} – maximum number parameters in Padé function (3);
 L – number of time-forward steps for RMSE evaluation.

Output: $\tilde{x}_1, \dots, \tilde{x}_M$ – time series normalized to the range of $[-1, 1]$;
 (m_*, n_*) – parameters of optimal Padé function for given time series.

```

1 Normalize time series:
2 for  $k = 1, \dots, M$  do
3    $\tilde{x}_k = \frac{2x_k - \max_{1 \leq l \leq M} x_l - \min_{1 \leq l \leq M} x_l}{\max_{1 \leq l \leq M} x_l - \min_{1 \leq l \leq M} x_l}$ ;
4 end
5
6 Minimize  $RMSE(m, n)$  using full sort:
7 for  $j = 2, \dots, \overline{M}$  do
8   for  $N = 2, \dots, M$  do
9     for  $m = \lfloor \frac{N}{2} \rfloor, \dots, N - 1$  do
10       $n = N - m - 1$ ;
11      form  $\mathbf{W}, \mathbf{x}$  with  $n, m$  and  $x_{\overline{M}-j+1}, \dots, x_{\overline{M}}$ ;
12      obtain parameters:  $\mathbf{p} = (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \mathbf{x}$ ;
13      compute Padé forecast for  $L$  forwards steps:  $\hat{x}_{\overline{M}+1}, \dots, \hat{x}_{\overline{M}+L}$ ;
14      compute error:  $RMSE(m, n) = \frac{1}{L} \sqrt{\sum_{k=1}^L (x_{\overline{M}+k} - \hat{x}_{\overline{M}+k})^2}$ ;
15    end
16  end
17 end
18 Choose parameters with smallest error:  $(m_*, n_*) = \arg \min_{m, n} RMSE(m, n)$ .

```

5 The construction of the fitness function

An evolutionary strategy is used in [25] to identify the algebraic skeleton sequence in the observation window of the predicted time series by removing the unknown additive noise. The idea is based on the assumption that the time series comprises some sort of deterministic skeleton describing the dynamics of the time series which is contaminated by the additive noise.

Let us denote $\tilde{x}_k = x_k - \varepsilon_k; k = 1, 2, 3, \dots$ as the corrected values of the sequence (ε_k are unknown corrections). The F-measure as in [26] becomes the fitness measure of a genetic algorithm that identifies predictive patterns in the sequence of events [27].

The F-measure consists of two parts that embody different objectives: PRECISION, the model precision, requires from the model that it faithfully recon-

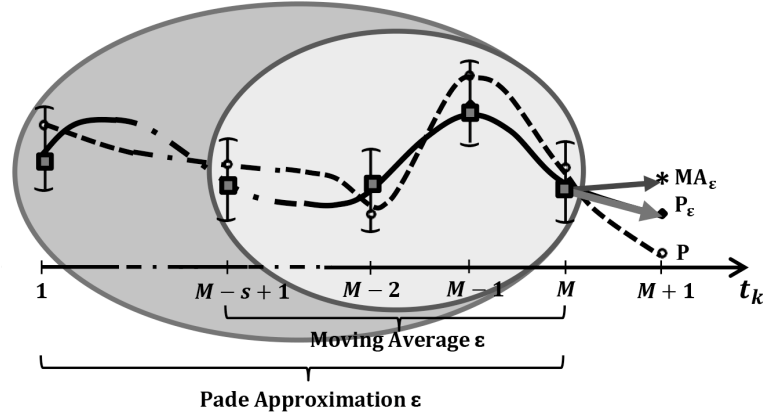


Fig. 1. Schematic diagram illustrating the proposed Padé method, where dots denote the original time series; squares denote the corrected time series; P^ε the Padé forecasting for corrected time series; P forecasting for original time series.

struct the last known time series values and RECALL requires that the prediction repeats past dynamical behaviour:

$$F(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M) = \frac{(\gamma^2 + 1) \cdot \text{PRECISION} \cdot \text{RECALL}}{\gamma^2 \cdot \text{PRECISION} + \text{RECALL}}. \quad (10)$$

In equation (10) the value γ controls the relative importance of precision to recall. If $\gamma = 0$ then the fitness function evaluates only the PRECISION part. If $\gamma = \infty$ then the fitness function evaluates the RECALL values only. In our case we build PRECISION and RECALL functions in such a way that the minimal value of the fitness function is reached when the corrections are small and the improved Bernstein extrapolation (through points \tilde{x}_k) is close to the moving average prediction (also based on \tilde{x}_k):

$$\text{PRECISION} = \frac{1}{M-1} \sum_{i=1}^M |x_i - \hat{x}_i|, \quad (11)$$

$$\text{RECALL} = \alpha \sum_{i=1}^M |\varepsilon_i| + \beta |MA_{M+1} - P_{M+1}^\varepsilon|; \alpha > 0, \beta > 0, \quad (12)$$

where an array $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_M$ represents near-optimal corrections of the original time series; MA_{M+1} stands for the moving average through the last s time series values; P_{M+1}^ε stands for the Padé extrapolation through last M values of \tilde{x}_k ; parameter α determines the penalty proportion between the sum of weighted corrections and the difference of forecasts based on MA_{n+1} and P_{M+1}^ε . Fig. 1 illustrates this technique.

To calibrate the fitness function in order to obtain optimal forecasts, an analysis on the impact of parameters α, β, γ to the fitness function must be performed. As shown in Fig. 5, the prediction results are closest to the original time series elements for $\alpha = 1, \beta = 0.5, \gamma = 2$ ($RMSE = 0.0480$). The parameter values obtained in this computational experiment are fixed for the subsequent computations.

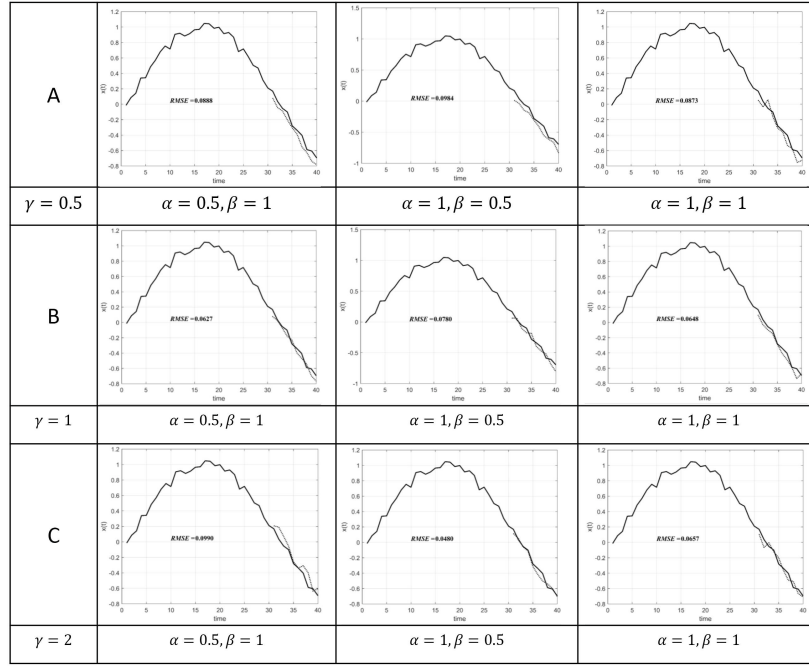


Fig. 2. Plots of the prediction accuracy for different values of the fitness function parameters α, β, γ .

6 Forecasting strategy for the Dow Jones time series

In previous section the optimal parameters of fitness function were selected. We apply this forecasting model with preselected fitness function parameters to real world time series: Dow Jones Industrial Average (DJIA) time series (data range provides 1896-05-26 to 2013-08-27 monthly index observations made up of 11 US stocks) [29]; DJIA time series is normed into the range $[-1; 1]$. The pre-processing is executed for $\{x_0, \dots, x_{30}\}$. The Padé model is built using the initial 21 observations; RMSE is computed for the last 10 observations. Padé polynomial parameters read: $M = 20, m = n = 4$. Results of the prediction for

these parameter values are displayed in Fig. 3. It can be noted that this model can be improved by taking into account dynamical nature of the Dow Jones time series.

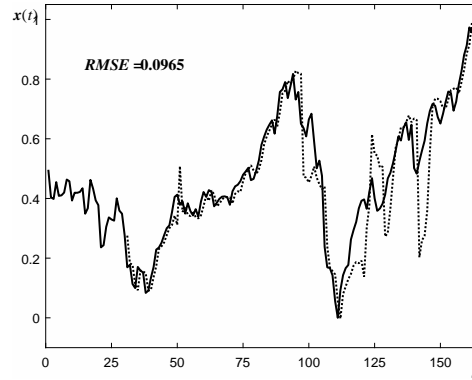


Fig. 3. Dow-Jones time series forecasting.

7 Adaptive forecasting of the Dow Jones time series

In this section, a strategy of adaptive selection of parameters M, m, n is proposed and used to predict different parts of time series. The proposed analysis is based on the idea of algebraic segmentation of short nonstationary time series [28]. The error level $\delta = 0.9$ is the key parameter that is preselected before the prediction is done. The prediction of the Dow Jones time series is shown in Fig. 4 (A); Fig. 4 (B) is the error plot for the prediction displayed in Fig. 4 (B).

Note that in the time interval $[31; 49]$, the time series has been predicted using Padé extrapolation with parameters $M = 20, n = m = 4$. In the interval $[50; 53]$, parameters $M = 20, n = m = 9$ have been used. In the next time series segment $[54, 96]$ optimal predictions are obtained for parameter values $M = 20, m = 5, n = 6$; for segments $[97; 99], [100; 106], [107; 117]$ the parameters values read $M = 11, m = 9, n = 1$; $M = 15, m = 8, n = 4$; $M = 20, n = m = 4$ respectively. For the remaining segment of the series, parameter values $M = 7, n = m = 1$ have been used. This parameter selection scheme has enabled the reduction of forecasting error down to $RMSE = 0.0707$.

8 Concluding remarks

Time series forecasting technique based on adaptive one step forward extrapolation of Padé extrapolants with internal smoothing is presented in this paper. Special optimization problem is developed for the identification of a near-optimal

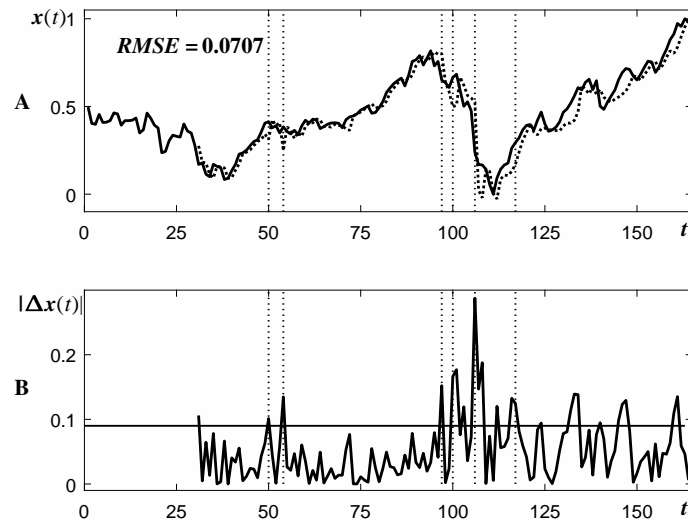


Fig. 4. Adaptive forecasting Dow Jones adaptive time series: part (A) is the time series forecast; part (B) is the error between the forecast and original time series values.

set of corrections used for the identification of the algebraic model in the observation window of the analyzed time series. Such an approach enables to unleash the power of Padé interpolants by ensuring the optimal smoothness of the extrapolant. Computational experiments with Dow Jones time series demonstrate the efficiency and the applicability of the proposed techniques for the prediction of real-world time series.

Acknowledgment

This research was funded by a grant (No. MIP078/2015) from the Research Council of Lithuania.

References

1. Parras-Gutierrez, E., Rivas, V. M., Garcia-Arenas, M., del Jesus, M. J.: Short, medium and long term forecasting of time series using the L-Co-R algorithm. *Neurocomputing* 128, 433–446 (2014)
2. Cao, Q., Leggio, K. B., Schniederjans M. J.: A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers & Operations Research*, Vol. 32(10), 2499–2512 (2005)
3. Lam K., Lam, K. C.: Forecasting for the generation of trading signals in financial markets. *Journal of Forecasting*, Vol. 19(1), 39–52 (2000)
4. Podsiadlo, M., Rybinski, H.: Financial time series forecasting using rough sets with time-weighted rule voting. *Expert Systems With Applications* 66, 219–233 (2016)

5. Taylor, J. W., de Menezes, L. M., McSharry, P. E.: A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, Vol. 22(1), 1–16 (2006)
6. Darbellay, G. A., Slama, M.: Forecasting the short-term demand for electricity – Do neural networks stand a better chance?. *International Journal of Forecasting*, Vol. 16(1), 71–83 (2000)
7. Nogales, F. J., Contreras, J., Conejo, A. J., Espnola, R.: Forecasting Next-Day Electricity Prices by Time Series Models. *IEEE TRANSACTIONS ON POWER SYSTEMS*, Vol. 17 (2), 342–348 (2002)
8. Kim, S., Shin, D. H.: Forecasting short-term air passenger demand using big data from search engine queries. *Automation in Construction* 70, 98–108 (2016)
9. Brezinski, C., Redivo-Zaglia, M.: Padé-type rational and barycentric interpolation. *Numerische Mathematik*, Vol. 125 (1), 89–113 (2013)
10. Baker, G. A. Jr., Graves-Morris P. R.: *Padé Approximants* Second edition, Cambridge University Press (1996)
11. Michael, D., Collins, A.: Split-step Padé Solution for the Parabolic Equation Method. *J. Acoust. Soc. Am.* 93, 1736 (1993)
12. Kamakura, T., Nomura, H., Clement, G. T.: Application of the split-step Padé approach to nonlinear field predictions *Ultrasonics* 53, 432–438 (2013)
13. Nagao, H.: The Padé Interpolation Method Applied to q-Painlevé Equations. *Letters in Mathematical Physics*, Vol. 105 (4), 503–521 (2015)
14. Baratchart, L., Yattselev, M. L.: Padé approximants to certain elliptic-type functions. *Journal d'Analyse Mathématique*, Vol. 121(1), 31–86 (2013)
15. Sablonnière, P.: Padé-type approximants for generalized Euler transforms. *Numerical Algorithms*, Vol. 66(2), 339–347 (2014)
16. Gruber, C., Luongo, O.: Cosmographic analysis of the equation of state of the universe through Padé approximations. *Physical Review D* 89, 103506 (2014)
17. Wang, Q., Moin, P., Iaccarino, G.: A Rational Interpolation Scheme with Superpolynomial Rate of Convergence. *SIAM Journal on Numerical Analysis*, Vol. 47(6), 4073–4097 (2010)
18. Gonnet, P., Pachon, R., Trefethen, L. N.: Robust rational interpolation and least-squares. *Electronic Transactions on Numerical Analysis* 38, 146–167 (2011)
19. Gel, Y. R., Fomin, V. N.: Identification of an unstable ARMA equation. *Mathematical problems in engineering* 7, 97–112 (2001)
20. Kumar, K.: Padé approximation and its application in time series analysis. *Applied Mathematics and Computation*, Vol. 48(2-3), 139–151 (1992)
21. Gil-Farina, M. C., Gonzalez-Concepcion, C., Pestano-Gabino, C.: Padé Approximation Modelling of an Advertising-Sales Relationship. *J. Service Science & Management* 3, 91–97 (2010)
22. Gonzalez-Concepcion, C., Gil-Farina, M. C.: Padé Approximation in Economics. *Numerical Algorithms* 33, 277–292 (2003)
23. Palivonaite, R., Ragulskis, M.: Short-term time series algebraic forecasting with internal smoothing. *Neurocomputing*, 127, 161–171 (2014).
24. Heath, M.: *Scientific Computing*. McGraw-Hill, New York (2000).
25. Palivonaite, R., Lukoseviciute, K., Ragulskis, R.: Short-term time series algebraic forecasting with mixed smoothing. *Neurocomputing* 171, 854–865 (2016)
26. van Rijsbergen, C. J.: *Information Retrieval*. Butterworths, London (1979)
27. Weiss, G. M.: Timeweaver: a Genetic Algorithm for Identifying Predictive Patterns in Sequences of Events. In *Proceedings of the Genetic and Evolutionary Computation Conference* 718–725 (1999)

28. Palivonaite, R., Lukoseviciute, K., Ragulskis, M.: Algebraic segmentation of short nonstationary time series based on evolutionary prediction algorithms. *Neurocomputing* 121, 354–364 (2013)
29. Federal reserve bank of st. Louis, <http://research.stlouisfed.org/fred2/series/STLFSI/downloaddata>.

The Dependence Structures of Non-Stationary Bivariate INAR(1) Processes: The Case of the Bivariate Poisson Innovations

Jowaheer Vandna*, Sunecher Yuvraj, and Mamode Khan Naushad.

University of Mauritius
Reduit Mauritius

{vandnaj@uom.ac.mu, ysunecher@umail.utm.ac.mu, n.mamodekhan@uom.ac.mu}

Abstract. There exists different mechanisms to induce the cross-correlation in the bivariate integer-valued autoregressive processes of order 1 (BINAR(1)). Some papers considered constrained processes where the interrelation between the two series is induced by correlated innovations only while in the unconstrained models an additional cross-correlation is borne by the relation between the observations from one series with the previous-lagged observations of the other series. However, in some unconstrained processes, some researchers have considered only a one-way cross-correlation by assuming the innovation series are mutually independent. This paper provides an analytical review of the two forms of unconstrained processes under the Poisson innovation assumptions and in particular under non-stationary moments. Monte Carlo simulations are implemented to compare the different BINAR(1) processes. These two unconstrained models are also applied to analyze real-life series of day and night accidents in Mauritius.

Keywords: Constrained, Unconstrained, BINAR(1), Poisson, GQL.

1 Introduction

In the recent decades, several bivariate integer-valued autoregressive of order 1 (BINAR(1)) processes were introduced in the literature. They differ mainly in the way their corresponding innovation series were specified. Originally, Pedeli and Karlis [3, 4] developed the BINAR(1) with Poisson and NB innovations respectively where the cross-correlation between the two series were only induced by the correlated innovation terms, thus making the model constrained. Later these authors proposed the full or unconstrained model with Poisson innovations that considers cross-correlation borne by the innovations and the relation of the current responses from one series with the previous-lagged observation of the other series [5]. In the same manner, Ristic et al. [6] and Nastic et al. [2] defined

* Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

another form of unconstrained BINAR(1) process with Geometric marginal but with innovation terms independent. Interestingly, in these unconstrained models, the corresponding marginal series exhibit same over-dispersion. Moreover, the model proposed by Ristic et al. [6] shows superior Akaike Information Criterion (AIC) than the constrained and unconstrained models of Pedeli and Karlis [3–5]. Overall, all these models were developed only under the stationary assumption, that is, under constant moments.

As for the non-stationary BINAR(1) processes, Mamode Khan et al. [1] and Sunecher et al. [7] developed the constrained BINAR(1) process with Poisson and NB innovations respectively. However, as at date, there is no unconstrained non-stationary BINAR(1) with Poisson innovations developed yet. Based on the findings of Nastic et al. [2], this paper aims at developing the non-stationary BINAR(1) with Poisson innovations under the two forms of unconstrained set-up. The parameters in these two processes are estimated via the GQL approach.

The paper is laid out as follows: In Section 2, the marginal and joint moments of two unconstrained non-stationary BINAR(1) with Poisson innovations are derived. In Section 3, three GQL equations are developed to estimate the regression and dependence parameters. In the same section, we derive the forecasting equations. This section is followed by a numerical evaluation where the performance of the GQL is assessed on the two unconstrained non-stationary BINAR(1) processes. In Section 5, the two BINAR(1) models are applied to analyze the day and night accidents along the motorway connecting the International Airport of Mauritius and tourist zone Grand-Bay in Mauritius. The conclusion is provided in Section 6.

2 The Unconstrained Non-Stationary BINAR(1) Process with Poisson Innovations (M1)

Consider

$$Y_t^{[1]} = \rho_{11} * Y_{t-1}^{[1]} + \rho_{12} * Y_{t-1}^{[2]} + R_t^{[1]} \quad (1)$$

$$Y_t^{[2]} = \rho_{21} * Y_{t-1}^{[1]} + \rho_{22} * Y_{t-1}^{[2]} + R_t^{[2]} \quad (2)$$

where $\rho_{kj} \in (0, 1)$ and $\rho_{kj} *$ are mutually independent binomial thinning operators such that $\rho_{kj} * Y_{t-1}^{[k]} = \sum_{i=0}^{Y_{t-1}^{[k]}} Z_i$ where $Z_i \sim \text{Bernoulli}(\rho_{kj})$. In the first instance, let us consider $\text{Corr}(R_t^{[1]}, R_t^{[2]}) = \alpha_{12,t}$ where $(R_t^{[1]}, R_t^{[2]})$ is bivariate Poisson with $R_t^{[k]} \sim \text{Poisson}(\lambda_t^{[k]})$, where

$$\lambda_t^{[1]} = (\mu_t^{[1]} - \rho_{11}\mu_{t-1}^{[1]} - \rho_{12}\mu_{t-1}^{[2]}) > 0 \quad (3)$$

$$\lambda_t^{[2]} = (\mu_t^{[2]} - \rho_{21}\mu_{t-1}^{[1]} - \rho_{22}\mu_{t-1}^{[2]}) > 0 \quad (4)$$

with $\mu_t^{[k]} = E(Y_t^{[k]})$. Under these assumptions,

$$\begin{aligned} \text{Var}(Y_t^{[1]}) &= \rho_{11}(1 - \rho_{11})\mu_{t-1}^{[1]} + \rho_{11}^2 \text{Var}(Y_{t-1}^{[1]}) + \rho_{12}(1 - \rho_{12})\mu_{t-1}^{[2]} + \rho_{12}^2 \text{Var}(Y_{t-1}^{[2]}) \\ &\quad + 2\rho_{11}\rho_{12}\text{Cov}(Y_{t-1}^{[1]}, Y_{t-1}^{[2]}) + \lambda_t^{[1]} \\ &= \mu_t^{[1]} + [2\rho_{11}\rho_{12}\text{Cov}(Y_{t-1}^{[1]}, Y_{t-1}^{[2]}) + \rho_{11}^2(\text{Var}(Y_{t-1}^{[1]}) - \mu_{t-1}^{[1]}) + \rho_{12}^2(\text{Var}(Y_{t-1}^{[2]}) - \mu_{t-1}^{[2]})] \end{aligned} \quad (5)$$

$$\begin{aligned} \text{Var}(Y_t^{[2]}) &= \rho_{21}(1 - \rho_{21})\mu_{t-1}^{[1]} + \rho_{21}^2 \text{Var}(Y_{t-1}^{[1]}) + \rho_{22}(1 - \rho_{22})\mu_{t-1}^{[2]} + \rho_{22}^2 \text{Var}(Y_{t-1}^{[2]}) \\ &\quad + 2\rho_{21}\rho_{22}\text{Cov}(Y_{t-1}^{[1]}, Y_{t-1}^{[2]}) + \lambda_t^{[2]} \\ &= \mu_t^{[2]} + [2\rho_{21}\rho_{22}\text{Cov}(Y_{t-1}^{[1]}, Y_{t-1}^{[2]}) + \rho_{22}^2(\text{Var}(Y_{t-1}^{[2]}) - \mu_{t-1}^{[2]}) + \rho_{21}^2(\text{Var}(Y_{t-1}^{[1]}) - \mu_{t-1}^{[1]})] \end{aligned} \quad (6)$$

$$\text{Cov}(Y_t^{[1]}, Y_t^{[2]}) = (\rho_{11}\rho_{22} + \rho_{12}\rho_{21})\text{Cov}(Y_{t-1}^{[1]}, Y_{t-1}^{[2]}) + \rho_{11}\rho_{21}\text{Var}(Y_{t-1}^{[1]}) + \rho_{22}\rho_{12}\text{Var}(Y_{t-1}^{[2]}) + \text{Cov}(R_t^{[1]}, R_t^{[2]}) \quad (7)$$

From the above, the corresponding formula for the second form of the unstrained process may be derived easily (M2). Note that the variance of the counting series $Y_t^{[k]}$ is greater than the expected mean which indicates that $Y_t^{[k]}$ is over-dispersed. Moreover, the marginal distribution of $Y_t^{[k]}$ based on equations (5) and (6) is rather difficult to identify.

As for the lag-covariances, they are computed similarly as in Pedeli and Karlis [5], where $\Sigma_{h,t} = \begin{bmatrix} \text{Cov}(Y_t^{[1]}, Y_{t+h}^{[1]}) & \text{Cov}(Y_t^{[1]}, Y_{t+h}^{[2]}) \\ \text{Cov}(Y_t^{[2]}, Y_{t+h}^{[1]}) & \text{Cov}(Y_t^{[2]}, Y_{t+h}^{[2]}) \end{bmatrix}$. In this paper, $\mu_t^{[k]} = \exp(x_t' \beta^{[k]})$, where $x_t = [x_{t1}, x_{t2}, \dots, x_{tp}]'$ is a $p \times 1$ vector of covariates influencing both $Y_t^{[1]}$ and $Y_t^{[2]}$ with corresponding regression coefficients $\beta^{[k]} = [\beta_1^{[k]}, \beta_2^{[k]}, \dots, \beta_j^{[k]}, \dots, \beta_p^{[k]}]'$.

3 Estimation Methods

The estimation of the regression parameters is performed using the GQL equation

$$D_\beta' \Sigma_\beta^{-1} (\mathbf{f} - \boldsymbol{\mu}) = 0 \quad (8)$$

where the score vector $\mathbf{f} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_t, \dots, \mathbf{f}_{t+h}, \dots, \mathbf{f}_T]$, with $\mathbf{f}_t = [Y_t^{[1]}, Y_t^{[2]}]'$ and $\boldsymbol{\mu}$ is the corresponding expected score. The t^{th} row of Σ_β is expressed as $c(\Sigma_{t-1,1}, \Sigma_{t-2,2}, \dots, \Sigma_{0,t}, \dots, \Sigma'_{t+h-t,t}, \dots, \Sigma'_{T-t,t})$ where

$$\Sigma_{t,h} = \begin{bmatrix} \text{Cov}(Y_t^{[1]}, Y_{t+h}^{[1]}) & \text{Cov}(Y_t^{[1]}, Y_{t+h}^{[2]}) \\ \text{Cov}(Y_t^{[2]}, Y_{t+h}^{[1]}) & \text{Cov}(Y_t^{[2]}, Y_{t+h}^{[2]}) \end{bmatrix} \text{ and } [\cdot]' \text{ is the transpose matrix.}$$

The derivative component $D_\beta = [D_1, D_2, \dots, D_t, \dots, D_T]'$ where $D_t = \text{diag}(\frac{\partial \mu_t^{[1]}}{\partial \beta^{[1]}}, \frac{\partial \mu_t^{[2]}}{\partial \beta^{[2]}})$

with $\frac{\partial \mu_t^{[k]}}{\partial \beta_j^{[k]}} = \mu_t^{[k]} x'_{tj}$. Using these component matrix, equation (8) is solved using the Newton-Raphson (NR) as in Mamode Khan et al. [1]. It is shown that $(\hat{\beta} - \beta)$ is asymptotic normal with mean 0 and covariance matrix $[\mathbf{D}'_{\beta} \boldsymbol{\Sigma}_{\beta}^{-1} \mathbf{D}_{\beta}]^{-1} [\mathbf{D}'_{\beta} \boldsymbol{\Sigma}_{\beta}^{-1} (\mathbf{f} - \boldsymbol{\mu})_{\beta} (\mathbf{f} - \boldsymbol{\mu})'_{\beta} \boldsymbol{\Sigma}_{\beta}^{-1} \mathbf{D}_{\beta}] [\mathbf{D}'_{\beta} \boldsymbol{\Sigma}_{\beta}^{-1} \mathbf{D}_{\beta}]^{-1}$ [1, 8, 9]. On the other hand, the estimation of the dependence coefficients $\hat{\rho}_{11}, \hat{\rho}_{12}, \hat{\rho}_{21}$ and $\hat{\rho}_{22}$ is performed using the GQL equation

$$\mathbf{D}_{\rho}^{*'} \boldsymbol{\Sigma}_{\rho}^{*-1} (\mathbf{f}^* - \mathbf{m}) = 0, \quad (9)$$

where the score function is re-defined as $\mathbf{f}^* = [\mathbf{f}_{1|0}^*, \mathbf{f}_{2|1}^*, \dots, \mathbf{f}_{t|t-1}^*, \dots, \mathbf{f}_{t+h|t+h-1}^*, \dots, \mathbf{f}_{T|T-1}^*]$,

where $\mathbf{f}_{t|t-1}^* = [Y_t^{[1]2} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}, Y_t^{[2]2} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}]'$ and \mathbf{m} is the corresponding expected score. The t^{th} row of $\hat{\boldsymbol{\Sigma}}_{\rho}^*$ is expressed as $c(\boldsymbol{\Sigma}_{t-1,1}^*, \boldsymbol{\Sigma}_{t-2,2}^*, \dots, \boldsymbol{\Sigma}_{0,t}^*, \dots, \boldsymbol{\Sigma}_{t+h-t,t}^{*'}, \dots, \boldsymbol{\Sigma}_{T-t,t}^{*'})$ where

$$\boldsymbol{\Sigma}_{h,t}^* = \begin{bmatrix} \text{Cov}(Y_t^{[1]2}, Y_{t+h}^{[1]2} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}, Y_{t+h-1}^{[1]}, Y_{t+h-1}^{[2]}) & \text{Cov}(Y_t^{[1]2}, Y_{t+h}^{[2]2} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}, Y_{t+h-1}^{[1]}, Y_{t+h-1}^{[2]}) \\ \text{Cov}(Y_t^{[1]2}, Y_{t+h}^{[2]2} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}, Y_{t+h-1}^{[1]}, Y_{t+h-1}^{[2]}) & \text{Cov}(Y_t^{[2]2}, Y_{t+h}^{[2]2} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}, Y_{t+h-1}^{[1]}, Y_{t+h-1}^{[2]}) \end{bmatrix} \quad (10)$$

The entries of the above matrices are computed using the 'working' multivariate normality assumption as in Sunecher et al. [7]. The derivative matrix

$$\mathbf{D}_{\rho}^* = [\mathbf{D}_1^*, \mathbf{D}_2^*, \dots, \mathbf{D}_t^*, \dots, \mathbf{D}_T^*]' \text{ where } \mathbf{D}_t^* = \left[\begin{pmatrix} \frac{\partial m_{t|t-1}^{[1]}}{\partial \rho_{11}} \\ \frac{\partial \rho_{11}^{[1]}}{\partial m_{t|t-1}^{[1]}} \\ \frac{\partial m_{t|t-1}^{[2]}}{\partial \rho_{12}} \\ \frac{\partial \rho_{12}^{[2]}}{\partial m_{t|t-1}^{[2]}} \end{pmatrix}, \begin{pmatrix} \frac{\partial m_{t|t-1}^{[2]}}{\partial \rho_{21}} \\ \frac{\partial \rho_{21}^{[2]}}{\partial m_{t|t-1}^{[2]}} \\ \frac{\partial m_{t|t-1}^{[1]}}{\partial \rho_{22}} \\ \frac{\partial \rho_{22}^{[1]}}{\partial m_{t|t-1}^{[1]}} \end{pmatrix} \right]$$

The entries of the derivative matrix are

1. $\frac{\partial m_{t|t-1}^{[1]}}{\partial \rho_{11}} = Y_{t-1}^{[1]} - 2\rho_{11}Y_{t-1}^{[1]} - \mu_{t-1}^{[1]} + 2(\rho_{11}Y_{t-1}^{[1]} + \rho_{12}Y_{t-1}^{[2]} + \mu_t^{[1]} - \rho_{11}\mu_{t-1}^{[1]} - \rho_{12}\mu_{t-1}^{[2]})(Y_{t-1}^{[1]} - \mu_{t-1}^{[1]})$
2. $\frac{\partial m_{t|t-1}^{[1]}}{\partial \rho_{12}} = Y_{t-1}^{[2]} - 2\rho_{12}Y_{t-1}^{[2]} - \mu_{t-1}^{[2]} + 2(\rho_{11}Y_{t-1}^{[1]} + \rho_{12}Y_{t-1}^{[2]} + \mu_t^{[1]} - \rho_{11}\mu_{t-1}^{[1]} - \rho_{12}\mu_{t-1}^{[2]})(Y_{t-1}^{[2]} - \mu_{t-1}^{[2]})$
3. $\frac{\partial m_{t|t-1}^{[2]}}{\partial \rho_{21}} = Y_{t-1}^{[1]} - 2\rho_{21}Y_{t-1}^{[1]} - \mu_{t-1}^{[1]} + 2(\rho_{21}Y_{t-1}^{[1]} + \rho_{22}Y_{t-1}^{[2]} + \mu_t^{[2]} - \rho_{21}\mu_{t-1}^{[1]} - \rho_{22}\mu_{t-1}^{[2]})(Y_{t-1}^{[1]} - \mu_{t-1}^{[1]})$
4. $\frac{\partial m_{t|t-1}^{[2]}}{\partial \rho_{22}} = Y_{t-1}^{[2]} - 2\rho_{22}Y_{t-1}^{[2]} - \mu_{t-1}^{[2]} + 2(\rho_{21}Y_{t-1}^{[1]} + \rho_{22}Y_{t-1}^{[2]} + \mu_t^{[2]} - \rho_{21}\mu_{t-1}^{[1]} - \rho_{22}\mu_{t-1}^{[2]})(Y_{t-1}^{[2]} - \mu_{t-1}^{[2]})$

The estimation of the cross-correlation parameter $\alpha_{12,t}$ is performed using:

$$\mathbf{D}_{\alpha_{12,t}}' [\text{Var}(Y_t^{[1]}Y_t^{[2]} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]})]_{(1 \times 1)}^{-1} [(Y_t^{[1]}Y_t^{[2]})_{(1 \times 1)} - (E(Y_t^{[1]}Y_t^{[2]} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]}))_{(1 \times 1)}] = 0 \quad (11)$$

The multivariate normality structure is used to approximate $\text{Var}(Y_t^{[1]}Y_t^{[2]} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]})$

and $\frac{\partial E(Y_t^{[1]}Y_t^{[2]} | Y_{t-1}^{[1]}, Y_{t-1}^{[2]})}{\partial \alpha_{12,t}} = \sqrt{\lambda_t^{[1]}} \sqrt{\lambda_t^{[2]}}$. The overall algorithm works as follows: for an initial value of $\hat{\beta}_j^{[k]}$, $\hat{\rho}_{kj}$ and $\hat{\alpha}_{12,t}$, we apply the first iterative equation

to obtain an updated estimate of the regression parameters and this updated estimate is used to estimate $\hat{\rho}_{kj}$ using the second iterative equation. Using this updated $\hat{\rho}_{kj}$ and $\hat{\beta}_j^{[k]}$, we estimate $\hat{\alpha}_{12,t}$ using the third iterative equation and then, a new set of regression parameters is re-evaluated using these updates and then this cycle continues until convergence of the three sets of parameters.

The forecasting equations are derived as follows:

By using the model in Section 2, we first write

$$Y_{t+1}^{[1]} = \rho_{11} * Y_t^{[1]} + \rho_{12} * Y_t^{[2]} + R_{t+1}^{[1]} \quad (12)$$

$$Y_{t+1}^{[2]} = \rho_{21} * Y_t^{[1]} + \rho_{22} * Y_t^{[2]} + R_{t+1}^{[2]} \quad (13)$$

and for given $Y_t^{[k]}$, the forecasting function is expressed as

$$E(Y_{t+1}^{[1]} | Y_t^{[1]}, Y_t^{[2]}) = \hat{\mu}_{t+1}^{[1]} + \hat{\rho}_{11}(Y_t^{[1]} - \hat{\mu}_t^{[1]}) + \hat{\rho}_{12}(Y_t^{[2]} - \hat{\mu}_t^{[2]}) \quad (14)$$

$$E(Y_{t+1}^{[2]} | Y_t^{[1]}, Y_t^{[2]}) = \hat{\mu}_{t+1}^{[2]} + \hat{\rho}_{21}(Y_t^{[1]} - \hat{\mu}_t^{[1]}) + \hat{\rho}_{22}(Y_t^{[2]} - \hat{\mu}_t^{[2]}) \quad (15)$$

and

$$Var(Y_{t+1}^{[1]} | Y_t^{[1]}, Y_t^{[2]}) = \hat{\rho}_{11}(1 - \hat{\rho}_{11})Y_t^{[1]} + \hat{\rho}_{12}(1 - \hat{\rho}_{12})Y_t^{[2]} + \hat{\mu}_{t+1}^{[1]} - \hat{\rho}_{11}\hat{\mu}_t^{[1]} - \hat{\rho}_{12}\hat{\mu}_t^{[2]} \quad (16)$$

$$Var(Y_{t+1}^{[2]} | Y_t^{[1]}, Y_t^{[2]}) = \hat{\rho}_{21}(1 - \hat{\rho}_{21})Y_t^{[1]} + \hat{\rho}_{22}(1 - \hat{\rho}_{22})Y_t^{[2]} + \hat{\mu}_{t+1}^{[2]} - \hat{\rho}_{21}\hat{\mu}_t^{[1]} - \hat{\rho}_{22}\hat{\mu}_t^{[2]} \quad (17)$$

4 Numerical Evaluation

In this section, we generate the BINAR(1) data using the models derived in Section 2, by using the following 2×1 time-dependent covariate matrix, where the first covariate is:

$$x_{t1} = \begin{cases} -1 + t & (t = 1, \dots, T/4), \\ \text{rnorm}(1, 0, 1) & (t = (T/4) + 1, \dots, 3T/4), \\ 1 + t & (t = (3T/4) + 1, \dots, T), \end{cases}$$

$$x_{t2} = \begin{cases} (1/t) & (t = 1, \dots, T/4) \\ (-1/t) & (t = (T/4) + 1, \dots, 3T/4) \\ t & (t = (3T/4) + 1, \dots, T) \end{cases}$$

The data are generated assuming $(\rho_{11}, \rho_{22}) = [0.3, 0.9]$ and $\rho_{12}, \rho_{21} = 0$ with dependence coefficient $\alpha_{12,t} = [0.3, 0.9]$ and $\beta_1^{[k]} = 0.3$ and $\beta_2^{[k]} = 0.7$. $(R_t^{[1]}, R_t^{[2]})$ are simulated using the codes developed by Mamode Khan et al. [1]. 5000 simulated runs are performed for each combination for $T=100, 500$ and 1000 .

$\alpha_{12,1}$	ρ_{11}	ρ_{22}	T	Models	$\hat{\beta}_1^{[1]}$	$\hat{\beta}_2^{[1]}$	$\hat{\beta}_1^{[2]}$	$\hat{\beta}_2^{[2]}$	$\hat{\rho}_{11}$	$\hat{\rho}_{22}$	$\hat{\rho}_{12}$	$\hat{\rho}_{21}$	$\hat{\alpha}_{12,1}$			
0.3	0.9	0.9	100	M1	0.2824 (0.0978)	0.2862 (0.0964)	0.6829 (0.0982)	0.6856 (0.0980)	0.8815 (0.1152)	0.8860 (0.1118)	0.0042 (0.1108)	0.0094 (0.1154)	0.2899 (0.1285)			
				M2	0.2820 (0.0990)	0.2809 (0.0910)	0.6818 (0.0992)	0.6827 (0.0988)	0.8811 (0.1167)	0.8844 (0.1130)	0.0055 (0.1119)	0.0098 (0.1170)	0.2860 (0.1296)			
			500	M1	0.2924 (0.0530)	0.2932 (0.0522)	0.6924 (0.0575)	0.6951 (0.0507)	0.8940 (0.0610)	0.8945 (0.0660)	0.0017 (0.0662)	0.0040 (0.0681)	0.2937 (0.0759)			
				M2	0.2914 (0.0550)	0.2920 (0.0536)	0.6911 (0.0590)	0.6938 (0.0522)	0.8927 (0.0619)	0.8933 (0.0675)	0.0025 (0.0673)	0.0031 (0.0693)	0.2926 (0.0750)			
			1000	M1	0.2982 (0.0139)	0.2990 (0.0145)	0.6978 (0.0132)	0.6994 (0.0113)	0.8954 (0.0274)	0.8974 (0.0231)	0.0005 (0.0259)	0.0011 (0.0232)	0.2980 (0.0354)			
				M2	0.2970 (0.0150)	0.2981 (0.0159)	0.6965 (0.0148)	0.6978 (0.0128)	0.8941 (0.0290)	0.8960 (0.0244)	0.0010 (0.0275)	0.0020 (0.0246)	0.2966 (0.0363)			
			0.3	0.3	0.9	100	M1	0.2888 (0.0910)	0.2814 (0.0965)	0.6838 (0.0921)	0.6850 (0.0977)	0.2866 (0.1136)	0.8868 (0.1179)	0.0020 (0.1152)	0.0043 (0.1112)	0.2826 (0.1264)
							M2	0.2875 (0.0902)	0.2810 (0.0956)	0.6831 (0.0910)	0.6845 (0.0961)	0.2857 (0.1124)	0.8855 (0.1163)	0.0031 (0.1140)	0.0052 (0.1103)	0.2817 (0.1252)
						500	M1	0.2919 (0.0545)	0.2958 (0.0527)	0.6929 (0.0536)	0.6936 (0.0510)	0.2950 (0.0654)	0.8916 (0.0661)	0.0010 (0.0670)	0.0016 (0.0647)	0.2923 (0.0740)
							M2	0.2910 (0.0560)	0.2949 (0.0535)	0.6919 (0.0547)	0.6926 (0.0520)	0.2940 (0.0669)	0.8910 (0.0674)	0.0018 (0.0682)	0.0023 (0.0660)	0.2914 (0.0758)
						1000	M1	0.2997 (0.0181)	0.2984 (0.0142)	0.6982 (0.0173)	0.6985 (0.0133)	0.2960 (0.0278)	0.8993 (0.0260)	0.0001 (0.0204)	0.0003 (0.0225)	0.2955 (0.0306)
							M2	0.2930 (0.0195)	0.2960 (0.0161)	0.6940 (0.0184)	0.6949 (0.0147)	0.2951 (0.0293)	0.8945 (0.0275)	0.0011 (0.0220)	0.0009 (0.0241)	0.2932 (0.0322)
0.3	0.3	0.3				100	M1	0.2851 (0.0961)	0.2811 (0.0987)	0.6802 (0.0944)	0.6830 (0.0923)	0.2819 (0.1175)	0.2811 (0.1114)	0.0068 (0.1108)	0.0076 (0.1151)	0.2824 (0.1245)
							M2	0.2840 (0.0973)	0.2805 (0.0999)	0.6800 (0.0956)	0.6824 (0.0930)	0.2807 (0.1186)	0.2801 (0.1127)	0.0080 (0.1121)	0.0089 (0.1165)	0.2815 (0.1259)
						500	M1	0.2950 (0.0526)	0.2959 (0.0531)	0.6966 (0.0540)	0.6933 (0.0575)	0.2919 (0.0662)	0.2920 (0.0672)	0.0038 (0.0661)	0.0041 (0.0609)	0.2952 (0.0734)
							M2	0.2940 (0.0538)	0.2944 (0.0542)	0.6952 (0.0554)	0.6920 (0.0588)	0.2910 (0.0676)	0.2911 (0.0686)	0.0045 (0.0670)	0.0049 (0.0619)	0.2939 (0.0747)
						1000	M1	0.2980 (0.0121)	0.2978 (0.0192)	0.6979 (0.0123)	0.6950 (0.0129)	0.2977 (0.0238)	0.2956 (0.0215)	0.0010 (0.0224)	0.0013 (0.0247)	0.2977 (0.0336)
							M2	0.2970 (0.0133)	0.2966 (0.0199)	0.6969 (0.0137)	0.6941 (0.0147)	0.2970 (0.0250)	0.2945 (0.0227)	0.0016 (0.0243)	0.0018 (0.0260)	0.2967 (0.0348)
			0.9	0.9	0.9	100	M1	0.2884 (0.0916)	0.2821 (0.0933)	0.6825 (0.0945)	0.6831 (0.0947)	0.8896 (0.1158)	0.8891 (0.1196)	0.0090 (0.1130)	0.0074 (0.1171)	0.8816 (0.1256)
							M2	0.2870 (0.0940)	0.2810 (0.0950)	0.6812 (0.0959)	0.6817 (0.0963)	0.8840 (0.1170)	0.8855 (0.1199)	0.0098 (0.1147)	0.0088 (0.1185)	0.8804 (0.1269)
						500	M1	0.2912 (0.0558)	0.2956 (0.0536)	0.6941 (0.0530)	0.6958 (0.0580)	0.8962 (0.0620)	0.8968 (0.0641)	0.0056 (0.0646)	0.0055 (0.0673)	0.8910 (0.0796)
							M2	0.2904 (0.0574)	0.2940 (0.0550)	0.6931 (0.0545)	0.6941 (0.0592)	0.8953 (0.0631)	0.8949 (0.0656)	0.0071 (0.0658)	0.0068 (0.0684)	0.8902 (0.0810)
						1000	M1	0.2997 (0.0123)	0.2982 (0.0169)	0.6975 (0.0129)	0.6973 (0.0152)	0.8981 (0.0231)	0.8977 (0.0220)	0.0011 (0.0243)	0.0026 (0.0234)	0.8998 (0.0348)
							M2	0.2988 (0.0140)	0.2970 (0.0180)	0.6966 (0.0142)	0.6960 (0.0160)	0.8972 (0.0244)	0.8968 (0.0233)	0.0020 (0.0252)	0.0035 (0.0247)	0.8983 (0.0360)
0.9	0.3	0.9				100	M1	0.2896 (0.0947)	0.2820 (0.0925)	0.6824 (0.0908)	0.6895 (0.0919)	0.2886 (0.1173)	0.8852 (0.1116)	0.0080 (0.1163)	0.0098 (0.1151)	0.8815 (0.1274)
							M2	0.2880 (0.0959)	0.2808 (0.0940)	0.6820 (0.0922)	0.6885 (0.0933)	0.2875 (0.1185)	0.8841 (0.1130)	0.0091 (0.1177)	0.0099 (0.1164)	0.8806 (0.1289)
						500	M1	0.2912 (0.0522)	0.2945 (0.0550)	0.6951 (0.0524)	0.6956 (0.0544)	0.2904 (0.0633)	0.8923 (0.0671)	0.0060 (0.0627)	0.0070 (0.0688)	0.8918 (0.0710)
							M2	0.2905 (0.0536)	0.2932 (0.0567)	0.6941 (0.0530)	0.6940 (0.0554)	0.2901 (0.0642)	0.8916 (0.0680)	0.0075 (0.0636)	0.0083 (0.0695)	0.8910 (0.0720)
						1000	M1	0.2993 (0.0121)	0.2981 (0.0122)	0.6977 (0.0105)	0.6984 (0.0130)	0.2962 (0.0255)	0.8991 (0.0250)	0.0031 (0.0259)	0.0037 (0.0259)	0.8989 (0.0343)
							M2	0.2980 (0.0130)	0.2971 (0.0136)	0.6968 (0.0116)	0.6972 (0.0144)	0.2955 (0.0263)	0.8979 (0.0261)	0.0040 (0.0275)	0.0052 (0.0269)	0.8975 (0.0351)
			0.9	0.3	0.3	100	M1	0.2882 (0.0912)	0.2821 (0.0930)	0.6831 (0.0961)	0.6812 (0.0955)	0.2824 (0.1145)	0.2845 (0.1115)	0.0095 (0.1126)	0.0083 (0.1119)	0.8841 (0.1295)
							M2	0.2871 (0.0920)	0.2811 (0.0944)	0.6815 (0.0970)	0.6802 (0.0960)	0.2820 (0.1150)	0.2833 (0.1133)	0.0099 (0.1141)	0.0095 (0.1129)	0.8830 (0.1299)
						500	M1	0.2950 (0.0589)	0.2913 (0.0558)	0.6927 (0.0563)	0.6912 (0.0511)	0.2952 (0.0660)	0.2907 (0.0675)	0.0070 (0.0633)	0.0065 (0.0678)	0.8959 (0.0728)
							M2	0.2940 (0.0598)	0.2909 (0.0565)	0.6919 (0.0570)	0.6905 (0.0525)	0.2944 (0.0675)	0.2901 (0.0686)	0.0080 (0.0640)	0.0079 (0.0689)	0.8950 (0.0736)
						1000	M1	0.2979 (0.0114)	0.2992 (0.0195)	0.6980 (0.0159)	0.6982 (0.0138)	0.2937 (0.0231)	0.2946 (0.0215)	0.0033 (0.0217)	0.0041 (0.0248)	0.8990 (0.0320)
							M2	0.2970 (0.0125)	0.2985 (0.0199)	0.6969 (0.0175)	0.6972 (0.0147)	0.2930 (0.0239)	0.2940 (0.0226)	0.0045 (0.0231)	0.0053 (0.0258)	0.8982 (0.0329)

Table 1. GQL estimates of the parameters and standard errors under non-stationary BINAR(1) process, based on 5000 Monte-Carlo replications for each combination ρ_{11}, ρ_{22} .

From the above, the GQL estimates under both M1 and M2 are consistent with M1 yielding lower standard errors (s.e) than M2.

5 Application

We analyze the number of day and night accidents along the motorway connecting the International Airport of Mauritius to the tourist zone Grand-Bay. Data were collected from 1st February 2015 to 31th December 2015, which makes a total of 334 paired observations and four explanatory variables that influence the two series were also collected: number of policemen (NP) deployed in this area monthly for patrol, number of speed cameras (NSC), number of traffic lights (TL) and number of roundabouts (RA). The figures below display the time series plots of the two data:

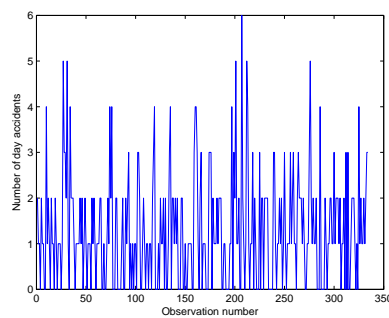


Fig. 1. Time series plot for day accidents

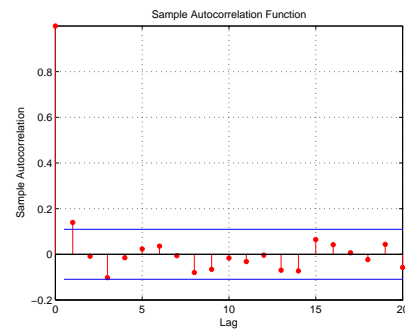


Fig. 2. ACF plot for day accidents

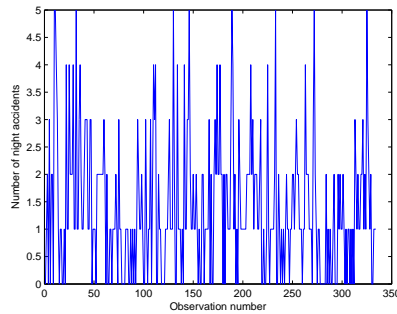


Fig. 3. Time series plot for night accidents

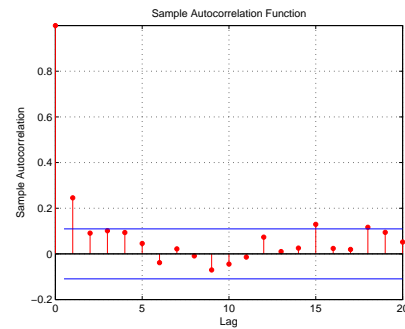


Fig. 4. ACF plot for night accidents

Day accident has a mean (variance) of 1.2754 (1.4314) and 1.3832 (1.4983) for night accident. Hence, both series are slightly over-dispersed with a cross-

correlation of 0.1164. Based on the ACF plots, we observe that the lag-1 autocorrelation is the most highly significant compared to lag-3, lag-15 and lag-18 for the accident data, which justify the application of the BINAR(1) model. We therefore fit the two unconstrained structures M1 and M2 and the GQL is applied to estimate the regressions effects and dependence parameters and the results are given in the table below:

Model		Intercept	TL	NSC	NP	RA	$\hat{\rho}_{12}$	$\hat{\rho}_{21}$	$\hat{\rho}_{11}$	$\hat{\rho}_{22}$	$\hat{\alpha}_{12,1}$
M1	$Y_t^{[1]}$	2.0067	-0.0315	-0.1151	-0.1213	0.0672	0.1178		0.0757		0.0759
	s.e	(0.2043)	(0.1272)	(0.1018)	(0.0531)	(0.1489)	(0.0761)		(0.0851)		(0.1166)
	p-values	0.1231	0.0158	0.0037	0.0093	0.0258					
	$Y_t^{[2]}$	1.1411	-0.0458	-0.0953	-0.1114	0.0874		0.1289		0.1956	
	s.e	(0.1873)	(0.1438)	(0.0860)	(0.0464)	(0.1356)		(0.0627)		(0.0979)	
	p-values	0.1729	0.0288	0.0106	0.0033	0.0118					
M2	$Y_t^{[1]}$	2.1581	-0.0430	-0.1212	-0.1265	0.0742	0.1270		0.0880		
	s.e	(0.2170)	(0.1285)	(0.1035)	(0.0550)	(0.1499)	(0.0775)		(0.0872)		
	p-values	0.1266	0.0190	0.0069	0.0115	0.0281					
	$Y_t^{[2]}$	1.1244	-0.0491	-0.0987	-0.1165	0.0944		0.1298		0.1970	
	s.e	(0.1883)	(0.1474)	(0.0884)	(0.0488)	(0.1376)		(0.0650)		(0.0990)	
	p-values	0.1788	0.0296	0.0130	0.0066	0.0135					

Table 2. Day and Night Accidents: Estimates of the regression and dependence parameters.

From the above estimates, M1 yields slightly better standard errors than M2. In the model M1, the day accidents reduces by an average of 10.9 percent, while the night accidents decreases by 9.1 percent if more speed cameras are installed. Similarly, if more Policemen are deployed, the expected decrease in the number of day accidents turns around 11.4 percent, while the number of night accidents decreases by 10.5 percent. Next, the RA estimates show that more roundabouts may induce further accidents, namely an expected increase of 7 percent for day accidents and 9.1 percent for night accidents. As for the traffic lights, its values are rather small and contribute less to the decrease in the number of accidents. Hence, more traffic lights tend to cause an expected decrease of 3.1 percent in the number of day accidents and 4.5 percent in the number of night accidents. Using the forecasting equations (14) and (15), we compute the one-step ahead in-sample prediction and the root mean square errors (RMSE) are presented in the table below for the two models:

Model	RMSE $Y_t^{[1]}$	RMSE $Y_t^{[2]}$
M1	0.2031	0.1826
M2	0.2045	0.1839

Table 3. RMSE in-sample values for number of day and night accidents.

6 Conclusion

This paper firstly treats the modelling of a non-stationary BINAR(1) series with Poisson innovations under unconstrained structures M1 and M2. Under M1, both cross-correlation structures were considered whilst under M2, we assume the inter-linkage between the two series was only borne by cross-correlation between the counting series. Parameter estimation under both models was conducted using some GQL equations which under some scores specification requires the multivariate normality 'working' structure. The outcomes of the simulation experiments and real-life study show that M1 provides slightly more efficient estimates with slightly lower RMSE than M2. However, in some complex bivariate time series process, the cross-correlation between the innovation terms may be ignored. In this situation, we still have the random vector $\{Y_t^{[1]}, Y_t^{[2]}\}$ consist of correlated random variables. Besides, this may yield some parsimony in estimating the parameters.

References

1. Mamode Khan, N., Sunecher, Y., Jowaheer, V.: Modelling a non-stationary BINAR(1) Poisson process. *Journal of Statistical Computation and Simulation* 86, 3106–3126 (2016b)
2. Nastic, A., Ristic, M., Popovic, P.: Estimation in a bivariate integer-valued autoregressive process. *Communication in Statistics-Theory and Methods* 45(19), 5660–5678 (2016b)
3. Pedeli, X., Karlis, D.: Bivariate INAR(1) models. Tech. rep., Athens University of Economics (2009)
4. Pedeli, X., Karlis, D.: A bivariate INAR(1) process with application. *Statistical Modelling: An International Journal* 11, 325–349 (2011)
5. Pedeli, X., Karlis, D.: Some properties of multivariate INAR(1) processes. *Computational Statistics and Data Analysis* 67, 213–225 (2013a)
6. Ristic, M., Nastic, A., Jayakumar, K., Bakouch, H.: A bivariate INAR(1) time series model with geometric marginals. *Applied Mathematical Letters* 25(3), 481–485 (2012)
7. Sunecher, Y., Mamodekhan, N., Jowaheer, V.: A gql estimation approach for analysing non-stationary over-dispersed BINAR(1) time series. *Journal of Statistical Computation and Simulation* (2017)
8. Sutradhar, B.: An overview on regression models for discrete longitudinal responses. *Statistical Science* 18(3), 377–393 (2003)

9. Sutradhar, B., Jowaheer, V., Rao, P.: Remarks on asymptotic efficient estimation for regression effects in stationary and non-stationary models for panel count data. *Brazilian Journal of Probability and Statistics* 28(2), 241–254 (2014)

Similarity Analysis of Time Interval Data Sets Regarding Time Shifts and Rescaling

Marc Haßler, Sabina Jeschke, and Tobias Meisen

Institute of Information Management in Mechanical Engineering,
RWTH Aachen University, Germany
`marc.hassler@ima.rwth-aachen.de`,
home page: <https://www.ima-zlw-ifu.rwth-aachen.de>

Abstract. Comparing things like objects, tasks, texts or audio is a common task in computer science. To do so, first a definition for similarity is required. In many fields of application, common and generic distance measures like the Minkowski distance or more specific measures like Dynamic Time Warping to compare temporal sequences are already defined and used. Based on our state of knowledge, there is no applicable measurement for calculating the similarity between time interval data sets in a manlike understanding.

In this paper, we present a novel method to compare time interval data sets while using an adapted distance measurement. With our approach we look at the data sets as the disjoint parts of a bigraph, such that we can use methods from graph theory. In particular, our solution provides the opportunity to take dynamic changes (like rescaling or time-shifting) into account and thus allows the comparison of real data in humanoid fashion. Hence, it allows to compare real data with e.g. scale models.

Keywords: time interval data set, TIDA, similarity analysis, graph theory, temporal displacement

1 Introduction and Motivation

Nowadays, process optimization is an essential feature in many areas of manufacturing [1]. The stability of these optimized processes is particularly important because continuous deviance could lead to aberration within the process management regarding its optimized parameters. This may result in unwanted time delays and additional costs as, due to the increasingly frequent on-demand production, there are no products in storage [2]. Today's optimized and timed procedures are more susceptible to irregularity, as a result of which, in addition to the optimization, the deviations themselves become more and more superficial. Since these deviations cannot always be avoided, a strategy for faster responses must be available.

At the moment, workers recognize deviations based on their experience and start appropriate counter measures. This procedure resembles a similarity analysis regarding past processes. Recognitions like these are not often part of the

computer-aided similarity analysis because at the moment there are only few limited possibilities to quantify interval similarities. Examples for specific similarity analysis already exist within certain scientific research areas. Those methods include the area of text or image processing, where similarity analysis is used to optimize search algorithms [3], within biology to compare genes or gene groups ([4] and [5]) or as a tool of audio recognition methods [6]. To our knowledge, basic considerations of similarities regarding time intervals are missing up to now. However, research regarding time interval data sets gained importance over the past years ([7] - [12]). While the similarity in the mentioned publications was derived from a sequence analysis and studies the existing data sets as a whole, we deduce the similarity of the data set from the individual similarities between the underlying intervals in our approach.

Therefore, we concentrate solely on the time intervals and at first construct a similarity measure to compare intervals with each other. This method allows for a detailed view on specific characteristics of the records saved in the time interval data set and a comparison even under big time offsets is possible. With this approach we are able to measure the similarity of two data sets with well known methods from graph theory [13]. To achieve our goal, we interpret the comparative data sets as the disjoint parts of an bigraph where each time interval is represented by a node within these parts and the weight of each edge represents the similarity measure of the corresponding intervals.

2 Related Work

One work regarding time interval data sets [7] compares two data sets in relation to the correlation of the intervals within the respective data set. Within this method, a difference regarding the interval length is not considered as long as it does not effect the correlation between the two intervals. The authors introduce seven interval correlations, which they used for their comparison (cf. figure 1).

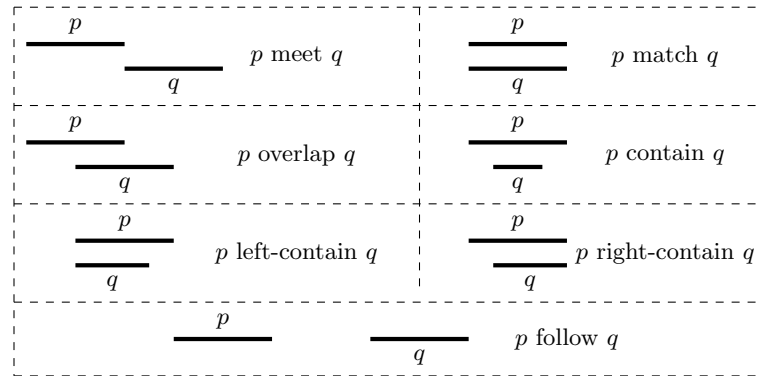


Fig. 1. Interval relation within a data set defined by Kostakis et al. [7]

Further authors [8] differentiate the similarity analysis into three distances, which are later combined to form the similarity measure. These distances are determined at a specific time t and are the following:

- 1) 'temporal order distance' compares the number of active intervals at time t .
- 2) 'temporal measure distance' matches the 'value' of all intervals at time t .
- 3) 'temporal relation distance' analyzes the relation of all intervals at time t .

This approach takes into account the lengths of the individual intervals, but only considers the data set for each evaluation at a certain point in time. Therefore, even small time shifts in one of the datasets are fully changing the outcome of the analysis.

The previously mentioned methods can be described as static comparisons, as depicted in Figure 2, yet global changes (like temporal displacements) are not regarded. Here, our method has a decisive advantage as we are able to allow global changes to be incorporated into the model by matching individual intervals.

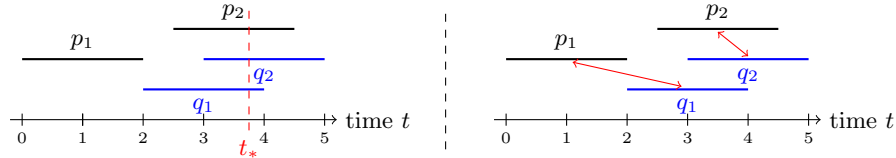


Fig. 2. *left:* time-based view on data sets by Meisen [8]; *right:* our interval approach

3 Similarities Between Time Intervals

In order to make sure that two time intervals are comparable, we take a closer look at the construction of these intervals. They consist of a start point and an end point and any amount of metadata, such as device class or hourly cost to run e.g. a specific process. In this paper, we assume that the metadata is available in mathematical form and is thus comparable (cf. chapter 3.2). We consider the following form for an interval p :

$$p := (s_p, e_p, M_{p_i} \mid i \in \mathbb{N})$$

or in *short form* $p := (s_p, e_p)$

where

$$\begin{aligned} s_p &:= \text{start point of the interval} \\ e_p &:= \text{end point of the interval} \\ M_{p_i} &:= i\text{-th metadata of the interval} \end{aligned}$$

The analysis is divided into three parts. At first, the geometrical data of each interval, such as length or position on the time axis, is compared to generate geometrical distances between two intervals. In the second part, the metadata as well as the possibility to address deadlines or earliest starting time is added into the interval similarity. In the end, all of these information define a similarity measure for two time intervals.

3.1 Geometrical Analysis

In the first step, we use the information for each interval to generate several distances with the possibility to evaluate each characteristic differently. For two intervals $p = (s_p, e_p)$ and $q = (s_q, e_q)$ as well as a norm $\|\cdot\|$, we conclude the following geometrical attributes.

- 1) Start point distance:

$$D_S(p, q) := \frac{\|s_p - s_q\|}{\|\max\{e_p, e_q\} - \min\{s_p, s_q\}\|} \quad (1)$$

- 2) End point distance:

$$D_E(p, q) := \frac{\|e_p - e_q\|}{\|\max\{e_p, e_q\} - \min\{s_p, s_q\}\|} \quad (2)$$

- 3) Lengths distance:

$$D_L(p, q) := 1 - \frac{\min\{\|e_p - s_p\|, \|e_q - s_q\|\}}{\max\{\|e_p - s_p\|, \|e_q - s_q\|\}} \quad (3)$$

- 4) Overlap:

$$D_O(p, q) := 1 - \frac{\|p \cap q\|}{\min\{\|e_p - s_p\|, \|e_q - s_q\|\}} \quad (4)$$

with the interval

$$p \cap q = \begin{cases} (\max\{s_p, s_q\}, \min\{e_p, e_q\}) & \text{for } \max\{s_p, s_q\} < \min\{e_p, e_q\} \\ 0 & \text{else} \end{cases}$$

- 5) Gap:

$$D_G(p, q) := \begin{cases} \frac{\min\{\|s_q - e_p\|, \|s_p - e_q\|\}}{\|\max\{e_p, e_q\} - \min\{s_p, s_q\}\|} & \text{for } \|p \cap q\| = 0 \\ 0 & \text{else} \end{cases} \quad (5)$$

Example 1. To visualize the geometrical attributes, we take a closer look at the intervals $p := (0, 10)$ and $q := (3, 7)$ and calculate their attributes:

- 1) $\|p\| = 10$, $\|q\| = 4$ and $\|\max\{e_p, e_q\} - \min\{s_p, s_q\}\| = 10$
- 2) $p \cap q = (3, 7)$ and therefore $\|p \cap q\| = 4$
- 3) $\|s_p - s_q\| = 3$, $\|e_p - e_q\| = 3$ and $D_G(p, q) = 0$

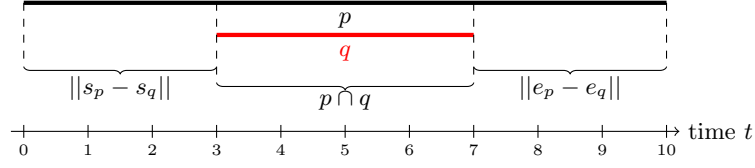


Fig. 3. Visualization of two intervals with their geometrical attributes

3.2 Metadata and Dealing With Deadlines

Like stated in the beginning, it is assumed that the metadata related to the considered time intervals p and q are in mathematically comparable form. That means that for every metadata i there is a continuous distance D_{M_i} with $0 < D_{M_i}(p, q) < 1$ available. The metadata is used to identify, if two intervals are comparable or not. If e.g. machine classes are considered, it measures whether the machines used within the intervals have equivalent functions and are therefore comparable.

A termination criterion regarding interval deadlines is also added, which means that, if interval p is compared with q , we want to make sure that interval q does not end after p has ended. The same goes for a start condition. Therefore, we defined the following two distances.

$$D_{END}(p, q) := \begin{cases} \min\{1, ||e_q - e_p||\} & \text{for } e_q > e_p \\ 0 & \text{else} \end{cases} \quad (6)$$

$$D_{START}(p, q) := \begin{cases} \min\{1, ||s_p - s_q||\} & \text{for } s_p > s_q \\ 0 & \text{else} \end{cases} \quad (7)$$

3.3 Similarity of Two Time Intervals

With the introduced distances, a distance measure for two time intervals is defined, where every characteristic is individually weighted. This measure is then used in chapter 4 to calculate the similarity between two data sets.

Definition 1 (distance between time intervals). *For two intervals p and q , the distance between them is measured by calculating the weighted sum of distances:*

$$S(p, q) := \sum_{i \in I} \lambda_i \cdot D_i(p, q) \quad (8)$$

Thus, the more similar the two intervals p and q are to each other, the smaller the value of $S(p, q)$ is. In the next step, two time interval data sets are compared and the similarity using this approach is evaluated.

4 Similarity Analysis Regarding Time Interval Data Sets

In this section, two interval data sets P and Q are compared. At first, the same cardinality for both P and Q is assumed, that means the number of intervals in each data set is the same. In Chapter 4.2, a procedure for dealing with different cardinalities is introduced. Furthermore, the intervals in P are specified with p_i and q_i for Q . For the remainder of the paper, data sets are considered as disjoint partial sets of a complete, weighted bipartite graph (cf. figure 4), in which the edge weight between two nodes corresponds to the interval similarity measure S .

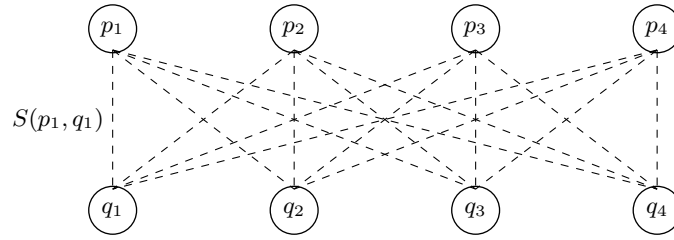


Fig. 4. Representation as a bipartite graph

Hence, the similarity of time interval data sets (STIDes) is equivalent to a perfect matching with minimal weight within our constructed bipartite graph.

Definition 2 (STIDes approach). *Let P and Q be two time interval data sets, $p_i \in P$, $q_i \in Q$ and $|P| = |Q| = n$. Furthermore, Π is the set of permutations of a set with n elements and $\pi \in \Pi$. The similarity between P and Q is determined by the following distance measure*

$$S(P, Q) := \min_{\pi} \left\{ \sum_{i=1}^n S(p_i, q_{\pi(i)}) \right\}_{\pi \in \Pi} \quad (9)$$

Such minimization problems in bipartite graphs can be solved within polynomial time by using for example the Hungarian algorithm [13]. Our approach is therefore capable of calculating a similarity measure within polynomial time while being able to prioritize certain characteristics and measure similarities even with existing time shift. In the next part, we expand this static approach for a dynamic similarity search, which also includes rescaling and shifting possibilities.

4.1 Dynamic Changes Within One Data Set

Until now, the previous static approach has difficulties in determining realistic similarities as soon as one of the time interval sets has big temporal shifts. In Figure 5 we recognize, that the pattern of p_1 and p_2 is the same as the pattern of

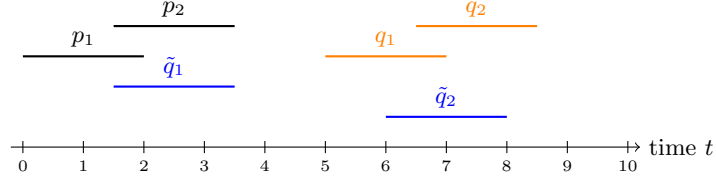


Fig. 5. Similarity regarding big temporal shifts

q_1 and q_2 . Our static approach would determine $\tilde{Q} = \{\tilde{q}_1, \tilde{q}_2\}$ as more similar to $P = \{p_1, p_2\}$. A comparison with true-to-scale model data sets is not provided in the basic configuration either, e.g. in Figure 6 the Set $Q = \{q_1, q_2\}$ is exactly like $P = \{p_1, p_2\}$, only compressed by factor $\frac{1}{2}$. The static algorithm would choose $\tilde{Q} = \{\tilde{q}_1, \tilde{q}_2\}$ like before. However the construction of the interval distances allows an extension of the desired properties. Therefore, we define two kinds of operations.

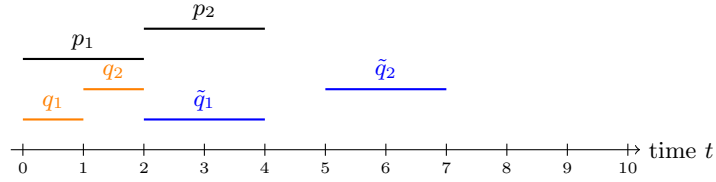


Fig. 6. Similarity regarding true-to-scale model data sets

Definition 3. Let $p = (s_p, e_p)$ be an time interval in short form. Furthermore, let $v \in \mathbb{R}$ be a shift parameter and $s \in \mathbb{R}_+$ a scaling factor. The functions

$$p + v := (s_p + v, e_p + v) \quad (10)$$

$$s \cdot p := (s \cdot s_p, s \cdot e_p) \quad (11)$$

map an interval onto a new interval, hence we can integrate these functions into our similarity measure.

For the similarity analysis of our data sets, this means that we have to solve the following minimization problems:

Definition 4. Let P and Q be two time interval data sets, $p_i \in P$, $q_i \in Q$ and $|P| = |Q| = n$. Furthermore, let Π be the set of permutations of a set with n elements, $\pi \in \Pi$, $v \in \mathbb{R}$ a shift parameter and $s \in \mathbb{R}_+$ a scaling factor. The degree of similarity taking into account global displacement (12) or global scaling

(13) can then be calculated with

$$S(P, Q + v) := \min_{\pi, v} \left\{ \sum_{i=1}^n S(p_i, q_{\pi(i)} + v) \right\}_{\pi \in \Pi, v \in \mathbb{R}} \quad (12)$$

$$S(P, s \cdot Q) := \min_{\pi, s} \left\{ \sum_{i=1}^n S(p_i, s \cdot q_{\pi(i)}) \right\}_{\pi \in \Pi, s \in \mathbb{R}_+} \quad (13)$$

Before dealing with an efficient solver of the above minimization problems, the solubility must be ensured. Therefore, the following is stated.

Lemma 1 (Existence of the Minimum).

Let the conditions of definition 4 be satisfied. The following functions are then continuous with a global minimum.

$$F_1(v) := \min_{\pi} \left\{ \sum_{i=1}^n S(p_i, q_{\pi(i)} + v) \right\}_{\pi \in \Pi} \quad (14)$$

$$F_2(s) := \min_{\pi} \left\{ \sum_{i=1}^n S(p_i, s \cdot q_{\pi(i)}) \right\}_{\pi \in \Pi} \quad (15)$$

Proof. For $\pi_k \in \Pi$ we define

$$f_{\pi_k}^1(v) := \sum_i^n S(p_i, q_{\pi_k(i)} + v) \quad (16)$$

Analogous we define $f_{\pi_k}^2(s)$.

Continuity:

We concentrate on the functions $f_{\pi_k}^1$, $f_{\pi_k}^2$ are valid analogously.

- 1) Let $\Pi = \{\pi_1\}$. $F_1(v) = f_{\pi_1}^1(v)$ is then continuous because it is a sum of continuous distance measures $D_*(p_i, q_{\pi_1(i)} + v)$.
- 2) Let $\Pi = \{\pi_1, \pi_2\}$.

$$F_1(v) = \min \{f_{\pi_1}^1(v), f_{\pi_2}^1(v)\} = \frac{f_{\pi_1}^1(v) + f_{\pi_2}^1(v) - |f_{\pi_1}^1(v) - f_{\pi_2}^1(v)|}{2} \quad (17)$$

is then continuous as a combination of continuous functions.

- 3) Let $F_1(v)$ be continuous for $|\Pi| = n$, then $|\Pi| = n + 1$ holds:

$$F_1(v) = \min \{f_{\pi_1}^1(v), \dots, f_{\pi_{n+1}}^1(v)\} \quad (18)$$

$$= \min \{f_{\pi_1}^1(v), \dots, f_{\pi_{n-1}}^1(v), \min \{f_{\pi_n}^1(v), f_{\pi_{n+1}}^1(v)\}\} \quad (19)$$

and therefore $F_1(v)$ is continuous for $|\Pi| = n + 1$

That means $F_1(v)$ and $F_2(s)$ are continuous functions.

Existence of a minimum:

To show the existence of a minimum, the well known extreme value theorem of Weierstrass¹ is used. The continuity of the functions $F_1(v)$ and $F_2(v)$ was already shown. The last step is to show that there exists an interval $[v_u, v_o]$ (or $[s_u, s_o]$) for which the values of the function $F_1(v)$ (or $F_2(s)$) outside the interval are greater than at least one within. These intervals for both functions are now constructed.

For $F_1(v)$ we define

$$v_u = -|| \max_i (e_{q_i} | q_i \in Q) - \min_j (s_{p_j} | p_j \in P) || \quad (20)$$

as well as

$$v_o = || \max_i (e_{p_i} | p_i \in P) - \min_j (s_{q_j} | q_j \in Q) || \quad (21)$$

Because of the construction of the geometrical distances, that means for every $v > v_o$ (or $v < v_u$):

$$F_1(v) \geq F_1(v_o) \text{ (or } F_1(v) \geq F_1(v_u) \text{)} \quad (22)$$

For $F_2(s)$ we define

$$s_u = \min \left\{ \frac{\min \{ ||p_i|| \}}{\max \{ ||q_i|| \}}, \frac{\min_i (s_{p_i} | p_i \in P)}{\max_j (e_{q_j} | q_j \in Q)} \right\} \quad (23)$$

as well as

$$s_o = \max \left\{ \frac{\max \{ ||p_i|| \}}{\min \{ ||q_i|| \}}, \frac{\max_i (e_{p_i} | p_i \in P)}{\min_j (s_{q_j} | q_j \in Q)} \right\} \quad (24)$$

And analogously $F_1(v) \geq F_1(v_o)$ (or $F_1(v) \geq F_1(v_u)$) holds for $s > s_o$ (or $s < s_u$). That a minimum for $F_1(v)$ (or $F_2(s)$) exists and is located within the interval $[v_u, v_o]$ (or $[s_u, s_o]$) is then shown by the extreme value theorem. \square

4.2 How to Deal With Different Cardinality

If the two disjoint parts of the bipartite graph do not have the same cardinality, the smaller of the two sub-sets is filled with additional nodes, dubbed "dummy nodes". Here, the edge weight of all nodes of the larger subset with the dummy node is set to the maximum occurring edge weight. With the help of this construction, we are able to use the Hungarian algorithm to find the perfect matching within our data sets. In this matching, all intervals which are

¹ A continuous function on an interval $[a, b]$ is bounded on that interval

connected to a dummy node, are not included in the perfect matching. In order to be able to use this result completely in the similarity analysis of time interval data sets, the distance measure must be adapted since the maximum distance measure of all interval pairs has been incorporated into the similarity measure by each dummy node. For convenience, these additional summaries are initially removed from the similarity measure.

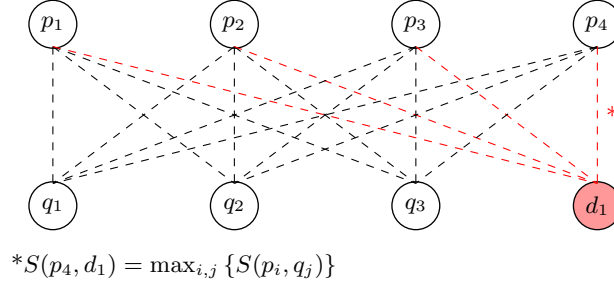


Fig. 7. Bigraph example with one dummy node d_1

In figure 7 we show an example with one dummy node. The adjusted calculation with the STIDES approach then is

$$S(P, \{Q \cup \{d_1\}\}) := \min_{\pi} \left\{ \sum_{i=1}^n S(p_i, q_{\pi(i)}) \right\}_{\pi \in \Pi} - |\{d_1\}| \cdot \max_{i,j} \{S(p_i, q_j)\} \quad (25)$$

The extent to which unmatched intervals influence the similarity measure must be considered according to the individual case and must be adapted accordingly. Another possibility to use data sets with different cardinality and therefore work with rectangular matrices within the Hungarian algorithm, is the algorithm presented by F. Bourgeois and J-C. Lassalle [14].

5 Discussion and Outlook

The STIDES approach is capable of processing different kinds of similarity views because of the capability to set different weight parameters λ_i according to each specific use case. However, this results in an additional effort in the basic setting of the method since the parameters must be set separately for each application. In addition, the creation of dummy nodes allows a determination of the similarity of two unequal data sets, but the remaining intervals do not yet influence the computation of similarity. Our approach incorporates the possibility to apply global changes (e.g. scaling or time shifts) to one of the data sets and we showed that for both scaling and shifting a optimal factor exists, such that the similarity between the two data sets is then optimal.

In the future, we will focus our research on these global changes like time shifts and scaling. We need to research, if the computing time regarding both defined functions 12 and 13 is still polynomial and how the solution can be computed efficiently. We will also investigate the combined effect of both time shifts and scaling. This combined influence can be represented by the structure of the method as a multidimensional function. However, to what extent this affects the complexity of the calculations must also be examined. The possibility to apply different shift and/or scaling factors to different groups of intervals within one data set is also an interesting case, which will be studied in future research. Within the future research, differences in the cardinality of the data sets will again be looked upon to be able to set influence parameters for the similarity measure.

6 Conclusion

At the beginning of this work, it was determined that in today's optimized production processes deviations can lead to unwanted time delays and additional costs. It turned out that a re-recognition of similar deviations from the past leads to a faster and more effective reaction possibility. In order to allow a quantification of similar situations, a similarity criterion on the basis of time interval data sets has been derived in this work, which compares the intervals themselves. For this purpose, a new similarity measure between two intervals was defined, which was transferred to a similarity measure of two data sets in a further step.

In this paper, a similarity measure depending on the relation of the intervals to each other was introduced. For this purpose, the properties of the intervals, such as the size of the overlap, start and end point distances were defined. From these properties distance values were derived, which in a weighted sum form the similarity measure of two intervals. This allows to individually weight each interval characteristic. On the basis of the weighted sum, the STIDes approach was defined, which compares two time interval data sets with one another. For this purpose, the minimum sum of the individual similarities is calculated over all possible interval pairs, which results in the defined similarity measure. An interval pair consists of an interval of each of the two considered time interval data sets. The possibility to weight each interval property is retained by this approach in the extended similarity measure of two data sets. In order to compensate for a possible cardinality difference between the data sets, dummy nodes were introduced, so that each interval can be assigned one partner from the other set and therefore the STIDes approach can be applied. The desired similarity measure of the possibly modified data sets is determined by the Hungarian algorithm in polynomial time ($O(n^3)$). The introduced methodology for identifying similarities also made it possible to incorporate global changes in the intervals of one data set into the analysis. In this context, it has been shown that the defined functions have a global minimum in order to be able to apply the

above described approach, but the complexity changes with the implementation of global changes is not yet researched.

Overall, the approach considered provides a versatile method for describing similarities, in which all properties of the intervals are included in the similarity analysis and, moreover, various types of dynamic changes within the data sets can be mapped. Due to the general representation of this methodology, the similarity analysis can be applied to a variety of problems and thus meets the goal of a general description of similarities between time interval data sets.

References

1. A. D. Jayal, F. Badurdeen, O.W. Dillon Jr. and I.S. Jawahir: *Sustainable manufacturing: Modeling and optimization challenges at the product, process and system levels*. CIRP Journal of Manufacturing Science and Technology 2.3 144-152 (2010)
2. W. B. Lee and H. C. W. Lau: *Factory on demand: the shaping of an agile production network*. International Journal of Agile Management Systems 1.2 83-87 (1999)
3. D. Metzler, S. Dumais and C. Meek: *Similarity measures for short segments of text*. European Conference on Information Retrieval, Springer Berlin Heidelberg (2007)
4. G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu and S. Wang: *GoSemSim: an R package for measuring semantic similarity among GO terms and gene products* Bioinformatics vol. 26, no. 7, 976 - 978 (2010)
5. H. Ogata et al: *A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters* Nucleic acids research 28.20, 4021 - 4028 (2000)
6. A. Wang: *An Industrial Strength Audio Search Algorithm*. ISMIR (2003)
7. O. Kostakis, P. Papepetrou and J. Hollm  n: *ARTEMIS: assessing the similarity of event-interval sequences*. Machine Learning and Knowledge Discovery in Databases, 229 - 244 (2011)
8. P. Meisen, D. Keng, T. Meisen, M. Recchioni and S. Jeschke: *Similarity Search of Bounded TIDASETS within Large Time Interval Databases*. International Conference on Computational Science and Computational Intelligence (2015)
9. J. Kruscall and M. Liberman: *The symmetric time warping algorithm: From continuous to discrete*. Time Warps, String Edits, and Macromolecules: The theory and Practice of String Comparison.
10. Y. Chen, M. Chiang and M. Ko: *Discovering time-interval sequential patterns in sequence databases* Expert Systems with Applications 25.3, 343 - 354 (2003)
11. R. Sadasivam and K. Duraiswamy: *Efficient approach to discover interval-based sequential patterns* Journal of Computer Science 9.2, 225 (2013)
12. C. Koncilia, T. Morzy, R. Wrembel and J. Eder: *Interval OLAP: Analytng Interval Data* International Conference on Data Warehousing and Knowledge Discovery (2014).
13. J. Munkres: *Algorithms for the assignment and transportation problems* Journal of the Society for Industrial and Applied Mathematics 5.1, 32 - 38 (1957)
14. F. Bourgeois and J-C. Lassalle: *An extension of the Munkres algorithm for the assignment problem to rectangular matrices*. Communications of the ACM 14.12, 802-804 (1971)

Financial variables and the real economy: Evidence using a data based procedure of Simultaneous Structural Model Design

Preliminary version currently under revision. Please
do not quote!

Roger Hammersland ^{*†}

Statistics Norway

August 28, 2017

Abstract

By using a new data based procedure of Simultaneous Structural Model Design this paper gives empirical evidence for the existence of a financial accelerator. Hence, credit to firms, asset prices and aggregate investments simultaneously interact over the business cycle in an empirical model of a dynamic economy. Moreover, in this model the interdependency between credit and asset prices creates a mechanism by which the effects of shocks persist and amplify. However, while innovations to credit and asset prices cause short run movements in investments – and vice versa – credit does not independently impinge upon the real trajectory of investments in the long run. A lasting asset price shock on the other hand will have long-term real consequences due to a Tobins Q effect. Besides contributing to reconcile the two opposing views in the literature related to real economy effects of the financial structure, these findings corroborate earlier findings indicating the existence of a long run causal link between asset prices and the business cycle.

Keywords: Financial variables and the real economy, Financial Accelerators, Business Cycles, Simultaneous Structural Model Design, Identification, Structural vector Error Correction modeling, Cointegration.

JEL Codes: C30, C32, C50, C51, C53, E44, E51

^{*}Email address of the corresponding author: Roger.Hammersland@ssb.no

[†]The analyses of this paper have been conducted by using PcGive 10.2 (Doornik and Hendry (2001))

1 Introduction

The idea that credit market conditions may have important effects on an economy's business cycle is today widely accepted, see e.g. Bernanke et al. (1999) and Hubbard (1998). A number of authors do in this context even talk about the existence of a financial accelerator where macro economic effects of shocks to credit conditions may be amplified at the macro economic level, see e.g. Kiyotaki and Moore (1997) and Bernanke and Gertler (1989). Spurred by these theories, a growing empirical literature has provided evidence supporting the existence of a link between indicators of credit availability and macroeconomic fluctuations, suggesting that credit market conditions tend to impact significantly on measures of real activity over the business cycle (for a recent survey see for instance Silvestrini and Zaghib (2015)).¹

However, from a theoretical point of view one may ask why credit should matter in the first place? After all, in a Modigliani and Miller (1958) world with perfect information and no credit constraints, the financial structure should both be indeterminate and irrelevant to real economic outcomes. A natural answer to such an objection would be the lack of realism in the premises of the Modigliani-Miller theory itself. Obviously, in the real world there is nothing like perfect information, and credit constraints are more or less omnipresent. However, to come to Modigliani and Miller's rescue one may plead that the standard assumption of financial structure irrelevance never has had the intention of being fully realistic and that it only must be viewed as a simplification, not to be taken too literally for the short-run evolution of the economy. In the long run, however, when frictions in financial and credit markets play a significantly more subdued role, its relevance should be more compelling. To be able to test the long run relevance of the Modigliani-Miller theorem one should therefore resort to methodologies that explicitly aims at distinguishing between the short- and long-run outcome of a model.

A Financial Accelerator mechanism would necessarily involve the potential existence of mutual causal links between a set of real and financial variables. To unveil such a mechanism would therefore necessitate resorting to a fully simultaneous and structural modeling procedure where the simultaneous causal structure is taken properly into account from the very outset on. However, in doing so, it is important to be aware of some potential pitfalls. For instance in the case of estimating simultaneous equation models that have been exactly identified through e.g. imposing a priori restrictions on their contemporary causal structure and assuming a diagonal structural covariance matrix, one certainly risks inducing a simultaneity bias in estimation through imposing an improper causal structure that does not lie in the data. The reason for this is related to the habit of adding the exact identifying restrictions on parts of the system with a significant bearing on its intrinsic causal structure and the fact that one can never test for the exactly identifying restric-

¹These works are largely based on general equilibrium models pertaining to the Real Business Cycle literature (RBC), see e.g. Kydland and Prescott (1982) and Hartley et al. (1998). To some extent financial accelerator mechanisms have also been implemented in so-called New Keynesian DSGE models, see Smets and Wouters (2007) and Christensen and Dib (2008). Few attempts have been made to incorporate such a mechanism in so-called structural macro-econometric models. An exception is Hammersland and Tr   (2014), where two reciprocal and interacting financial accelerator mechanisms are implemented in a macro econometric model (B  rdsen and Nymoen (2009)) to study the effect of different types of shocks to the financial stability of the Norwegian economy.

tions of a structural model in the first place.² The case where the system is made up of equations that have been individually designed by a process of single equation reductions in a preliminary step – before possibly being put together as a system – bear on the other side witness to the fact that the estimated simultaneous equation model might be the outcome of a design process that is by itself plagued by an intrinsic simultaneity bias. The idea of getting rid of a potential simultaneity bias by putting individually designed equations together in a system and then to re-estimate them simultaneously – after they have found their final form – apparently suggests that the single equation design process itself must have been affected by a simultaneity bias in the first place. Otherwise, there would be no need trying to get rid of it at a later stage.

To address some of the empirical issues raised in the above and to help us in the search for a potential Financial Accelerator mechanism, this paper advocates the use of a pragmatic – and in some sense pluralistic – data based approach where theory and data is set to play harmoniously together in an attempt of identifying the economic structure best at reconciling the information contained in the two independent sources of model design and construction. Theory by contributing to put up an extended theoretical possibility set, and data by playing the role of a judge in choosing among the various alternatives in this possibility set. In the process of model construction this approach is then coupled with a fully simultaneous structural model design procedure, occasionally referred to in the following as the procedure of Simultaneous Structural Model Design (SSMD) (Hammersland and Jacobsen (2008)). In this procedure the preferred simultaneous equation model not only is estimated simultaneously, but is itself the outcome of a fully simultaneous and structural reduction, or design process, where the causal structure of the data has been taken properly into account from the very onset on. It is worth noting that this amounts to an approach where all the behavioral or structural equations of a structural system are reduced and designed jointly. This is therefore an exercise that differs from the less involved one-equation-at-the-time general to specific approach associated with the LSE school of econometrics (see e.g. Hendry (1993, 1995) and Ericsson and Tran (1990)), or for that sake from a SVAR approach where little room is left for design beyond what is implied by the process of exact identification.

Admittedly, the outcome of such a process of Simultaneous Structural Model Design will involve an element of arbitrariness in that it will depend on how the structural model was exactly identified in the first place. To add to the reliability of the final outcome it is therefore prudent to give credence to the identification scheme being used and in doing so,

²When talking about structural models and shocks in the following I am not restricting my models to be derived from an explicit utility maximizing rational representative agent (RA) framework. This means that a structural model, its constituent behavioral equations and shocks are given a far wider interpretation than is given to these concepts in modern micro-based macro theory and refer in principle to theory-driven structural representations in general; be that structures based on more old fashioned type of macro-informed models, models based on emerging macro properties or structures informed by a combination of theory and common sense, including in this structural representations based on an explicit representative agent utility maximizing framework. A consequence of this is that the concept of being structural loses its un-ambiguity as several types of models and shocks can rightly be claimed to have a structural interpretation, though the way they are defined or interpreted as structural will differ across models. This also contributes of course to dilute the proper meaning of the word behavioral, as it makes structural relations based on relational and descriptive macro theories juxtaposed to micro based structural relations. Despite this fact I have in the text chosen to use the word behavioral throughout, keeping in mind that a more appropriate connotation perhaps would have been relational.

not only to the restrictions being imposed, but also to the extent that the auxiliary tools being used to exactly identify the system makes sense, in the sense of having arguably a structural or behavioral interpretation.³In any case, to explicitly describe how the system was identified is clearly associated with a couple of distinct advantages: not only does it serve to guarantee against sweeping the problem of identification under the carpet, but also provides us with a test for over-identifying restrictions that later can be used to inform the structural design process and, in this respect, the imposition of causal restrictions in particular. This stands in contrast to what is common practise in e.g. the SVAR literature where a priori restrictions on the contemporary causal feedback matrix are used to exactly identify the model, often based on some a priori perceived view of delayed reaction. As there is no way to test for these exactly identifying restrictions this necessarily introduces a significant trace of arbitrariness in model design and specification.

To improve on such an obvious short coming this paper advocates resorting to a classical identification scheme that renders importance to a much wider set of identification sources than what is common practise in e.g. the SVAR literature. As far as identification is concerned, such an idea is rooted in a firm belief that one never should rule out any kind of information - extrinsic or otherwise - relevant for the exact identification of a system on purely a priori grounds. Rather, as long as it does not interfere with or directly impede on the ability of coming up with an unbiased analysis, one should in my view seek to throw as much information there is at the problem of resolving the issue of exact identification, and then to subject the outcome based on a given and specific identification scheme to a thorough robustness test where it's validity is tested against alternative ways to accomplish the aim of exact identification.

Some might object that such a strategy is as arbitrary and dependent on the exact identifying restrictions as the procedure I aim at criticizing. However, though I am aware of the fact that there is no such thing as a free lunch when it comes to how one goes about to exactly identify a simultaneous equation model, it is nevertheless my firm belief that ignoring additional and extrinsic information, when it exists, is clearly disadvantageous to using it when it comes to exact identification of structural representations. *In particular, such a strategy will help us avoid laying the exact identifying restrictions on information laden parts of the model – like the contemporaneous feedback matrix – and instead leave such kind of restrictions at the discretion of the data.* An eventual issue with arbitrariness may in this context be sought alleviated by demonstrating the models' robustness to alternative identification schemes.

To study the mutual interplay between financial variables and the real economy and in this respect to see whether it is possible to identify a financial accelerator, a simultaneous structural equation model is constructed on Norwegian aggregate data using our SSMD procedure.⁴ In doing so, particular emphasis has been placed on investigating (i)

³It is in this context important to emphasize that the procedure promoted in this paper is a fully simultaneous structural reduction process with the scope of ending up with an over-identified and parsimonious structural representation. As such, this should render the problem related to ambiguity less of an issue as it directly combines the process of exact identification to the initialization of a search process rather than to the scope of directly ending up with an a priori predetermined final structural representation.

⁴I have chosen to focus on Norway, which happens to be my country of origin and thus the country I know best, both institutionally and otherwise. This choice, however, is also related to the fact that the SSMD procedure advocated in this paper is cumbersome and involved, not to say downright complex,

whether a financial accelerator mechanism has empirical relevance and – if so is the case – (ii) whether it is possible to reconcile such a mechanism with a long run structure that renders the financial structure all but irrelevant for the real economic outcome of the model. To be able to utilize the SSMD procedure advocated herein I have in this respect been forced to keep the dimension of the model down to a minimum due to a relatively few number of observations. The model is thus necessarily simple, and my analysis should therefore be viewed as an attempt to obtain qualitative insights based on data, rather than to provide an empirical description of real financial interactions that aims at being fully realistic.⁵ None the less, in the case of Norway, it turns out that to illustrate the working of a financial accelerator in the setting of a fully simultaneous equation model that adequately and congruently portrays the evolvement of the real economy one can do with a surprisingly small information set.⁶ In fact, in addition to real investments, the information set that forms the basis of my preferred structural vector error correction model comprises only stock prices (domestic as well as global), an indicator for, respectively, domestic credit and the repurchasing cost of capital, interest rates and oil prices.

As regards the outcome of my procedure of Simultaneous Structural Model Design it contributes to reconcile the two opposing views of the literature. In particular I do find evidence of a financial accelerator that is amplified by a credit-asset price spiral in the short run. However, while temporary innovations to asset prices and credit do cause short run movements in production, and vice versa, credit do not independently impinge upon the real trajectory of investments in the long run. A lasting asset price shock on the other hand will have long term real consequences as a result of a Tobins Q effect in the relation pertaining to the long-run investment relationship. Noteworthy, this is in accordance with Beaudry and Portier (2005, 2006), where shocks to stock prices have lasting long-run effects on the US and Japanese real economy.

The remaining sections of the paper are structured as follows. In Section 2, in addition to give some background information, I present some stylized facts related to a potential link between financial variables and the real economy. Section 3 is devoted to a critical discussion of the procedure that these days more or less has got the status of a *come-il-faut* when it comes to how to proceed when exactly identifying structural representations. This is a discussion that has clear implications for the line of approach chosen in the data based design procedure advocated in this paper. In Section 4 I then set up the empirical model framework and run through a modeling exercise with the aim of illustrating the potential of a data-based structural model design procedure and demonstrating how it can be used to shed light on the sources of economic fluctuation, both in the long and short run. Finally, Section 5 offers some concluding comments.

due to a complete lack of automatic structural model reduction procedures. Analyzing another country utilizing the procedure advocated herein, would therefore surely require a non-negligible effort, if not a separate analysis on its own.

⁵However, having said this, one must bear in mind that one of the main purposes of this analysis is illustrative in the sense of aiming at illustrating how to use a particular method of structural modeling and design. To get the point across this often requires making certain sacrifices on the altar of realism.

⁶To my knowledge this is the first analysis to confirm such a mechanism on Norwegian data using a fully simultaneous structural modeling framework (for a further discussion on this point see e.g. Hammersland and Tr   (2014)).

2 Background, Stylized Facts and the Data

2.1 Theoretical Background

It has long been recognized in the literature that in an environment with informational asymmetries, internal finance has a cost advantage over external finance for an entrepreneur considering undertaking a project. Hence, the Modigliani and Miller (1958) theorem does not apply, as internal funds, new debt or equity finance are not perfect substitutes. Lenders who are less informed about e.g. borrower types, borrower action or project quality, will demand a premium when providing uncollateralized loans. This external finance premium will be increasing in the size of the uncollateralized loan, causing financing costs to be higher than if the loan was fully collateralized. Since the agency problem raises the costs of external finance, it will affect wealth-constrained entrepreneurs' willingness to undertake projects. If increased borrower net worth renders possible more internal finance to the funding of the project and/or to raise collateral, then agency costs will be curbed. Thus, a positive shock to net worth will reduce the agency problem and may in turn lower financing costs and increase investments. This inverse relationship between net worth and agency costs of investment finance has a decisive role for many theoretical model predictions. Bernanke and Gertler (1989) develop an overlapping-generations model with costly state verification as in Townsend (1979). The asymmetry of information between lender-investors and borrower-entrepreneurs creates an agency problem where the optimal financial contract is characterized with a deadweight loss due to agency costs. A positive shock to borrower net worth reduces agency costs and increases physical investment. This induces a persistent investment upturn which is not present in the first-best perfect-information case. As a positive shock to net worth is likely to be procyclical, a financial accelerator effect emerges: The positive shock to net worth stemming from a business cycle upturn amplifies the boom.

Other theoretical studies have identified a financial accelerator mechanism where the accelerator - as described in the previous paragraph - is amplified through the working of a credit-asset price spiral. This last feature refers to a mechanism where higher asset prices spur higher credit which in turn leads to still higher asset prices and so forth, due to procyclical asset prices and the behavior of credit constrained investors when given the opportunity to take on more credit. Important contributions in this respect are the seminal articles by Kiyotaki and Moore (1997) and Bernanke et al. (1999).

Summarizing, the models in Bernanke and Gertler (1989), Kiyotaki and Moore (1997) and Bernanke et al. (1999) are all modified real business cycle models where a financial accelerator mechanism may cause large and persistent business cycle fluctuations. In the following, I will refer to the financial accelerator as the mutually reinforcing interaction between asset prices, credit and economic activity. I investigate whether a financial accelerator has empirical relevance, using Norwegian quarterly data for the past twenty years. More specifically, I examine the possibility of interdependence between net worth, credit and investments using classical estimation methods not imposing a priori restrictions (distributional or otherwise) on the model parameters. In doing so I draw heavily on the general to-specific- principle of Hendry (1993), though the design scheme has been cast in a new and fully structural and simultaneous framework. The variables I use are real domestic share prices (an Oslo Stock Exchange index), total credit to non-financial firms

in mainland Norway and real investments mainland Norway. In the analysis I also include an indicator for the repurchasing cost of capital, real oil prices in Norwegian Kroner and a global stock market index, the last two variables commonly seen as being important for developments at the Oslo Stock Exchange. The repurchasing cost of capital, together with the global stock market index and oil prices are all treated as exogenous in the empirical analysis. I search for long-term relationships within the framework of a multi-variate cointegration analysis, and I aim to identify a structural, dynamic Simultaneous Equations Model.

2.2 Stylized facts and the data

As already noted, the empirical analysis comprises the following variables: Real credit to non-financial firms, real share prices, an indicator of real replacement costs, investments in mainland Norway, banks' lending rate and real oil prices in Norwegian currency. This section illuminates a few stylized facts. Figure 1, panel a, below shows developments in real total credit to non-financial firms and a real Oslo Stock exchange index in the period from 1986 to 2014, while Figure 1, panel b, shows the real investment level in mainland Norway and real credit to non-financial firms over the same period. In Figure 2 on the other hand, the evolvement of investments and the real share price index are compared to, respectively, the ratio of share prices to an indicator of repurchasing costs (panel a) and to the real oil price in Norwegian kroner and a global share price index (panel b). Figure 3, panel a and b, illustrates the co-variation between real share prices, real credit and investments measured as percentage change over four quarters.

The Norwegian credit market was deregulated in the early and mid-1980's while interest rates were politically controlled at fairly low levels until end-1986. Not surprisingly, this spurred a sharp rise in credit growth and asset prices. Without discussing causal factors, the fact remains that the government had to deal with a severe banking crisis only a few years later. The banking crisis in the early 1990's coincided with a substantial downturn in the Norwegian economy. After a sharp drop in interest rates following the ERM-crisis in 1992 and breakdown of the fixed exchange rate regime, the economy eventually started to pick up. This was also reflected in rising share prices from 1992, and after a period of economic revival, firms started to increase their debt markedly from end-1996 onwards. As the dot.com bubble burst, Norwegian stock prices fell from 2000 and credit growth stabilized. In 2003, interest rates started to decline to a very low level, and economic activity and share prices boosted until we got the Bear Stearns Bankruptcy early in 2008 and later the financial crisis. Credit to non-financial firms also picked up from 2005 onwards and did not change tack before the financial crisis fully struck with the outbreak of the Lehman Brothers bankruptcy in September 2008. The financial crisis that began in 2008 caused a significant decline in the global real economy and as shown in Figures 1 and 2, took a heavy toll on Norwegian investments which fell by nearly 40 percent from the start of the crisis in early 2008 until the beginning of 2010. By way of comparison, credit only made a knee-jerk correction. After a sharp drop in interest rates, expansionary fiscal policy and a return to growth among Norway's trading partners, investments and the overall economy eventually started to pick up again, but only to be replaced by a violent contraction in oil investments later on by the end of 2013, beginning of 2014. With the subsequent precipitous fall in oil prices from mid-2014 until

early 2016, this is a process that has been going on until our days, though higher oil prices and somewhat improved prospects indicate that Norway might see brighter times ahead.

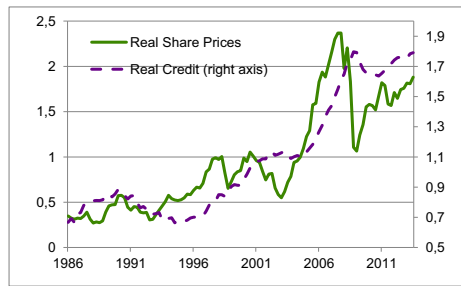
Overall, the figures above indicate a positive correlation on the one hand between share prices and credit to firms and between investment and credit on the other. However, the co-variation of investment and credit seems to be much weaker than the one that applies to share prices and credit, and looking at Figure 2, panel a, one can easily get the impression that the first correlation could be the result of the last, as investments are highly correlated to the ratio of share prices to the repurchasing costs of capital (Tobins Q). Also, though there evidently is a relationship between share prices and credit, Figure 2, panel b, suggests that there are considerably more to share prices than credit, as oil prices and the global share prices seem to capture much of the variation in the Norwegian share price index over time.

My preferred measure of credit to firms and investment relates to mainland-Norway. However, the Oslo Stock Exchange index I use in the empirical analysis also includes offshore activities. Due to the structure of the Norwegian economy, oil prices in addition to a global trend for shares prices, are commonly seen to have a significant bearing on developments at the Oslo Stock Exchange. In addition to a global share index I have therefore chosen to include oil prices denominated in Norwegian kroner as an exogenous variable in my analysis. Figure 3, panel a and b, shows the annual percentage changes in real credit, the Oslo Stock exchange price index and investment. The figure displays a clear and positive correlation between all the time series, though based on pure eyeballing it is difficult to draw any definitive conclusions as regards the causal relationship between the variables. The exceptionally high credit growth in 2000, pictorial in Figure 3, is due to extremely large loan-raising by two Norwegian firms (Telenor and Norske Skog).

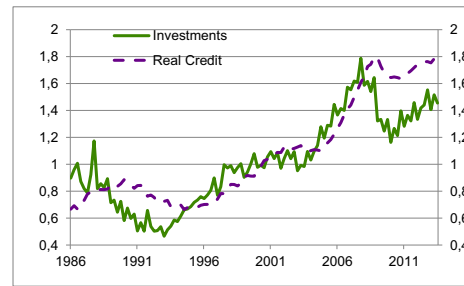
3 Data vs. a priori information in model design

These days a priori information has more or less completely got the upper hand on data in the process of structural model identification and design. For instance, in the structural vector autoregressive (SVAR) and simultaneous equation (SEM) model literature it has been, and still is, common to exactly identify the system by combining the imposition of a diagonal structural form covariance matrix of the errors with either (non-testable) a priori restrictions on the contemporaneous feedback matrix or analogous restrictions on the matrix of parameters that characterizes the long run solution of the system.⁷ Often these kind of restrictions imply a lower or upper block triangular contemporaneous feedback matrix which gives importance to the ordering of the variables in the block diagonal part of the system in that the short run responses implied by the lower or upper triangularity should be in accordance with some perceived a priori view of "delayed"

⁷There is a huge and growing literature in this area and to render justice to all of its contributors is clearly outside the scope of this paper. However, not to mention Sims (1980) seminal paper where he introduces the idea of exact identification through recursive identification would indisputably have to be characterized as an oblivion. Papers that deserve mention for the introduction of restrictions on the systems long run properties are, respectively, Blanchard and Quah (1989), Shapiro and Watson (1988) and Gali (1992).



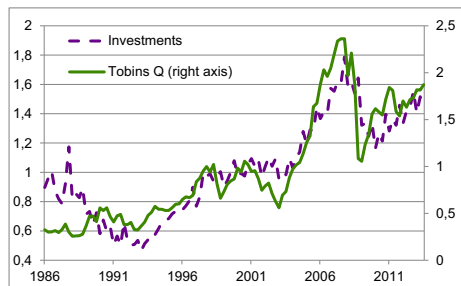
(a) Real share prices and real credit to firms.



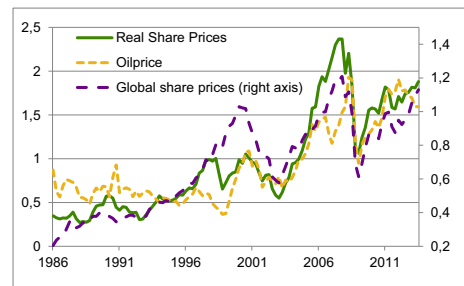
(b) Investments in mainland Norway and real credit to firms.

Indices, 2000Q1=1. 1986Q1-2013Q3.

Figure 1: Credit to non-financial firms, investments and share prices in levels



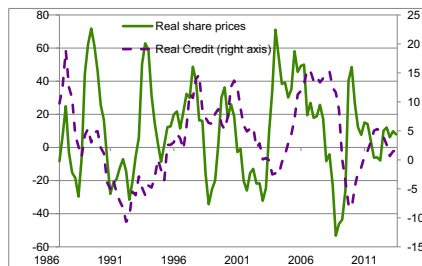
(a) Investments and the ratio of share prices to repurchasing costs (Tobins Q).



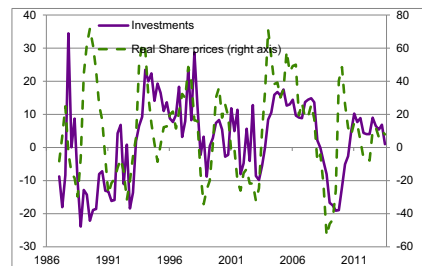
(b) Real domestic and global share prices and the real NOK oil price.

Indices, 2000Q1=1. 1986Q1-2013Q3.

Figure 2: Investments, the ratio of share prices to repurchasing costs (Tobins Q), Domestic and global share prices and the oil price, all in levels



(a) Real share prices and credit to firms.



(b) Investments in mainland Norway and real credit to firms.

Annual growth. Per cent. 1987Q1-2013Q3.

Figure 3: Credit to non-financial firms, Investments and share prices as percentage change over four quarters

reaction.⁸ In general little attention is paid in this literature to the issue of model design beyond what is implied by this process of identification. That is, when the model is exactly identified it has also more or less found its final form.

The inherent problem of structural model design is that there is no way to test for the exact identifying restrictions of a structural or simultaneous equation model. As long as the exact identifying restrictions reflect subjective a priori information of substantial interest and consequence for the properties of the model, this introduces necessarily a significant trace of arbitrariness in model design and specification. In fact, in some cases one might even speak of design where the outcome is more or less fully driven by the researcher's a priori subjective belief or wishful thinking.

Moreover, though imposing the covariance matrix of the structural model's disturbances to be diagonal is theoretically substantiated, the matter presents itself quite differently when constructing empirical models on real data as there is little to suggest that the empirical covariance matrix of an estimated structural form model should inherit the stochastic properties of its theoretical counterpart. This follows both as a consequence of utilizing empirical proxies for theoretical variable constructs and due to the fact that empirical models in most cases are linear approximations of non-linear theoretical equivalents. Add to this the inherent problem of omitted variables and the fact that theories, after all, are revised in light of ongoing scientific theory, there should be no lack of reasons to substantiate why one should be careful with laying the identifying restrictions on the covariance matrix of the disturbances of an empirical model. When all comes to all such a practise would contribute to impair the possibility of developing a data congruent model as it contributes to make the model less elastic when confronted with data. In particular the price paid for securing a structural interpretation of shocks *ex ante* in this respect could be unduly high in terms of miss-specification and lack of congruency.

To reduce the degree of arbitrariness inherent in structural modeling the procedure advocated in this paper strikes a blow for classical identification techniques aimed at giving more emphasis to data in the process of structural model specification and design. The strategy is based on the idea of making the models "more elastic" when confronted with data and thus to avoid laying the exact identifying restrictions on information laden parts of the model and on parts that would make it harder to come up with an admissible and congruent deterministic structure, like the covariance matrix. The advantage of such a strategy should be obvious as after the system is exactly identified tests for over-identifying restrictions are at ones disposal and one can enter into a design process where the data are allowed to speak, i.e. a process where both the ordering of the variables and the contemporaneous structure of the model is the outcome of a testable dialog with the data and not a priori information. As regards the covariance matrix, this advocates a strategy where the structural shock restrictions are tested for and potentially imposed *ex post*, i.e. after the deterministic part of the model has found its final structural form.

Ruling out the use of the contemporaneous feedback matrix and the covariance matrix of the disturbances as sources of exact identification limits the set of ways to exactly identify the system. However, it is important to point out that several alternatives still

⁸Notably there are authors that have tried to avoid the recursive identification scheme, see e.g. Bernanke (1986) and Blanchard and Watson (1986) among others who introduced non-recursive restrictions on the contemporaneous interactions among variables for identification. For sign restrictions see e.g. Paustian (2007) and Uhlig (2005).

remain at our disposal. A classical approach to the problem would for instance imply that one puts to use exogenous information and information about structural breaks. In addition one might utilize information – inferential or otherwise – regarding the long run feedback structure of a model. A priori assumptions related to the lag structure of the structural model is another option (Hammersland (2008)). As regards the first of these alternatives – and to help us with the exact identification part of the model building process in this paper – we have in particular chosen to utilize additional and exogenous information with only a minor qualification; that this information should be structural, in the sense of having a "behavioral" information content or interpretation. To legitimate this being the case one often has to resort to some ad hoc reasoning, a fact that clearly illustrates that there in general is no such thing as a free lunch when it comes to exact identification. Whether one combines the imposition of a diagonal covariance matrix with SVAR-like restrictions on the contemporaneous feedback matrix or utilizes extrinsic exogenous information in the form of structural breaks and exogenous variables one will never be able to fully free oneself from the curse of arbitrariness. However, to ignore using identification promoting exogenous information when it exists, is clearly not optimal in this respect and would represent a huge disservice to the aim of constructing models informed by data. In particular, such kind of extrinsic information would enable us to avoid laying the exact identifying restrictions on information laden parts of the model, and to leave such kind of restrictions at the discretion of the data. A potential problem with ambiguity should in this respect be addressed by requiring that a final structure should be robust as to alternative ways of how to accomplish exact identification.

4 Simultaneous Structural Model Design

To save space I will in this part seek to illustrate the potential of my so-called data based Simultaneous Structural Model Design procedure by running through an explicit modeling exercise, aimed at revealing the structural interplay between real and financial variables. However, before starting on this I will first give a rough outline of the steps involved.

4.1 The procedural steps

The first step of the procedure starts out with the specification of a congruent reduced form VAR model of all model endogenous variables, contingent on a set of exogenous variables, some of which may conceivably be given a structural interpretation in the sense of having a structural rationale related to a subset of the equations of the structural model. To help with the transformation of the reduced form model to a simultaneous equation model or structural form representation later on, a subset of these - so-called structural dummies - are here included in the information set.⁹ The next step of the

⁹Whether a dummy or a variable can be characterized as a carrier of structural information is related to whether the information on which it is based can be regarded as something that is intrinsic to the structural equations of which it is intended to inform, a corollary of this being that a structural variable can only inform a subset of the endogenous variables in a structural form model. As far as the structural dummies used in the analysis are concerned their economic rationale is in this sense related to one and only one of the equations of the structural form. Though not stated explicitly in the above, a defining

procedure then consists of reducing this general reduced form representation down to a parsimonious model and then to use this to identify and estimate the long-run structure of the model. Given this long-run structure the reduced form version of the model is then transformed into an exactly identified simultaneous structural system version thereof, more precisely a Structural Vector Equilibrium Correction Model (SVECM), utilizing restrictions on the feedback matrix related to long-run equilibrium imbalances and/or on exogenous variables as instruments of exact identification, possibly in combination with restrictions on the structural dummies included in the first step of the procedure.

In the last step this exactly identified SVECM is so used as the point of departure for what I believe to be a new design procedure. That is, a kind of simultaneous structural general to specific design process where the entire structural model is reduced down to a parsimonious and over-identified structural specification within a fully simultaneous and structural model framework. As distinct from one-equation-at-the-time model design procedures this final process of reduction takes on a fully simultaneous and structural perspective where all the structural equations are reduced and designed jointly, a corollary of this implying a constant search for uncovering whether identified reduced form relationships reflect direct structural causal links in themselves or are the consequence of effects feeding through via contemporaneous structural causal links between the model endogenous variables. Naturally, in such a context, a restriction imposed on a parameter belonging to one of the behavioral equations in the system would potentially spill over and have consequences for parameters belonging to all or some of the other behavioral equations in the system. A restriction imposed early on may also have clear implications for the type of restrictions it would be possible to impose at a later stage in the reduction process. In the process of model design this kind of simultaneous interdependence therefore involves a substantial degree of trial and error, a feature that contributes to make the process of model design time consuming as well as involved, not least due to the fact that it has been undertaken by hand.¹⁰

characteristic of a structural dummy is also that the information on which it is based comes from sources outside the system under consideration. A structural dummy should thus in principle neither be caused by any of the model endogenous variables nor be correlated with the structural errors of the equations it is intended to structurally inform, the last requirement hinting at an additional absence of so-called nonsense correlations due to omitted variables, like e.g. variables intercepting anticipatory movements (for a discussion of how to get rid of a potential potential bias in estimation due to endogenous and anticipatory movements in single equation conditional models see e.g. Romer and Romer (2004) and Romer and Romer (2010)).

¹⁰The fact that this procedure of reduction is highly informed by theory and a desire of ending up with a model with good interpretable properties is what makes it difficult to automatize. As one reduction imposed early on in the process might turn out to have dire consequences for the possibility of ending up with a model with the desired properties, the process of design will necessarily imply a lot of back and forth searching with theory and interpretation as the rule of conduct. Also, as I in the process of reduction have given priority to theory and interpretation, I have occasionally had to resort to brute force, in the sense of accepting partial reductions that would otherwise have been *marginally* rejected if one exclusively gave priority to the outcome of tests or information criteria. This further complicates the use of automatic reduction procedures as it involves a great deal of judgement as to whether the end justifies the means in some of the individual steps considered in the reduction process. However, what is important to realize in this context is that the structural model I eventually end up with - the so called final structural representation - should lie in the space spanned by its exactly identified point of departure, in the sense of not being rejected by the overall test of over identifying restrictions.

4.2 Structural Model Design: an illustrative example

As regards my illustrative example, I have chosen as a point of departure the error correction version of the vector autoregressive model written in reduced form. In the general case this can be given the following representation:

$$\Delta X_t = \Pi Y_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \Phi D_t + \epsilon_t, \quad (1)$$

where X_t represents a $p \times 1$ vector of endogenous variables, $Y_t = (X_t', Z_t')'$ a $(p+q) \times 1$, vector where Z_t is a $q \times 1$ vector of exogenous variables and k the order of the VAR. D_t is a vector composed of contemporaneous and lagged differences of the model exogenous variables, Z_t , deterministic variables like dummies, a trend and a constant. ϵ_t is a Gaussian white noise term with covariance matrix Ω . The rank of the Π matrix gives us information about the cointegration properties of the model, and in the case the rank, r , is less than full, i.e. less than p , the Π matrix may be written as the product of a $p \times r$ matrix, α , and a $(p+q) \times r$ matrix, β , with full column rank equal to $r < p$. The level term in equation (1) can then be written as $\Pi Y_{t-1} = \alpha \beta' Y_{t-1}$ where $\beta' Y_{t-1}$ represents the r cointegrating linear combinations of the variables while the α matrix has got the interpretation of a coefficient matrix with error correction coefficients or loadings. The cointegration analysis in connection with the preparation of the SVECM¹¹ is based on a three dimensional conditional VAR of order 3,¹² where all the variables are specified as logarithms of the original level series and a trend restricted to lie in the space spanned by the α matrix.¹³ Since I am utilizing unadjusted data centered seasonal dummies were specified to enter unrestrictedly together with a constant and dummies for certain important historical events. As will be evident from the subsequent discussion, some of these dummies can arguably be considered as carriers of structural information in the sense of informing one and only one of the behavioral equations. This is what motivates their role as auxiliary tools of exact identification in the following, though it is important to

¹¹To distinguish the type of structural model developed in this paper from the SVAR model type I have chosen to use the term Structural Vector Equilibrium Correction Model interchangeably with the statistical concepts of a structural form and a simultaneous equation model (SEM).

¹²The VAR of order 3 amounts to a valid reduction of a data congruent VAR of order 6. In this VAR(3) none of the individual equation hypotheses for normality or absence of autocorrelation and heteroscedasticity are rejected at conventional significance levels. The system diagnostics of the VAR(3) model, given below, where the figures in parentheses are the respective tests' significance probabilities, do neither give rise to any concern.

Vector AR 1-5 test:	$F(45, 113)$	=	1.2211[0.1990]
Vector Normality test:	$\chi^2(6)$	=	5.4893[0.4893]
Vector Heterosc. test:	$F(96, 222)$	=	1.0327[0.4166]

The F-test statistic for the elimination of all lags greater than 3 from the model is $F(63, 84) = 1.2966[0.1325]$, where the figure in parenthesis is the test's significance probability. Nor where any of the partial reductions of the model reduction scheme - from a VAR of order 6 down to a VAR of order 3 rejected.

¹³The VAR is conditional in the sense that the model is contingent on the real Norwegian krone price of oil, a global equity index, real interest rates and an indicator of repurchasing cost of capital being exogenous

emphasize that the final structural model in no way is dependent upon these for exact identification. The VAR was then estimated by full maximum likelihood.¹⁴

As regards the historically motivated dummies, several of these turns out to have a potential structural interpretation in the sense of being related structurally to one of the behavioral equations. For instance if we look at the "behavioral" credit equation, the dummy, D2000Q3, is fairly straight-forward in this respect as it represents the influence on the full amount of credit provided to firms of two extraordinary big corporate credit expansions in the third quarter of 2000 (see Section 2 for details). Accordingly, it takes the value of one in 2000Q3 and zero otherwise.^{15,16} Likewise, as far as the aggregate asset price equation is concerned it is hard to disregard two obvious candidate identification instruments, respectively the Bear Stearns bankruptcy on March 17, 2008, and the subsequent collapse of Lehman Brothers only six months later, on September 15. Though the effect of both of these events to a certain extent can be said to already have been taken into account by including a global share price index in our analysis, the possibility of asymmetric responses across countries could be taken to legitimize the inclusion of two distinct dummies with a potential structural interpretation for the purpose of exact identification.¹⁷ Respectively, a dummy that is 1 and -1 in respectively the first and second quarter of 2008 and an impulse dummy that is one in the fourth quarter of 2008, the -1 in the second quarter of 2008 representing the effect of the subsequent stock market rally when it later was learned that Bear Stearns had been rescued by J.P. Morgan Chase. As regards the second dummy we have chosen to place the event in the fourth quarter of 2008 despite the fact that the Lehman Brothers debacle happened in the end of the third, the reason for this being the fact that the index we use represents averages based on daily close values and not end of quarter close values. In principle it is much harder to find shocks with a clear structural origin when dealing with real economy variables like investments or production. However, in the case of Norway I happen to know of an extraordinary big revision made to the original figures due to the sale of used aircrafts abroad in the first part of 2002. To be able to take this into account, I have included a dummy which equals 1 in the second quarter of 2002 as an auxiliary identification tool.

As regards the line of reasoning being used to legitimize whether a dummy is to be

¹⁴In this context it is, as pointed out by Johansen (2006), worth noting that there is a price paid by using maximum likelihood in estimating VARs. Namely that the model must fit the data in the sense of constituting a congruent representation of the data generating process (DGP). However, in light of Footnote 12, this requirement does not seem to represent any cause for concern in this case.

¹⁵There is little to indicate that this credit expansion was related to the dot.com crisis in early 2000, a contingency that would have clear negative implications for its use as a valid identification instrument.

¹⁶In this paper I have not tried to identify separate equations for credit supply and demand, the main reason for that being a lack of specific aggregate supply and demand data. Admittedly I could have circumvented this problem by utilizing the procedure advocated in this paper on the sub structure related to the provision of aggregate credit. However, as this - not least due to a limited number of observations - would have complicated matters considerably without necessarily leading to new insights as far as the relevance of the financial accelerator is concerned, I have deliberately left this for future research. As a consequence I cannot tell whether the dummy, D2000Q3, has a structural interpretation in the sense of being related either to the supply or demand of credit. It could be related to both.

¹⁷Alternatively one could choose to impose the identifying restriction on the global share index directly, restricting it to only inform the structural asset price equation. Based on the final structural model in (4) this turns out to be a restriction that would amount to a valid alternative to using the two structural dummies.

considered as a carrier of structural information I have no problem admitting that it in a couple of the instances referred to above, is rather ad hoc. I therefore want to stress the importance of looking at the possibility of supplementing – or even substituting – the use of extrinsic information in terms of structural dummies with other kind of identification promoting restrictions, like e.g. restrictions placed on the long-term feedback- or even lag structure - of the system or on exogenous variables with a definite structural interpretation. After all – and a point I implicitly seek to convey in this paper – one should in my opinion never on a priori grounds rule out any kind of information, extrinsic or otherwise, relevant for the exact identification of a system. Rather, as long it does not directly interfere with or directly impede on the ability of coming up with an unbiased analysis, one should seek to throw as much information there is at the problem of resolving the issue of exact identification, not least to be able to test for a potential path dependency later on. In other words, in the presence of a perceived threat to the validity of the outcome of the analysis this stresses the importance of testing for the robustness of the identification scheme being used, be it whether one uses structural dummies or not.

Compared to alternative approaches, it is important to be aware of the fact that the exact identification procedure advocated in this paper comes with an advantage. That is, it renders possible formal tests of causal restrictions without resorting to the imposition of incredible restrictions on either the structural empirical covariance matrix or the contemporaneous feedback matrix. Put differently, by resorting to what amounts to no less than a purely classical identification scheme utilizing various kinds of extrinsic information, we have been able to move the act of exact identification to a higher level in the design process, rendering possible data informed testing and design of the model's contemporaneous causal structure without resorting to un-testable restrictions on information laden parts of the model.

Based on the discussion used to legitimate the appropriateness of the auxiliary tools used to help with the exact identification of the structural representation of this paper, I therefore move on to a structural analysis, feeling rather confident that the identification scheme being proposed would represent a contribution to the goal of revealing important aspects of the true underlying causal structure. However, before doing so, I will first return to the reduced form analysis and the identification of the model's long-run structure.

The results of the reduced form cointegration analysis is given in Table 1 and Table 2 and give unambiguous support for the existence of three cointegrating vectors. Moreover, the F-test for the number of over-identifying restrictions in Table 2, shows that the identified system, consisting of three cointegrating relationships, constitutes a valid restriction of a corresponding exactly identified long-run structure.¹⁸

The first of the structural long-run relationships in Table 2 implies that aggregate

¹⁸As indicated by the final and over-identified log-run structure, the long-run structure can in principle be exactly identified in many alternative ways without compromising the final outcome of our long-run analysis. A relatively uncontroversial set of exactly identifying restrictions in this respect is the following: 1) to assume that real equity or asset prices only affect investments as far as a change in equity prices represents a change relative to the replacement costs of capital and that there is no effect of global equities on investments. And 2) to assume that credit in the long run is neither driven by the repurchasing cost of capital or the foreign equity price index. And finally 3) to assume that asset prices in the long run are homogeneous of degree one in oil prices and global equity prices, at the same time as there is no direct effect of replacement costs on asset prices.

Table 1: **Johansen's test for the number of cointegrating vectors**

VAR order: 3, constant and trend restricted to lie in the α space, unrestricted centered seasonal dummies. Estimation period: 1990 Q4 to 2013 Q3.

Trace Eigenvalue test:		
H_0	H_1	Values of test statistics
$r=0$	$r \leq 3$	130.22[0.000]**
$r \leq 1$	$r \leq 3$	63.771[0.000]**
$r \leq 2$	$r \leq 3$	27.599[0.000]**

¹⁾ The values in parentheses are the respective tests' significance probabilities.

²⁾ * and ** signify that the test is significant at a level of 5 and 1%, respectively.

investments in the mainland part of the Norwegian economy behave in accordance with the Tobin's Q theory. Accordingly, for a given interest rate, an increase in the ratio of equity prices to the replacement cost of capital is identified to affect investments positively, a one percent increase estimated to lead to an aggregate increase in investments of approximately 0,34 percent. The investment equation also identifies a separate real interest rate effect. For a given Q, a one percentage point increase in the real interest rate is here estimated to lead to an investment decline of about 4 per cent.¹⁹

The second cointegrating relationship implies on the other hand that a weighted ratio of domestic credit of enterprises to equity prices is constant over time, which due to the logarithmic specification and a small abuse of terminology amounts to saying that a percentage increase in the equity price feeds into an increase in domestic credit of enterprises in the long run, estimated to 0,7 per cent. To substantiate what was here hinted at, namely that the causal link between credit and equity prices goes from equity prices to credit, requires a full-fledged structural analysis. However, before starting on such a task one may get some idea as to how the causal structure might look like by taking a closer look at the error correction coefficient matrix, α , of the reduced form. I will return to this immediately after having discussed the third long-run cointegrating relationship.

Finally, the third cointegrating relationship is a long-run asset price relationship homogeneous of degree one in real Norwegian krone oil- and global equity prices and with a separate negative real interest rate effect. A one percent increase in real oil prices for a given level on the global equity price index is in this equation estimated to lead to an increase in equity prices of approximately 0,67 per cent, leaving the effect of a similar change to the global equity index - keeping oil prices fixed - at 0,33 per cent. A one percentage point increase in the real interest rate is in this equation estimated to lead to a decline of 5,6 per cent in equity prices. Given the significant role played by oil in

¹⁹It is worth noting that the output gap, as estimated in this way, is fairly similar to that presented in Norges Bank (2006) from 1996 up until 2006 (See Inflation Report 1/06 on <http://www.norges-bank.no/>). It is also almost identical to the output gap relationship estimated in Hammersland (2008), using a slightly different information set.

Table 2: **The identified system of cointegrating linear combinations given $r=3$, the loading matrix and a test of overidentifying restrictions** ¹⁾

The identified long run structure given 3 cointegrating relations:

$$\hat{\beta}' \begin{pmatrix} Y_t \\ \text{TREND}_t \end{pmatrix} = \begin{pmatrix} i_t + 0.043 RR_t - 0.34 \{s - rc\}_t \\ (0.005) \\ c_t - 0.7 s_t \\ s_t + 0.056 RR_t - 0.33 msci_t + 0.67 poil_t \\ (0.007) \quad (0.002) \end{pmatrix}$$

Error correction coefficient matrix:

$$\begin{matrix} \Delta i \\ \Delta c \\ \Delta s \end{matrix} : \begin{pmatrix} \hat{\alpha}_{11} & \hat{\alpha}_{12} & \hat{\alpha}_{13} \\ \hat{\alpha}_{21} & \hat{\alpha}_{22} & \hat{\alpha}_{23} \\ \hat{\alpha}_{31} & \hat{\alpha}_{32} & \hat{\alpha}_{33} \end{pmatrix} = \begin{pmatrix} -0.24 & 0.037 & -0.047 \\ (0.02) & (0.06) & (0.05) \\ 0.087 & -0.142 & -0.035 \\ (0.021) & (0.023) & (0.018) \\ 0.04 & 0.037 & -0.265 \\ (0.07) & (0.064) & (0.062) \end{pmatrix}$$

LR-test of overidentifying restrictions: $\chi^2(12) = 20.150[0.0643]$ ¹⁾ The value in parenthesis under each coefficient is the estimated coefficient's standard error while the value in parenthesis following the test of over-identifying restrictions refers to the test's significance probability. Note that the test of over-identifying restrictions refers to the restrictions one will have to impose on an exactly identified structure to arrive at the final structure given by the system's right hand side. The variables i_t , c_t , s_t , $msci_t$, rc_t and $poil_t$ stand for, respectively, real mainland investments, real domestic credit to enterprises, real equity prices, a global equity index, the repurchasing cost of capital and the real price of oil in Norwegian kroner, lower case letters indicating that all the quantities are logarithmic transformations of the original variables referred to in the text. RR_t stands for the real interest rate and has not been transformed logarithmically. The vector to the left of the loadings matrix and before the colon refers to the individual equations in the corresponding reduced form VECM representation. As usual the Δ symbol stands for the first difference operator.

the Norwegian economy, the fact that oil price fluctuations contribute significantly to explain the evolution of Norwegian equity prices should hardly be surprising, neither the importance of the global equity price index.

As regards the loading matrix, most of its entries are significantly estimated. This contributes to hamper its usefulness as a device to come up with qualified guesses as to the shaping of the contemporaneous feedback matrix of the model's structural form in the following. However, the fact that the second error correction term does not seem to enter the reduced form real investment equation could be taken to indicate that the direction of contemporaneous dynamic causality goes from activity towards credit and not vice versa. Otherwise we do observe that the first error correction term enters with positive coefficients in the reduced form equations of both credit and asset prices, the positive

coefficient indicating that the output gap could be playing an independent behavioral role in both equations, another alternative being that it only enters in one of the behavioral equations and feeds into the other variable's reduced form equation through a contemporaneous causal link. As regards the causal link between credit and asset prices, the fact that the second cointegrating vector (the ratio of credit to asset prices) is insignificant in the reduced form asset price equation at the same time as the deviation of asset prices from its own long-run relationship is found to affect aggregate credit, suggest a one-way causal dynamic relationship between credit and asset prices, going from asset prices to credit and not vice versa. Notably, this would be in accordance with the kind of causal structure hinted at when interpreting the long run cointegrating relationship between asset prices and credit earlier on. However, despite these presumptive deliberations I have chosen to uphold the possibility of a fully simultaneous and unfettered causal structure as a starting point when designing my structural model in the following.

The model that so far has been analyzed is a reduced form representation of the variables in our information set. To be able to explicitly address the topic of dynamic contemporary causality and to construct a model that is more in accordance with the idea of economic data generating processes by nature being simultaneous and structural, we will now move on and, on the basis of the reduced form analysis, develop a simultaneous structural equation model for our three model-endogenous variables. However, before presenting the results of this modeling exercise I will first turn to a brief discussion of the scheme being used to exactly identify the structural system.

The structural form or SEM representation of the reduced form is obtained by multiplying (1) by a contemporary response matrix B . This results in the simultaneous equation system:

$$B\Delta X_t = B\Pi Y_{t-1} + \sum_{i=1}^{k-1} B\Gamma_i \Delta X_{t-i} + B\Phi D_t + B\epsilon_t,$$

or after having set $B\Pi = B\alpha\beta' = \alpha^*\beta'$, $B\Gamma_i = \Gamma_i^*$, $B\Phi = \Phi^*$ and $B\epsilon_t = u_t$

$$B\Delta X_t = \alpha^*\beta' Y_{t-1} + \sum_{i=1}^{k-1} \Gamma_i^* \Delta X_{t-i} + \Phi^* D_t + u_t \quad (2)$$

Given the three previously estimated long run relationships and the fact that the cointegration analysis was undertaken on a VAR(3), (2) will have the representation given by (3) where we have normalized the contemporary response- or feedback matrix such that the coefficients along the main diagonal is equal to one. Furthermore, in (3) we have split the exogenous variable vector, D_t , into two parts. One containing exclusively the structural dummies used to help with the exact identification of our structural model and another one, \tilde{D}_t , containing contemporaneous and lagged differences of the exogenous variables and a couple of non-structural historic dummies. The constant and seasonal dummies have been suppressed for expositional purposes.²⁰ As regards the two non-structural historical dummies these are, respectively, D1992Q4 and D2003Q1, the first one representing a dummy for the collapse of the ERM exchange rate system in the fourth quarter of 1992

²⁰In the continuation I will refer to the matrices $\begin{pmatrix} \lambda_{11}^* & \lambda_{12}^* & \lambda_{13}^* & \lambda_{14}^* \\ \lambda_{21}^* & \lambda_{22}^* & \lambda_{23}^* & \lambda_{24}^* \\ \lambda_{31}^* & \lambda_{32}^* & \lambda_{33}^* & \lambda_{34}^* \end{pmatrix}$ and

and the second one a dummy for the abrupt change in monetary policy at the end of 2002, beginning of 2003. Both of these dummies take the value 1 in the quarter when the event actually took place.

$$\begin{aligned}
& \begin{pmatrix} 1 & b_{12} & b_{13} \\ b_{21} & 1 & b_{23} \\ b_{31} & b_{32} & 1 \end{pmatrix} \begin{pmatrix} \Delta i_t \\ \Delta c_t \\ \Delta s_t \end{pmatrix} = \sum_{i=1}^2 \begin{pmatrix} \gamma_{11.i}^* & \gamma_{12.i}^* & \gamma_{13.i}^* \\ \gamma_{21.i}^* & \gamma_{22.i}^* & \gamma_{23.i}^* \\ \gamma_{31.i}^* & \gamma_{32.i}^* & \gamma_{33.i}^* \end{pmatrix} \begin{pmatrix} \Delta i_{t-i} \\ \Delta c_{t-i} \\ \Delta s_{t-i} \end{pmatrix} \\
& + \begin{pmatrix} \alpha_{11}^* & \alpha_{12}^* & \alpha_{13}^* \\ \alpha_{21}^* & \alpha_{22}^* & \alpha_{23}^* \\ \alpha_{31}^* & \alpha_{32}^* & \alpha_{33}^* \end{pmatrix} \begin{pmatrix} i_t + 0.043 RR_t - 0.34 \{s - rc\}_t \\ c_t - 0.7 s_t \\ s_t + 0.056 RR_t - 0.33 msci_t + 0.67 poil_t \end{pmatrix}_{t-1} \\
& + \begin{pmatrix} \phi_{1,1}^* & \phi_{1,2}^* & \phi_{1,3}^* & \phi_{1,4}^* & \phi_{1,5}^* & \phi_{1,6}^* & \phi_{1,7}^* & \phi_{1,8}^* & \phi_{1,9}^* & \phi_{1,10}^* & \phi_{1,11}^* & \phi_{1,12}^* & \phi_{1,13}^* & \phi_{1,14}^* \\ \phi_{2,1}^* & \phi_{2,2}^* & \phi_{2,3}^* & \phi_{2,4}^* & \phi_{2,5}^* & \phi_{2,6}^* & \phi_{2,7}^* & \phi_{2,8}^* & \phi_{2,9}^* & \phi_{2,10}^* & \phi_{2,11}^* & \phi_{2,12}^* & \phi_{2,13}^* & \phi_{2,14}^* \\ \phi_{3,1}^* & \phi_{3,2}^* & \phi_{3,3}^* & \phi_{3,4}^* & \phi_{3,5}^* & \phi_{3,6}^* & \phi_{3,7}^* & \phi_{3,8}^* & \phi_{3,9}^* & \phi_{3,10}^* & \phi_{3,11}^* & \phi_{3,12}^* & \phi_{3,13}^* & \phi_{3,14}^* \end{pmatrix} \tilde{D}_t \\
& + \begin{pmatrix} \lambda_{11}^* & \lambda_{12}^* & \lambda_{13}^* & \lambda_{14}^* \\ \lambda_{21}^* & \lambda_{22}^* & \lambda_{23}^* & \lambda_{24}^* \\ \lambda_{31}^* & \lambda_{32}^* & \lambda_{33}^* & \lambda_{34}^* \end{pmatrix} \begin{pmatrix} D2000Q3 \\ D2002Q2 \\ ID2008Q1 \\ D2008Q4 \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix}
\end{aligned} \tag{3}$$

However, as regards estimation of (2) and (3), there evidently is a puzzle to resolve as neither of the two representations are identified – in the sense of representing a one-to-one mapping of the corresponding reduced form – without imposing further restrictions.²¹ In the SVAR literature this problem is solved by assuming: i) a lower or upper triangular response matrix and ii) a diagonal empirical structural covariance matrix. However, as already mentioned in Section 3, there is an inherent and insuperable problem associated with imposing exactly identifying restrictions on a structural form in this way as the exactly identifying restrictions never can be tested for. Thus, one evidently runs the risk of imposing a dynamic contemporary structure that is not supported by data. This

as, respectively, the Λ and Φ matrix. $\tilde{D}_t = (\Delta poil_t, \Delta poil_{t-1}, \Delta poil_{t-2}, \Delta msci_t, \Delta msci_{t-1}, \Delta msci_{t-2}, \Delta RR_t, \Delta RR_{t-1}, \Delta RR_{t-2}, \Delta rc_t, \Delta rc_{t-1}, \Delta rc_{t-2}, D1992Q4, D2003Q1)$ (See Table 2 for a definition of all the variables involved).

²¹Note that if we multiply (2) with an arbitrary non-singular F-matrix, the corresponding reduced form will still be equal to (1). This illustrates that there in general does not exist a one-to one mapping between the reduced form and a SEM, or structural form representation. Only in the case where the only admissible transformation matrix, F, is equal to a diagonal matrix, or in the case of (3) where we have normalized the coefficients along the main diagonal of the feedback matrix to be equal to one, the identity matrix, will the simultaneous equation system be identified.

advocates a strategy where one leaves the parts of the system perceived to be of minor importance for the purpose of exact identification and then to consider over-identifying restrictions to test for restrictions on more information laden parts of the model and parts that would make it harder to come up with an admissible and congruent deterministic structure, like, respectively, the feedback- and the covariance matrix.

To help out with the exact identifying part of the modeling building process this paper resorts to classical identification techniques where additional and extrinsic information in the form of structural breaks and exogenous variables – possibly coupled with restrictions on the long-run feedback structure of the system – plays the role as auxiliary tools of exact identification. By initially limiting ourselves to look at the case where we only resort to structural breaks this means that restrictions are exclusively placed on the coefficients of the Λ matrix in (3) and in such a way that the dummies only affect the behavioral equation of which they are intended to structurally inform. Given the set of four structural dummies specified in (3) and the fact that we have chosen to stick to the 2000Q3 and 2002Q2 dummies both being genuinely structural in the sense of informing, respectively, only the structural credit and investment equation, this gives us two different ways to exactly identify our SVECM. The first one using the first three-tuple of dummies (D2000Q3, D2002Q2, ID2008Q1) and a second one where the dummy for the Bear Stearns bankruptcy in the first quarter of 2008, ID2008Q1, is substituted with the dummy for the Lehman Brothers debacle on September 15 2008, D2008Q4. In line with such an identification scheme and using the first three-tuple of dummies, an exactly identified SEM representation is given by a version of (3) where the Λ matrix is restricted such

that it equals $\begin{pmatrix} 0 & \lambda_{12}^* & 0 & \lambda_{14}^* \\ \lambda_{21}^* & 0 & 0 & \lambda_{24}^* \\ 0 & 0 & \lambda_{33}^* & \lambda_{34}^* \end{pmatrix}$.^{22,23,24}

As already mentioned, the structural status of some of the dummies referred to above is rather unclear. Combined with an imminent danger of rendering the identification scheme weak, not least due to a general lack of break observation periods, this makes it prudent to look for alternative and supplementary sources of exact identification, like e.g. the long-run feedback structure of the system and some of its exogenous variables. As far as the first of these sources is concerned there are often good reasons to believe that error correction is a structural property, in the sense of representing a mechanism that is specific to one -or at least not more than to a limited subset- of the structural equations involved. For example in the error correction structure identified in Table 2 there is much

²²The alternative way to exactly identify the system using structural dummies, beyond the one used as a default would have given a Λ matrix equal to $\begin{pmatrix} 0 & \lambda_{12}^* & \lambda_{13}^* & 0 \\ \lambda_{21}^* & 0 & \lambda_{23}^* & 0 \\ 0 & 0 & \lambda_{33}^* & \lambda_{34}^* \end{pmatrix}$.

²³This statement is related to the order condition. However, the claim that the system is exactly identified hangs evidently on the rank condition also being fulfilled, which thus is tacitly and implicitly assumed in this assertion.

²⁴To anticipate events somewhat I may already at this point mention that the final outcome of my procedure of Simultaneous Structural Model Design by chance turns out to be robust as to both identification schemes just mentioned. That, is whether the first or second alternative referred to above is used to exactly identify my system the final outcome will nonetheless be the same and accepted by the test of over-identifying restrictions. More importantly, and an issue to which I now wish to draw the reader's attention, the final model also turns out to be robust with respect to many other ways to achieve exact identification

to indicate that the last cointegrating vector should pertain to the structural asset price equation only, while the second one in principle could be related to both the structural credit and asset price equation, but not the structural investment equation. The status of the first cointegrating vector is, however, more questionable in this respect due its alternative interpretation of an output gap.

In conformity with such an arrangement I present an alternative identification scheme where the process of exact identification is sought accomplished through a combination of imposing restrictions on the long-run feedback structure of the system and on the coefficients of some of its exogenous variables. As far as the long-run cointegration structure is concerned I have restricted the second and third cointegrating vector to only enter the structural equations to which they – on a priori grounds – are supposed to structurally inform. To be more explicit this means that I have imposed the third cointegrating vector to only enter the structural asset price relationship while the second one in principle is allowed to inform both the structural credit and asset price relationship. To avoid wavering a direct testing of the structural property of the output gap effect in the reduced form asset price relation I have at the outset tried to avoid laying any restrictions on the feedback structure related to the first cointegrating vector.²⁵ Combined with a zero restriction on one of the ϕ coefficients related to the dynamic effect of contemporary changes to global equity prices in the credit equation, these restrictions should be sufficient to accomplish exact identification of respectively the first and second structural relationship in (3). What remains is to identify the last equation as a linear combination containing the two first equations still will be illegible as a candidate structural equity price equation without further restrictions being imposed. However, a close look at the partially identified structure reveals that we may overcome such a problem by simply imposing some additional restrictions on the third row of the Φ matrix, implying e.g. that the lag structure of the repurchasing costs variable does not affect the structural equity price relation directly. Hence, by combining the imposition of restrictions on the coefficients related to some of the exogenous variables with restrictions on the long-run feedback structure of the system we have in this way been able to exactly identify the system without resorting to structural breaks. The exactly identified SVECM representation would in this case be given by a version of (3) where the α matrix has been restricted to be lower diagonal and the Φ matrix equal to:

$$\begin{pmatrix} \phi_{1,1}^* & \phi_{1,2}^* & \phi_{1,3}^* & \phi_{1,4}^* & \phi_{1,5}^* & \phi_{1,6}^* & \phi_{1,7}^* & \phi_{1,8}^* & \phi_{1,9}^* & \phi_{1,10}^* & \phi_{1,11}^* & \phi_{1,12}^* & \phi_{1,13}^* & \phi_{1,14}^* \\ \phi_{2,1}^* & \phi_{2,2}^* & \phi_{2,3}^* & 0 & \phi_{2,5}^* & \phi_{2,6}^* & \phi_{2,7}^* & \phi_{2,8}^* & \phi_{2,9}^* & \phi_{2,10}^* & \phi_{2,11}^* & \phi_{2,12}^* & \phi_{2,13}^* & \phi_{2,14}^* \\ \phi_{3,1}^* & \phi_{3,2}^* & \phi_{3,3}^* & \phi_{3,4}^* & \phi_{3,5}^* & \phi_{3,6}^* & \phi_{3,7}^* & \phi_{3,8}^* & \phi_{3,9}^* & 0 & 0 & \phi_{3,12}^* & \phi_{3,13}^* & \phi_{3,14}^* \end{pmatrix}.$$

Based on different varieties of the two exactly identified example structures referred to above where the two sets of exactly identifying restrictions might be combined and amended in a number of different ways, we are now ready to get down to the process of reducing our exactly identified simultaneous equation model down to a parsimonious structural representation of the information contained in our data set. The fact that

²⁵ As the final results will illustrate I could alternatively have utilized a more stringent set of a priori restrictions related to the long-run feedback structure of the system, one of these implying that there should be no feedback effect in the asset price equation related to long-run disequilibrium investment imbalances. For that matter this is also suggested by the fact that the error correction coefficient associated with the first cointegrating vector (the output gap) in the reduced form equity price relationship -though correct sign - is not significantly estimated.

we by resorting to restrictions on the long-run feedback matrix and on exogenous variables with a plausible structural interpretation, possibly in combination with the use of structural dummies, have managed to avoid laying the exactly identifying restrictions on the contemporaneous response matrix – or for that sake the covariance matrix – implies that we by way of tests for over-identifying restrictions now are in full command when designing the contemporaneous causal structure of the model. The result of this process of simultaneous structural model design is our preferred parsimonious structural model, given by (4).

The first thing to notice in (4), and as already alluded to in the second last paragraph, is that the final structural model is robust to many different ways of how to initiate the structural design or search process, including in this the three alternative example schemes explicitly referred to in the text. As far as the three example schemes are concerned this follows as a consequence of the fact that the union of the identifying restrictions imposed when resorting exclusively to structural breaks (including both alternatives) or a combination of restrictions on the long-run feedback structure with restrictions on the exogenous variables, by chance happens to be fully contained in (4).²⁶ Combined with the totality of restrictions imposed on the dynamic structure of our final structural model representation in (4), this contributes to make the final outcome of our SSMD procedure robust against a fairly big set of alternative ways to initiate the structural design process, or put differently, of how to exactly identify our structural model in the first place.

Furthermore, it is important to bring attention to the fact that the test of the over-identifying restrictions does not reject the null hypothesis of the final parsimonious simultaneous equation model in (4) constituting a valid reduction of an exactly identified version of the model. The system diagnostics given below my preferred structural representation also indicate that the system describes data fairly well, as none of the standard vector tests indicate presence of autocorrelation, non-normality or heteroscedasticity. Moreover, the single equation and vector stability tests in an accompanying note to this paper demonstrate that the system as such is relatively stable over the estimation period as in fact none of the recursive tests breaks a test level of 1%. Moreover, when estimating the model on data up to and including 2007 Q2 and using it to simulate dynamically 23 periods ahead, Figure 4 demonstrates that the model performs surprisingly well given the fact that the structural dummies related to the financial crisis both were taken out before running the simulations. In particular the model is able to foresee both the unexpectedly strong drop in asset prices following in the aftermath of the Lehman Brothers default on September 15 2008, and the subsequent recoil from 2009 and onwards.

In the SVECM model of (4) another important thing to note is that the contemporaneous feedback matrix is neither upper nor lower triangular. Given that the structural covariance matrix is far from diagonal this gives implicit support to the invalidity of an identification scheme where these kind of restrictions are used to initialize the design process. In particular, the contemporaneous feedback matrix reveals in this context a con-

²⁶In particular, looking at the final over-identified structural model in (4), we see that all the structural break dummies enter the system in accordance with their a priori structural intention or rationale. For instance the dummies D2008Q1 and D2008Q4, earlier being both characterized as carriers of structural information related to the process governing equity prices, do only enter the structural equity price equation. Likewise, the dummies characterized as carriers of structural information related to, respectively, investments, D2002Q2, and credit, D2000Q3, do only enter the structural investment and credit equation in the final model setup.

temporaneous two-directional causal link between credit and real activity as measured by investments. Accordingly, a productivity shock that momentarily leads to higher investments and activity, will feed into a contemporaneous increase in credit growth. Higher credit growth will on the other hand spur further growth in investments, a feature that reveals a mechanism through which the initial shock to productivity is amplified. Evidently, (4) is thus characterized by the existence of a financial accelerator mechanism in the very short run. Furthermore, in (4) growth in real domestic credit to firms is contemporaneously affected by growth in equity prices, at the same time as equity prices are affected by credit growth. The dynamic interaction between credit and asset prices thus turns out to be a transmission mechanism by which the effects of shocks could persist and amplify. This is a feature that is given some support by looking at the impulse responses to shocks in the accompanying note to this paper. The impulse response analysis demonstrates moreover that the model implies cyclical fluctuations and persistence that gradually dies out in the long term in the wake of shocks. Independent innovations to borrower net worth²⁷ are thus initiating sources of real fluctuations. This stands in contrast to the perfect information case, but is consistent with a model where agency costs introduces cyclical fluctuations into an environment which is not rigged to exhibit such a feature in the long run, when agency costs are not present (see e.g. Bernanke and Gertler (1989)).²⁸ As regards the long-run structure of our model there is as we have seen no long-run link between credit and investment. Hence, while temporary innovations to stock prices and credit do cause short run movements in investments, credit do not independently impinge upon the real trajectory of investments in the long run. A lasting asset price shock on the other hand will have long term real consequences as a result of the Tobins Q effect in the investment relation pertaining to the long run outcome of the model.²⁹

²⁷ And which in the wake of induced changes to agency costs, would lead to a redistribution of income between borrowers and lenders.

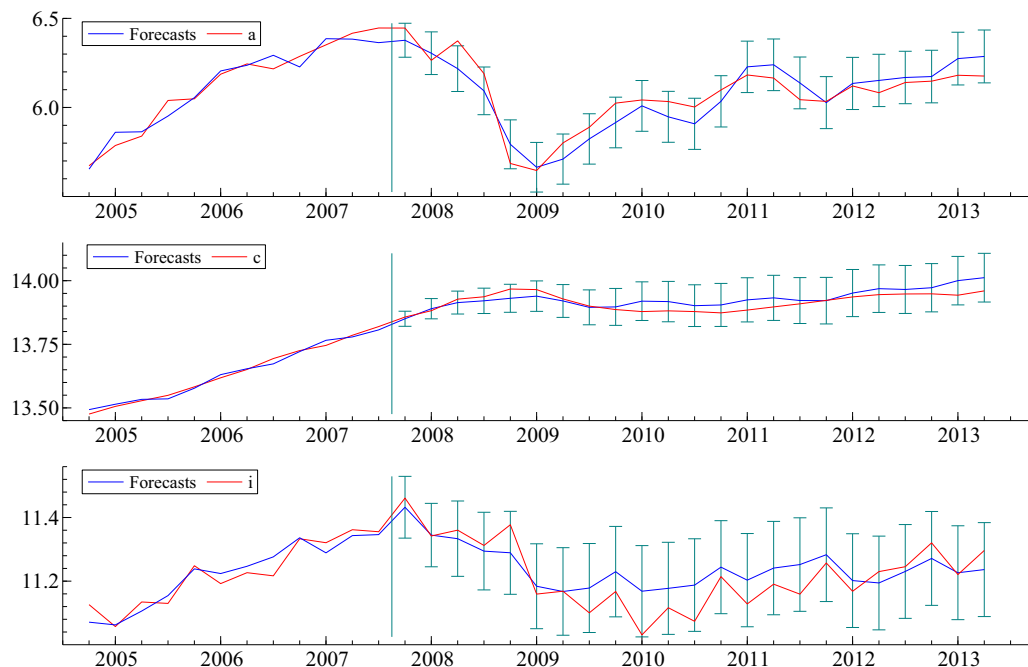
²⁸ Noteworthy, in the wake of sequential shocks to borrower net worth, the credit-asset price spiral reinforced financial accelerator of our SVECM would for some time contribute to bring the economy further and further away from its equilibrium path. However, as the process goes on, eventually the economy would reach a crossroads where the equilibrium correcting forces start to dominate the forces that until then have contributed to drag the economy still further away from its equilibrium path. From then on the disequilibrium position will start to unwind, well supported by a financial accelerator put in reverse. As suggested by the impulse responses, this unwinding of former excesses will not necessarily happen through a smooth reversal to the long run equilibrium path of the economy, but go through an interim period where the economy undershoots its long run equilibrium path before converging towards it in the long run.

²⁹ Furthermore, looking at the error correction coefficients we do see that the "output gap" does not play a structural role in the asset price relation. The positive though insignificant output gap coefficient of the reduced form asset price equation in Table (2) is therefore confirmed to be due to a contemporaneous causal link between credit and asset prices and not a separate structural effect originating from the asset price relation per se. Otherwise, according to (4) higher oil prices affects credit negatively in the short run, only mitigated partially by its positive effect on asset prices. Such an effect of higher oil prices on credit is interpreted to represent a cost effect. In the long-run, however, the effect of higher oil prices on credit comes exclusively via its effect on asset prices and is strongly positive. In fact a one percent rise in oil prices is estimated to increase credit in the long run by just below 0.5 per cent.

$$\begin{pmatrix} 1 & 0 & -0.17 \\ (0.05) & & \\ -0.12 & 1 & -0.037 \\ (0.037) & & (0.02) \\ -0.44 & -1.32 & 1 \\ (0.20) & (0.34) & \end{pmatrix} \begin{pmatrix} \Delta i_t \\ \Delta c_t \\ \Delta s_t \end{pmatrix} = \begin{pmatrix} -0.55 & 0.91 & 0.17 \\ (0.08) & (0.21) & (0.05) \\ 0 & -0.30 & 0 \\ (0.09) & & \\ 0.36 & 0 & -0.18 \\ (0.13) & & 0.08 \end{pmatrix} \begin{pmatrix} \Delta i_{t-1} \\ \Delta c_{t-2} \\ \Delta s_{t-2} \end{pmatrix} \\
+ \begin{pmatrix} -0.21 & 0 & 0 \\ (0.04) & & \\ 0.12 & -0.10 & 0 \\ (0.019) & (0.0098) & \\ 0 & 0 & -0.28 \\ (0.036) & & \end{pmatrix} \begin{pmatrix} i_t + 0.043 RR_t - 0.34 \{s - rc\}_t \\ c_t - 0.7 s_t \\ s_t + 0.056 RR_t - 0.33 msci_t - 0.67 poil_t \end{pmatrix}_{t-1} \\
+ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -0.036 & 0 & 0 & 0 & 0 & 0 \\ (0.015) & & & & & \\ 0.34 & -0.10 & 0.34 & 0.36 & 0.15 & -0.026 \\ (0.05) & (0.04) & (0.06) & (0.06) & (0.06) & (0.007) \end{pmatrix} \begin{pmatrix} \Delta poil_t \\ \Delta poil_{t-1} \\ \Delta msci_t \\ \Delta msci_{t-1} \\ \Delta msci_{t-2} \\ \Delta RR_t \end{pmatrix} \\
+ \begin{pmatrix} 0 & 0.14 & 0 & 0 \\ (0.04) & & & \\ 0.035 & 0 & 0 & 0 \\ (0.014) & & & \\ 0 & 0 & -0.11 & -0.15 \\ (0.031) & (0.05) & & \end{pmatrix} \begin{pmatrix} D2000Q3 \\ D2002Q2 \\ ID2008Q1 \\ D2008Q4 \end{pmatrix}
\tag{4}$$

System diagnostics and test of restrictions

LR-test for over-identifying restrictions:	$\chi^2(52)$	=	49.723[0.5639]
Vector test for autocorrelation of order 1-5:	F(45, 184)	=	1.1040[0.3187]
Vector test for normality:	$\chi^2(6)$	=	3.9178[0.6878]
Vector test for heteroscedasticity:	F(198, 203)	=	0.80135[0.9625]



¹⁾ The forecasts are represented by the red line while the blue lines represent the actual values.

Figure 4: "Ex ante forecasts" of Investments Mainland Norway, i , real domestic credit to enterprises, c , and real equity prices, a , 23 quarters ahead from 2007Q4. Logarithmic scale. 2007Q4 to 2013Q2

5 Conclusion

This paper addresses how to enhance the role of data in structural modeling by proposing a procedure of Simultaneous Structural Model Design. In this procedure particular emphasis is placed on handling the inherent problem of a potential simultaneity bias in design by resorting to an approach where all the behavioral equations of a system are reduced and designed jointly. A central ingredient of such a procedure is the use of extrinsic information and admissible long-run feedback structures as auxiliary tools of exact identification. This renders possible a design process where data are allowed to speak, i.e. a process where both the ordering of the variables and the contemporaneous structure of the model is the outcome of a testable dialog with the data and not the imposition of non-testable restrictions.

In general, the outcome of such a process of Simultaneous Structural Model Design will involve an element of arbitrariness in that it depends on how the structural model was exactly identified in the first place. To add to the reliability of the final outcome it is therefore imperative to give credence to the identification scheme being used. In

so doing, not only to the restrictions being imposed, but also to the extent that the auxiliary tools being used to exactly identify the system makes sense in the sense of having a potential structural interpretation. To legitimate this being the case I have in this paper had to resort to some ad hoc reasoning, a fact that clearly illustrates that there is no such thing as a free lunch when it comes to exact identification. Whether one combines the imposition of a diagonal covariance matrix with SVAR-like restrictions on the contemporaneous feedback matrix or utilizes extrinsic information in the form of structural breaks and assumptions related to the long-run feedback structure, one will never be able to fully free oneself from the curse of arbitrariness. However, to ignore using identification promoting extrinsic information when it exists, is clearly not optimal in this respect, as it would represent a huge disservice to the aim of constructing models informed by data. In particular, such kind of information would enable us to avoid laying the exact identifying restrictions on information laden parts of the model, and to leave such kind of restrictions at the discretion of the data. Of no less importance, by rendering possible a series of tests aimed at revealing the robustness related to the outcome of a structural search process based on an arbitrary and specific set of exactly identifying restrictions, it would also make it possible to explicitly address the issue of arbitrariness.

To illustrate the procedure and to study the simultaneous interplay between financial variables and the real side of the economy, a simultaneous equation model is constructed on Norwegian aggregate quarterly data. As far as the results are concerned the model substantiates the leading indicator properties of the financial variables through the identification of a financial accelerator that is amplified by a credit-asset price spiral in the short run. In the long run, however, the model is driven by a simultaneous structure where asset prices are driven by a set of model exogenous variable at the same time as credit is related to asset prices through a one-directional causal link and investments driven by a Tobins Q effect. The fact that credit is related only to asset prices in the long run means that there is no long-run link between credit and investment. Thus, while temporary innovations to stock prices and credit cause short-run movements in investments - and vice versa - credit does not independently impinge upon the real trajectory of investments in the long run. A lasting asset price shock on the other hand will have long-term real consequences due to the Tobins Q effect. As mentioned in the introduction to this paper this corroborates the results in Beaudry and Portier (2005, 2006) - where they demonstrate that shocks to stock prices may have a lasting long-run effect on the US and Japanese real economy. The results also contributes to reconcile the two opposing views of the literature in the sense that the short run outcome of the model is characterized by a financial accelerator while the financial structure, as represented by credit, is irrelevant for the model's real trajectory in the long run.

References

- Bårdsen, G and R Nymoen (2009), *Macroeconometric modeling for policy*, Vol. 2, Palgrave-Macmillan, pp. 851–916.
- Beaudry, P. and F. Portier (2005), ‘The ‘news’ view of economic fluctuations: Evidence from aggregate Japanese data and sectoral U.S. data’, *Journal of the Japanese and International Economies* **19**(4), 635–652.

- Beaudry, P. and F. Portier (2006), ‘Stock prices, news and economic fluctuations’, *American Economic Review* **96**(4), 1293–1307.
- Bernanke, B. (1986), ‘Alternative explanations of the money-income correlation’, *Carnegie-Rochester Conference Series on Public Policy* **25**, 49–99.
- Bernanke, B. and M. Gertler (1989), ‘Agency costs, net worth, and business fluctuations’, *American Economic Review* **79**(1), 14–31.
- Bernanke, B., M. Gertler and S. Gilchrist (1999), The financial accelerator in a quantitative business cycle framework, in J. Taylor and M. Woodford, eds, ‘*Handbook of Macroeconomics*’, Vol. 1, Elsevier Science B.V., pp. 1341–1393.
- Blanchard, O.J. and D. Quah (1989), ‘The dynamic effects of aggregate demand and supply disturbances’, *American Economic Review* **79**, 655–673.
- Blanchard, O.J. and M.W. Watson (1986), *Are Business Cycles all alike?*, Chicago, University of Chicago Press, pp. 123–156.
- Christensen, I and A. Dib (2008), ‘The financial accelerator in an estimated new keynesian model’, *Review of Economic Dynamics* **1**, 155178.
- Doornik, J. A. and D.F. Hendry (2001), *PcGive 10: Empirical Econometric Modelling using PcGive*, London, Timberlake Consultants Press.
- Ericsson, N.R. and H.A. Tran (1990), ‘Pc-give and David Hendry’s econometric methodology’, *Revista de Econometria* **10**, 7–117.
- Gali, J. (1992), ‘How well does the IS-LM model fit postwar U.S. data?’, *Quarterly Journal of Economics* **107**(2), 709–738.
- Hammersland, R. (2008), ‘Classical identification: A viable road for data to inform structural modelling?’. Statistics Norway, Discussion Papers. No. 562.
- Hammersland, R and C Bolstad Træe (2014), ‘The financial accelerator and the real economy: A small macroeconometric model for norway with financial frictions’, *Economic Modelling* **36**, 517–537.
- Hammersland, R. and D.H. Jacobsen (2008), ‘The financial accelerator: Evidence using a procedure of structural model design’. Discussion Paper No. 569, Statistics Norway.
- Hartley, J, K. Hoover and K. D. Salyer (1998), *Real business cycles: a reader*, Routledge.
- Hendry, D.F. (1993), *Econometrics: Alchemy or Science? Essays in Econometric Methodology*, Oxford, Blackwell Publishers.
- Hendry, D.F. (1995), *Dynamic Econometrics*, Oxford, Oxford University Press.
- Hubbard, R.G. (1998), ‘Capital market imperfections and investment’, *Journal of Economic Literature* **36**, 193–225.

- Johansen, S. (2006), Confronting the economic model with the data, in D.Colander, ed., ‘*Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic General Equilibrium Model*’, Cambridge University Press, pp. 287–300.
- Kiyotaki, N. and J. Moore (1997), ‘Credit cycles’, *Journal of Political Economy* **105**(2), 211–248.
- Kydland, F. E. and E. C. Prescott (1982), ‘Time to build and aggregate fluctuations’, *Econometrica* **50**(6), 1345–1369.
- Modigliani, F. and M. Miller (1958), ‘The cost of capital, corporation finance, and the theory of investment’, *American Economic Review* **48**, 261–297.
- Paustian, M. (2007), ‘Assessing sign restrictions.’, *The B.E. Journal of Macroeconomics (Topics)* **7**(1), 1–37.
- Romer, Christina D. and David H. Romer (2004), ‘A new measure of monetary shocks: Derivation and implications’, *American Economic Review* **94**(4), 1055–1084.
- Romer, Christina D. and David H. Romer (2010), ‘The macroeconomic effects of tax changes: Estimates based on a new measure of fiscal shocks’, *American Economic Review* **100**(3), 763–801.
- Shapiro, M.D. and M.W. Watson (1988), *Sources of Business Cycle Fluctuations*, Cambridge, MA: MIT Press.
- Silvestrini, A. and A. Zaghini (2015), ‘Financial shocks and the real economy in a nonlinear world: From theory to estimation’, *Journal of Policy Modeling* **37**(6), 915–929.
- Sims, C. A. (1980), ‘Macroeconomics and reality’, *Econometrica* **48**, 1–48.
- Smets, F. and R. Wouters (2007), ‘Shocks and frictions in us business cycles: A bayesian d sge approach’, *American Economic Review* **97**(3), 586–606.
- Townsend, R. M. (1979), ‘Optimal contracts and competitive markets with costly state verification’, *Journal of Economic Theory* **21**, 265–293.
- Uhlig, H. (2005), ‘What are the effects of monetary policy on output? results from an agnostic identification procedure’, *Journal of Monetary Economics* **52**(2), 381–419.

Logical Comparison Measures in Classification of Data

Kalle Saastamoinen*

Department of Military Technology, National Defence University,
P.O. Box 7, FI-00861 Helsinki, Finland
kalle.saastamoinen@mil.fi <http://www.puolustusvoimat.fi/en/>

Abstract. Traditionally measures used for comparison have been metric based similarities. In this approach we will present logical comparison measures that have been created using t -norms and t -conorms, which are compensated with generalized means.

We will use classification task as our test bench for the suitability of these measures created. We will compare results achieved with these new measures to the ones achieved with pseudo equivalences and show that these new measures tend to give better results.

Keywords: Ionos, Iris, Wine, Similarity, Comparison measure, Logical, Classification

1 Introduction

In this article we will present how previously presented results in [1] can be improved by the use of simple logic based combined comparison measures presented in this paper.

It is a common belief that measures for comparison should hold true for some properties of metric spaces. This belief originates from the blinkered view that the comparison of objects should always have something to do with distance. This has been questioned in many papers [2–6]. In practice, it seems that properties of distance have little or no affect at all on the results that can be achieved from the use of different comparison measures. This becomes empirically clear when one looks at the test results presented in this paper.

Much of the fuzzy set theory's original inspiration and further developments originate from the problems of pattern classification and cluster analysis. Essentially, this is the reason why classification is chosen to be the test bench for many valued logic based comparison measures in this article. In classification, the question is not whether a given object is or is not a member of a class, but the degree to which the object belongs to the class. This means that most classes in real situations are fuzzy in nature [7]. This fuzzy nature of real world

* I feel gratitude to the National Defence University which have given me time to do my research.

classification problems may shed some light on the general problem of decision making [8].

Motivation for this article is to suggest a general definition for comparison measure and show how previous results presented in article [1] achieved with pseudo equivalences can get better.

Article is organized as follows. In the first section, logical comparison measures, combined comparison measure (CCM) (4) and a little theory behind them is presented. The second section presents classification schemata and data sets used for testing. The third section presents results achieved and these results are compared vs. to the previous results achieved. In the fourth section conclusions are done and some future directions are given.

2 Logical Comparison Measures

Definition 1. *A set function g defined on X , where X is a fuzzy set and has the following properties is called a fuzzy measure:*

1. $g(\emptyset) = 0, g(X) = 1$
2. If $A, B \in X$ and $A \subseteq B$ then $g(A) \leq g(B)$
3. If $A_n \in B, A_1 \subseteq A_2 \subseteq \dots \subseteq A_{n-1} \subseteq A_n$ then $\lim_{n \rightarrow \infty} g(A_n) = g\left(\lim_{n \rightarrow \infty} A_n\right)$

A general definition is given below of what in this paper is meant by a comparison measure.

It is suggested here that the comparison measures used in fuzzy sets, where comparison is done feature by feature and then these comparisons are aggregated, could actually be any measures which fulfil the following properties:

- Property 1.*
1. The comparison measure used has a clear logical structure e.g. it is an Archimedean t-norm or t-conorm (like Frank (1), (2)) or S -equivalence [1].
 2. The comparison measure is monotone. This condition ensures that a decrease (or increase) in any values that are to be compared cannot produce an increase (or decrease) in the comparison result.
 3. The comparison measure is associative. This guarantees that the final comparison results are independent of the grouping of the arguments and that one can expand these comparison to more than two arguments.
 4. The comparison measure is continuous. This guarantees that one can safely compute with the values that are to be compare.

The idea behind using logical structures instead of, for example, simple distances lies in the fact that logical structures always have some kind of linguistic content inside them. For example t-norms and t-conorms can be seen as corresponding to the words "and" and "or", equivalence as corresponding to the expression "if and only if". One can see that just by using these logical measures it is possible to give some linguistic meaning to the comparison procedure.

Some criteria for comparison measures are suggested here. The following criteria are almost the same as Lowen gives for aggregation operators [9] and originally they are presented by Bellman-Giertz [10]. It has also been suggested that not all of these criteria are necessary [11]. One can, however, see that the criteria by Bellman R. and Giertz M. also applies well to the comparison measures presented in this article.

- Criterion 1**
1. **Axiomatic strength.** *It is suggested here that the operator is better if the axioms the operator satisfies are less limiting, this is equivalent to Lowen [9]. It is seen that depending on the choice of the logical structure used this will fit well with the definition given in (1).*
 2. **Flexibility.** *Through the flexibility three things are met that are of an empirical fit, adaptability and compensation. Adaptability comes from the fact that all comparison measures created in this article are parameterized. Compensation property follows from the use of a generalized mean to combine the different values. Empirical fit follows then from the three things and these are the use of logical structures, adaptability and compensation. Empirical fit can naturally only finally be proven by empirical testing, as is done in this article.*
 3. **Numerical efficiency.** *Some operators such as min and max are numerically more efficient than, for example, Frank's t-norm and t-conorm. In large problems this will always be problematic to some degree. However, it is gradually becoming less of a problem as computers computing power is constantly increasing.*
 4. **Range of compensation.** *In general, the larger the range of compensation the better the compensatory operator. In some comparison measures presented in this article the range of compensation has been increased by combining t-norms and t-conorms and in all comparison measures a generalized mean has been used.*
 5. **Aggregating behavior of the comparison measure.** *Aggregating behavior can in the comparison measures presented here, be adjusted by the use of proper mean value in the generalized mean. For example, if a parameter value of 0 is used with a generalized mean a geometric mean will be obtained, which is to say that one attains the product of the values and subsequently each value "added" normally decreases the resulting aggregate degrees of membership.*
 6. **Required scale level of membership functions.** *Comparison measures presented in this article have very little restrictions concerning scale levels.*

2.1 T-norms and T-conorms as the Measures for Comparison

In the paper [12] measures have been defined based on the use of the generalized mean, weights, t-norms and t-conorms. Below these results are added to the definition of the combination measure of the t-norm and t-conorm.

Connectives play an important role when trying to model reality by equations. For example, when linguistic interpretations such as "AND" or "OR" are used for connectives in conjunction and disjunction, quite often this does not

require or mean crisp connectives, but that these connectives are only needed to some degree. In such cases connectives called t-norms or t-conorms may be used. The t-norm gives minimum compensation, while the t-conorm gives maximum compensation. This means that t-norms tend to give more value for the low values, while t-conorms give more value for the high values in the interval in which they are used. In practice, neither of these connectives fit the collected data appropriately. There is still a lot of information that is left in between of these two connectives. An important issue when dealing with t-norms and t-conorms is the question of how to combine them in a meaningful way, since neither of these connectives alone give a general compensation for the values where they are adapted. For this reason one should use a measure that somewhat compensates this gap in between values of these two norms. Article [13] shows how the generalized mean works as the compensative connective between minimum and maximum connectives. The scope of aggregation operators is demonstrated in Figure (1).

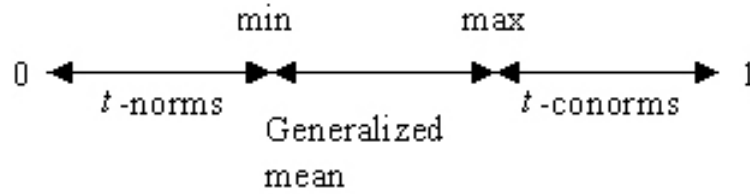


Fig. 1. Compensation of t-norms and t-conorms

The first researchers to try the compensation of t-norms and t-conorms were Zimmermann and Zysno in [14]. They used the weighted geometric mean in order to compensate the gap between fuzzy intersections and unions. When one uses the geometric mean equal compensation is allocated to the all values, and problems might occur if some of the values combined are relatively very low or high.

Created Comparison Measures From T-norms and T-conorms The following is a brief representation of the algebraic equations that can be created by combining weights into some important t-norms and t-conorms and then the combining values are given that were achieved by aggregating them with a generalized mean. Archimedean t-norms and t-conorms are a good choice since they are continuous and monotonic [15].

The comparison measure (4) has been tested by combining it with the following comparison measures (1) and (2). The measure (4) has been tested without weights ω_{ci} and ω_{di} , since the weighting process was too time consuming with differential evolution. All the comparison measures mentioned in this sub-chapter

have been tested in classification tasks. T-norms and t-conorms are tested with weights, where a generalized mean has been used to aggregate and compensate the values.

Parameterized families of t-norms and t-conorms are used here. Families tested in classification were the Dombi family [16], Frank family [17], Schweizer-Sklar family [18], Yager family [19] and Yu family [20]. The Frank and Schweizer-Sklar families of t-norms are also copula families [21] so they have some good statistical properties see Fisher 1997 [22].

From the Frank family we have created the following comparison measures.

Definition 2. *Measure based on Frank (1979) [17] class of t-norm with generalized mean and weights:*

$$T_F \langle f_1(i), f_2(i) \rangle = \left(\sum_{i=1}^n \omega_{ci} \left(\log_p \left[1 + \frac{(p^{f_1(i)} - 1)(p^{f_2(i)} - 1)}{p - 1} \right] \right)^m \right)^{\frac{1}{m}}, \quad (1)$$

where $p > 0$, $p \neq 1$ and $i = 1, \dots, n$.

Definition 3. *Measure based on Frank (1979) [17] class of t-conorm with generalized mean and weights:*

$$S_F \langle f_1(i), f_2(i) \rangle = \left(\sum_{i=1}^n \omega_{di} \left(1 - \log_p \left[1 + \frac{(p^{1-f_1(i)} - 1)(p^{1-f_2(i)} - 1)}{p - 1} \right] \right)^m \right)^{\frac{1}{m}}, \quad (2)$$

where $p > 0$, $p \neq 1$ and $i = 1, \dots, n$.

Definition 4. *Combined comparison measure (CCM) based on the t-norm and t-conorm with a generalized mean and weights [30]:*

$$C \langle f_1, f_2 \rangle = \left(\sum_{i=1}^n \left(w_i T_i^p \langle f_1(i), f_2(i) \rangle + (1 - w_i) (S_i^p \langle f_1(i), f_2(i) \rangle) \right)^m \right)^{\frac{1}{m}} \quad (3)$$

where $i = 1, \dots, n$, p is a parameter combined to the corresponding class of fuzzy intersections T_i and unions S_i and w_i are weights and $i = 1, \dots, n$.

3 Classification

Many time there are given a set of data which is already grouped into classes and the problem is then to predict which class each new data belongs to. This is normally referred to as classification problem. First set of data is referred to as training set, while this new set of data is referred to as test set [23]. Classification is seen as comparison between training set and test set.

3.1 Description of the Similarity Based Classifiers

Objects, each characterized by one feature vector in $[0, 1]^n$, is classified into different classes. The assumption that the vectors belong to $[0, 1]^n$ is not restrictive since the appropriate shift and normalization can be done for any space $[a, b]^n$. The comparison measures can be used to compare objects to classes. Below is the used classifier in the algorithmic form:

SIMILARITY BASED CLASSIFIER

Require: *data*

scale *data* between $[0, 1]$

Require: *test, learn*[1...*n*], *weights, dim*

for *i* = 1 to *n* **do**

idealvec[*i*] = *IDEAL*[*learn*[*i*]]

$$maxcomp[i] = \left(\frac{1}{dim} \right)^{1/m} \left(\sum_{j=1}^{dim} weights[j] (CCM(idealvec[i, j], test[j]))^m \right)^{1/m}$$

end for

class = $\arg \max_i maxcomp[i]$

In the algorithm, the combined comparison measure (CCM) with a generalized mean is used. *IDEAL* is the vector that best characterizes the class *i* and here the generalized mean vector of the class as an *IDEAL*-operator has been used.

When we choose to use randomized weights (RW) instead on using differential evolution (DE), we achieve a significant saving in computing time. RW is approximately 150000-times faster.

Evolutionary algorithm is used because of its diversity and robustness to find weights in classification process, information about evolutionary algorithms in general can be found for example from [24], [25], [26] and [27]. Obviously, other optimizers can be used as well. Evolutionary algorithm used here is based on differential evolution [28]. DE is a simple population based stochastic function minimizer. The objective of DE is to iterate each member of the population and compare its value to the trial member value, and the superior member stays for the next iteration. The evolution strategy defines the way in which a trial member is generated. DE tries to seek weights that will give the maximal similarity compared to the values set by experts. This is done so that DE tries to minimize the value of the objective function with trial member values. The objective function is the total difference between classification defined by experts and the classification defined by similarity used here for all learning data sets. Finally, DE gives the optimal weight values. The basic action of used differential evolution is demonstrated in figure (2).

The classification task has been described more clearly in the flowchart (3). Here classification procedure uses part of the data (*learning*) for weight optimization either using differential evolution or randomized weights depending of the choice. After this rest of the data (*test*) is used for classification and then

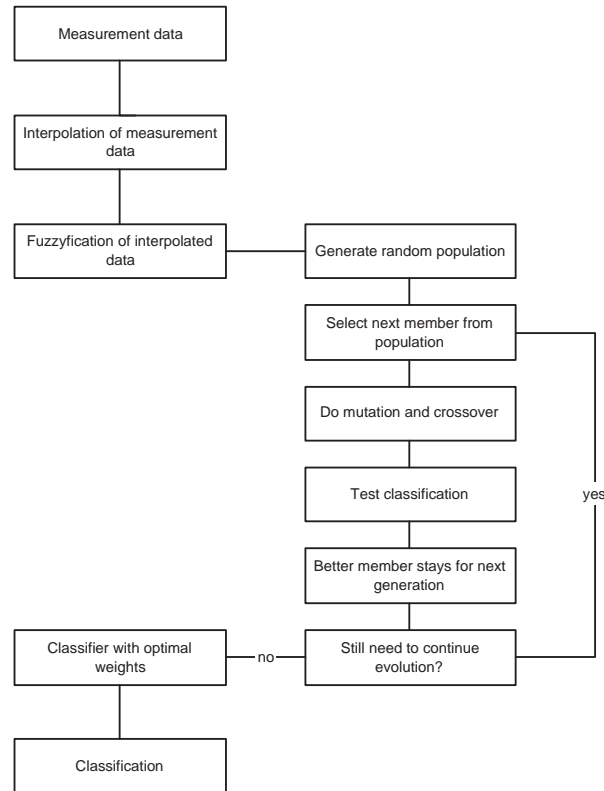


Fig. 2. Simplified computational model for DE

this result is saved, now if loop is done N -times max , min and $mean$ values are saved and then this same classification procedure is done from the beginning for the next parameter value p . After all p -values have been done we start from the next mean value the loop again.

3.2 Data sets

We tested our measures with three different data sets which are available from the [29]. The data sets chosen for the test were: Ionosphere, Iris and Wine. These sets differ greatly in the magnitude of instances and the number of predictive attribute values.

Ionos: This is radar data, where the targets were free electrons in the ionosphere. Here are two classes: "Good" and "Bad". "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. The number of instances is 351. The number of attributes is 34 plus the class attribute.

Iris: Perhaps the best-known database to be found in the pattern recognition literature. The number of attributes is 4 and the class. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.

Wine: The data is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The number of instances: class 1 59, class 2 71, class 3 48.

4 Results

In our classification tasks we have tested five different types of combined comparison measures (CCM), with t -norms and t -conorms from Dombi, Frank, Schweizer-Sklar, Yu and Yager families [30]. In all these tested families combination of Frank norms always managed to give the best results. From the table (1) one can see comparison between the previous best true positive classification results with many-valued pseudo equivalences [1] vs. to the best true positive classification results with CCM presented here. We tested our classifications in both with weights that were randomly selected 200 times (RND) and with weights that were optimized 10 times (DE) for each p - and m -value.

Table 1. Mean (Av), Maximum (Max) Classification Results with different comparison measures (CM) Optimized (DE) and Randomized (RND) Weights and Variances (VAR) vs. previous best results with pseudo equivalences

CM	Iono _{Av}	Iono _{Max}	Iris _{Av}	Iris _{Max}	Wine _{Av}	Wine _{Max}
Frank CCM _{DE}	83.68%	94.89%	91.68%	100%	90.47%	100%
Frank CCM _{RND}	79.86%	94.32%	71.61%	100%	87.22%	100%
Frank CCM _{VAR}	0	0.0075833	0	0.022331	0.00013855	0.022344
Equivalences	80.16%	93.75%	98.84%	100%	96.35%	100%

Table 1 shows the mean and maximum of the classification results from all combinations of weights and parameters. Results were better with Iono-sphere data set than when we used pseudo equivalences for the classification. All the other results were also comparable. Variances with this new measure were also relatively low indicating that combined comparison measure (4) with Frank norms (1, 2) is also quite stable.

5 Conclusions

In classification and the development of expert systems, the problem of choosing the right functions for comparison is often faced. When data has different dependencies, different operators should be used. Usually the simplest operators are selected, which are not normally the optimal choice. As a solution to this

problem this paper has offered combined comparison measure 4 that combines with generalized mean t -norm and t -conorm. This measure is on a logically sound basis. It has been shown that these comparison measures give reasonable results.

It has been shown that the comparison measures introduced in this article consistently give good and stable results in classification, which can be seen from the following table (1). Combined comparison measure (4) based on Frank type of t -norm (1) and t -conorm (2) gave the best classification results, which are the same or better than those attained from the pseudo equivalences. One can also see that the improvements in classification results due to changing to the right comparison measures were quite significant.

From the tested combined comparison measures (4) use of a combination of Frank type t -norm and t -conorm is recommended.

These new comparison measures can be used in, for example, pattern recognition, clustering, expert systems, medical diagnosis systems, decision support systems, fuzzy control etc. Classification results were not only good but also stable, which makes these comparison measures usable.

References

1. Saastamoinen K., Classification of data with Similarity Classifier, International work-conference on Time Series, keynote speech, ITISE2016 Proceedings.
2. Tversky, A., Krantz, D. H.: The Dimensional Representation and the Metric Structure of Similarity Data. *Journal of Mathematical Psychology*. 7(3), pp. 572-596 (1970).
3. Santini, S., Jain, R.: Similarity Measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 21(9), pp. 871-883 (1999).
4. France, R. K.: Weights and measures: an axiomatic model for similarity computations. Internal Report. Virginia Tech, (1994).
5. De Cock, M., Kerre, E.: Why Fuzzy T-Equivalence Relations do Not Resolve the Poincar Paradox, and Related Issues. *Fuzzy Sets and Systems*. 133(2), pp. 181-192, (2003).
6. Tversky A., Features of similarity, *Psychological Review*, 84(4), pp. 327-352, 1977.
7. Zadeh L.A., *Fuzzy Sets and Their Application to Pattern Classification And Clustering Analysis*, in J. Van Ryzin (Ed.): Classification and Clustering, Academic Press, pp. 251-299, 1977.
8. Pal S.K. and D.K. Dutta-Majumder, *Fuzzy Mathematical Approach to Pattern Recognition* John Wiley & Sons (Halsted), N. Y. 1986.
9. Lowen R.R., *Fuzzy Set Theory: Basic Concepts, Techniques, and Bibliography*, Kluwer Acad. Publishers, Dordrecht, 1996.
10. Bellman R. and Giertz M., On the Analytic Formalism of the Theory of Fuzzy Sets, *Information Sciences*, 5, pp. 149-165, 1973.
11. Kang T. and Chen G., Modifications of Bellman-Giertz's theorem, *Fuzzy Sets and Systems*, Vol. 94, Issue 3 (16), pp. 349-353, 1998.
12. Saastamoinen, K., Ketola, J. Using Generalized Combination Measure from Dombi and Yager type of T-norms and T-conorms in Classification, *Proceedings of the ECTI-CON 2005 conference*, 2005.
13. Dyckhoff H. and Pedrycz W., Generalized Means as Model of Compensative Connectives, *Fuzzy Sets and Systems*, 14, pp. 143-154, 1984.
14. Zimmermann H.-J. and Zysno P., Latent connectives in human decision making, *Fuzzy Sets and Systems*, 4, pp. 37-51, 1980.
15. Bilgiç T. and Türkşen I.B., Measurement-theoretic Justification of Connectives in Fuzzy Set Theory, *Fuzzy Sets and Systems*, 76 (3), pp. 289-308, 1995.
16. Dombi J., Basic Concepts for a theory of evaluation: The aggregative operator, *European Journal of Operational Research*, 10, pp. 282-293, 1982.
17. Frank M.J., On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$, *Aequationes Math.*, (19), pp. 194-226, 1979.
18. Schweizer B. and Sklar A., Associative functions and abstract semigroups, *Publ. Math. Debrecen*, 10, pp. 69-81, 1963.
19. Yager R.R., On a General Class of Fuzzy Connectives, *Fuzzy Sets and Systems* 4, pp. 235-242, 1980.
20. Yu Y., Triangular norms and TNF-sigma algebras, *Fuzzy Sets and Systems*, 16, pp. 251-264 1985.
21. Sklar A., Fonctions de répartition á n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, 8, pp. 229-231, 1959.
22. Fisher N.I., Copulas, In Kotz S., Read C.B., and Banks D.L. eds., Update Vol. 1, pp. 159-163, *Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York, 1997.

23. Hastie T. and Tibshirani R., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, 2001.
24. Goldberg D.E., Real-coded genetic algorithms, virtual alphabets, and blocking, *Technical Report 9001*, University of Illinois at Urbana- Champaign, 1990.
25. Mantel B., Periaux J. and Sefrioui M., Gradient and genetic optimizers for aerodynamic desing, *ICIAM 95 Conference*, Hamburgh, 1995.
26. Michalewics Z., *Genetic algorithms + data structures = evolution programs Artificial Intelligence*, Springer-Verlag, New York, 1992.
27. Grefenstette J.J., Optimization of control parameters for genetic algorithms, *IEEE Transactions on Systems, Man and Cybernetics*, 16(1), pp. 122-128, 1986.
28. Price K.V., Storn R.M., Lampinen J.A., *Differential Evolution - A Practical Approach to Global Optimization*, Springer, Natural Computing Series, 2005.
29. UC Irvine Machine Learning Repository, <http://archive.ics.uci.edu/ml/>, [Accessed January 15, 2017].
30. Saastamoinen K., *Many Valued Algebraic Structures as Measures of Comparison*, Acta Universitatis Lappeenrantaensis, PhD. Thesis, 2008.

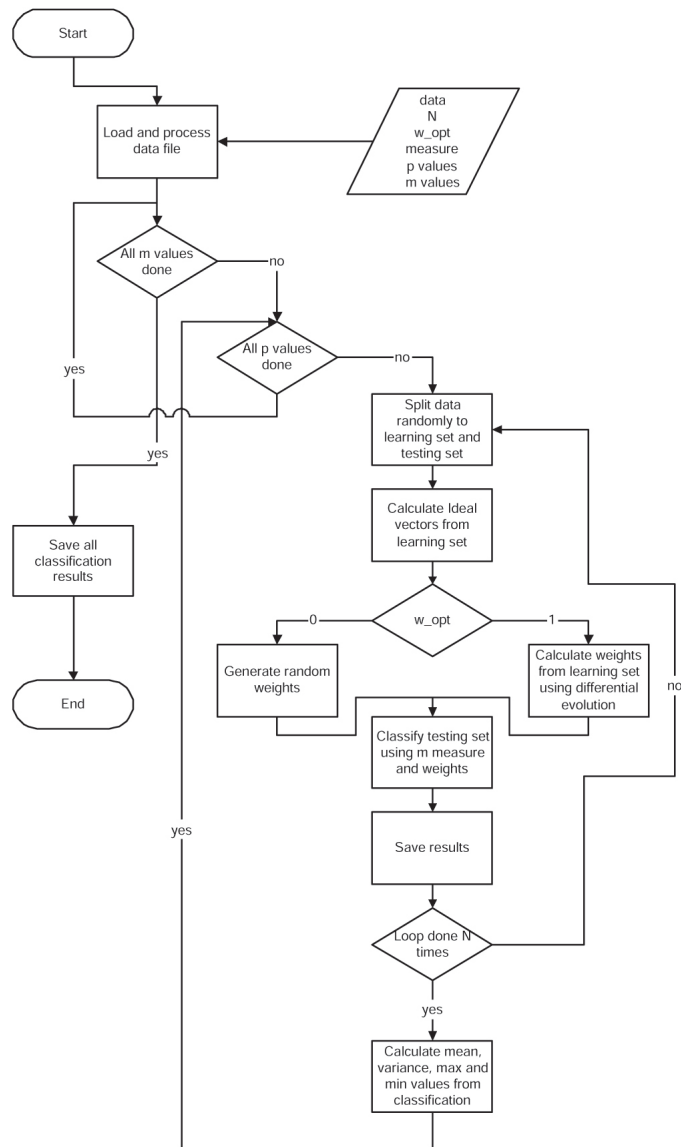


Fig. 3. Simplified Flow Chart of the Classification Procedure

Macroeconomic Forecasting using Approximate Factor Models with Outliers

RAY YEUTIEN CHOU, TSO-JUNG YEN, YU-MIN YEN.

Institute of Economics, Academia Sinica. Address: 128 Academia Road, Section 2, Nankang, Taipei 115, Taiwan.

rhou@econ.sinica.edu.tw

Institute of Statistical Science, Academia Sinica. Address: 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan.

tjyen@stat.sinica.edu.tw

Department of International Business, National Chengchi University, 64, Sec. 2, Zhi-nan Road., Wenshan, Taipei 116, Taiwan.

yyu_min@nccu.edu.tw

Abstract. Approximate factor models assume the mean of the data generating mechanisms is a linear combination of the relevant common factors and error terms. They can extract useful information from a large number of relevant variables. Due to this flexibility, approximate factor models and their extensions are popular in economic analysis and forecasting [10, 3, 7]. [9] list advantages on using approximate factor models. For example, approximate factor models can fit macroeconomic data well. The structure of latent factors embedded in approximate factor model obeys theories in dynamic macroeconomic models, which makes it validate to be used on macroeconomic forecasting and impulse analysis. Some statistical learning methods, such as the lasso, can also handle regression models with a large number of candidate predictors. Using such statistical learning methods relies on the assumption of sparsity in the candidate predictors. However, using approximate factor models does not need to impose the special assumption. This is one of the reasons why approximate factor models are the main tool used for big data in macroeconomics [9]. In terms of forecasting important macroeconomic variables, a completed survey conducted by [6] shows that approximate factor models indeed have a superior performances on forecasting exchange rates and inflation than other methods.

Approximate factor models usually assume common factors are non-observable. As a result of that, an important goal in estimating the approximate factor models is to identify the latent common factors and their factor loadings. Methods for carrying out this task include the maximum likelihood estimation (MLE), Markov Chain Monte Carlo (MCMC) and Principal Component Analysis (PCA). Econometricians often estimate the approximate factor models with high dimensional data, and in this circumstance the PCA method is more preferred as it is less computational intensive than the MLE and MCMC.

By using the PCA method on estimating approximate factor models, the latent common factors can be viewed as a linear combination the original predictors. When the number of original predictors becomes large, this

kind of cross sectional linearly combination can reduce estimation errors in the estimated latent common factors. More accurate estimated latent common factors makes approximate factor models an efficient method on empirical applications.

Although the PCA method has the aforementioned advantages, it may fail to provide a correct estimation on the latent factors and factor loadings when data are subject to large and sharp spikes [5]. To overcome this difficulty, we present a simple and efficient method for estimating approximate factor models with data containing such large and sharp spikes. We call the proposed estimation method “P-PCA method” (*Penalized least squares plus PCA method*). This method formulates the estimation problem as a penalized least squares problem in which a norm penalty function is imposed on those large and sharp spikes. Such a formulation allows us to model both the latent common factors and the sharp and large spikes, and therefore can reduce estimation biases in the latent common factors and their factor loadings. To solve the estimation problem, we decompose it into two optimization problems: a Principal Component Analysis (PCA) problem and a one dimensional shrinkage estimation problem. We then develop an algorithm to iteratively solve the two problems. This algorithm can deliver reliable numerical estimates and is flexible in incorporating methods for selecting the number of common components.

There exist different approaches to formulating and estimating the approximate factor models. [8] proposed a multilevel factor model for large panel data with between-block variations and idiosyncratic noise. They developed an estimation procedure which can identify block-level shocks and genuinely common factors, and therefore achieve dimension reduction. [1] proposed a multifactor model with a large number of observable factors and unobservable common and group-specific pervasive factors. Their estimation procedure for such a model can simultaneously select relevant observable factors and determine the number of common and group-specific unobservable factors. [4] developed a factor model in which both factor loadings and number of factors can have a structure break. They adopted a shrinkage estimation that can consistently identify the number of common factors before and after the structure break. Their estimation procedure can be implemented by solving a convex optimization with the principal components of data matrix as input.

As compared with the aforementioned research, our method focuses on a situation in which data are subject to large and sharp spikes. For example, observations may occasionally be blurred by extreme large signals, like asset price jumps in financial data. It is different from the situations in which data generating mechanisms are broken by a permanent change of common factors or factor loadings. Indeed, under suitable assumptions on the idiosyncratic uncommon components [2, 10], the factors and factor loadings might still be consistently estimated by using the PCA method. However, in term of finite sample efficiency, we show that the proposed P-PCA method can outperform the conventional PCA method on estimating the latent common factors. We demonstrate such advantages by carrying out intensive simulations under a wide range of model settings. We then use the proposed method performs on predicting yearly

growth of important macroeconomic variables and find that it can deliver comparable performances as the traditional methods. Throughout these works, we believe the proposed method can serve as a complementary tool for robust estimations rather than a competitive approach to those established approximate factor models.

Keywords: Approximate Factor Model, Forecast, Norm Penalty, Principal Component Analysis

References

1. Ando, T. and J. Bai: Asset Pricing with a General Multifactor Structure, *Journal of Financial Econometrics*, 13, 556–604 (2015).
2. Bai, J. and S. Ng : Determining the Number of Factors in Approximate Factor Models, *Econometrica*, 70, 191–221 (2002).
3. Bernanke, B., J. Boivin, and P. S. Elias (2005): Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach, *The Quarterly Journal of Economics*, 120, 387–422 (2005).
4. Cheng, X., Z. Liao, and F. Schorfheide: Shrinkage Estimation of High-Dimensional Factor Models with Structural Instabilities, *The Review of Economic Studies*, 83, 1511–1543 (2016).
5. Jolliffe, I. T.: *Principal Component Analysis*, Springer, second ed. (2002).
6. Kavtaradze, L. and M. Mokhtari: Factor Models and Time-Varying Parameter Framework for Forecasting Exchange Rates and Inflation: A Survey, *Journal of Economic Surveys*, forthcoming (2017).
7. Ludvigson, S. C. and S. Ng: Macro Factors in Bond Risk Premia, *Review of Financial Studies*, 22, p. 5027–5067 (2009).
8. Moench, E., S. Ng, and S. Potter: Dynamic Hierarchical Factor Model, *The Review of Economics and Statistics*, 95, 1811–1817 (2013).
9. Stock, J. and M. Watson: Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics, Elsevier, vol. 2 of *Handbook of Macroeconomics*, 415–525 (2016).
10. Stock, J. H. and M. W. Watson: Forecasting using Principal components from a Large Number of Predictors, *Journal of the American Statistical Association*, 97, 1167–1179 (2002).

Testing Granger-causality on macroeconomic time series: a bootstrap approach

Matteo Farné, Angela Montanari

University of Bologna

Abstract. In this short paper, we present a bootstrap test for Granger-causality in the frequency domain, particularly suitable for short macroeconomic series. In particular, we improve upon the testing approach of Ding et al. (2006) proposing a bootstrap test for comparing unconditional and conditional Granger-causality. This allows to test the relevance of a conditioning variable in respect to the causality structure. A relevant application to the study of co-movements of money stock and GDP in the Euro-Area is described, in order to provide some answers about the effectiveness of the ECB monetary policy before and after the financial crisis of 2008.

Macroeconomic time series often have a quarterly frequency, which may result in a very short sample size T . In this short paper, we describe a bootstrap testing approach for Granger-causality (GC) in the frequency domain, specifically thought for short series. We discuss the rationale behind this method and we show a relevant application on the study of the mutual relationship between GDP and money stock (both M3 and M1 aggregate) in the Euro area, before and after the financial crisis of 2008.

Causality measures in the frequency domain were first proposed in Pierce (1979) as R^2 measures for time series. In Geweke (1982) and Geweke (1984) the concept of unconditional and conditional Granger-causality in the frequency domain was introduced. Its measures were extended in Hosoya (1991) and Hosoya (2001) respectively. In Breitung and Candelon (2006), a test for Granger-causality in the frequency domain is proposed. Its very elegant formulation in the VAR and cointegrated VAR context presents two shortcomings for our analysis: the convergence rate is $O(T^{-1/2})$ and the power is decreasing as the distance of the frequency of interest from $\frac{\pi}{2}$ increases. These drawbacks may lead to misleading conclusions in presence of a small sample size.

For this reason, the inference on Granger-causality spectra in the frequency domain is still an open problem. Differently from the time-domain quantities, the limiting distribution for unconditional and conditional spectra is unknown (see Barnett and Seth (2014), par. 2.5). In Ding et al. (2006) bootstrap thresholds are derived for Geweke's unconditional and conditional GC measures in the context of neurological data. A further extension of that approach can be found in Wen et al. (2013), and relevant applications in the neurophysiological context include Brovelli et al. (2004), Roebroek et al. (2005) and Dhamala et al. (2008), where explicit VAR estimation is avoided by a nonparametric approach.

In order to test the significance of Granger causalities at each frequency, a possible approach is to retain the maximum across frequencies as in Ding et al. (2006), and to build the bootstrap distribution generating stationary bootstrap time series as in Politis and Romano (1994), from which it is possible to draw the desired quantiles. We stress that unconditional and conditional spectra must be assessed separately, because their magnitude can in general be different.

In Ding et al. (2006), the comparison between unconditional and conditional Granger causalities is performed using the t-test on the bootstrapped series of the peak across frequencies. This approach can be suitable for their case, where they deal with psychological/neurological data: on the contrary, in the economic context, we can not assume that the two populations are normal and independent. For this reason, we propose to take the signed maximum difference in absolute value for each run. Then we have a bootstrap distribution of the signed difference. We can test the observed difference between unconditional and conditional Granger-causality spectra at each frequency comparing it with the bootstrap quantiles computed over all the runs returning only stationary VAR models. This approach works on both directions, and is suitable both for two-tail and one-tail tests.

The key innovation in our approach is now explained. If the unconditional GC at a particular frequency is significant, we can say that the cause variable causes the effect variable. If the conditional GC at a particular frequency is significant, we can say that the cause variable causes the effect variable once the effect of the conditioning variable has been removed. Nevertheless, we can not say in any case that the conditioning variable has a relevant impact on the causality structure in absence of a specific test. Since unconditional and conditional GC at the same frequency have unknown and dependent distributions (we are in a time-dependent data context) we can not use the comparison test of Ding et al. (2006), which works for serially independent data. Our proposed test fills this gap, allowing to determine if the conditioning variable has a significant impact on the causal relationship (amplification-annihilation) or not.

The described method allows to make statistically founded considerations on the nature of the relationship between economic output and money stock in the Euro Area, even taking into account further series like the unemployment rate (UN) and the inflation rate (HICP). Our analysis focuses on the period 2001-2014, when the monetary union has become effective. Our data come from the ECB Real Time Database, where all figures are harmonized respect to the changing composition of the Euro Area across time (see Giannone et al. (2012) for the details).

We have conducted GC analysis separately for time periods 2001-Summer 2008 (31 quarters), Autumn 2008-2014 (25 quarters) in order to obtain locally mean and covariance stationary series (Box et al., 2015). In fact, pre-crisis and post-crisis samples have extremely different characteristics, so that we must conduct the analysis separately for the two periods.

Unconditional spectra show that one significant relationship is the causal relationship from M3 to GDP in the pre-crisis sample. Its characteristic period

is around 1 year. Another one is the causal relationship from M1 to GDP in the post-crisis sample. In that case, the significance is on a wide period range, from 3 to 8 quarters approximately.

If we remove the effect mediated by HICP, any causal relationship between M3 and GDP disappears, while the causal relationship from M1 to GDP is weakened as it is still present only for a period of 4/5 quarters approximately. We have evidence of causal relationship from GDP to M1 both in the pre-crisis and in the post-crisis sample at very high frequencies (period 2 quarters), thus proving that money stock can also react to shocks in the economic output in the short run. If we remove the effect mediated by UN any significant causality disappears for both settings (M1-GDP and M3-GDP).

We have evidence of strong difference between unconditional and conditional spectra only in two cases, both in the post-crisis sample: the unconditional causality from GDP to M1 is significantly smaller than the same causality conditional on HICP, and the unconditional causality from M1 to GDP is significantly larger than the same causality conditional on UN. We can note that the strength of the effect of GDP to M1 is amplified removing the indirect effect of HICP, while the strength of the effect of M1 to GDP is annihilated removing the indirect effect of UN.

Bibliography

- Barnett, L. and A. K. Seth (2014). The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference. *Journal of neuroscience methods* 223, 50–68.
- Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Breitung, J. and B. Candelon (2006). Testing for short-and long-run causality: A frequency-domain approach. *Journal of Econometrics* 132, 363–378.
- Brovelli, A., M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences of the United States of America* 101, 9849–9854.
- Dhamala, M., G. Rangarajan, and M. Ding (2008). Analyzing information flow in brain networks with nonparametric granger causality. *Neuroimage* 41, 354–362.
- Ding, M., Y. Chen, and S. L. Bressler (2006). 17 granger causality: Basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications*, 437.
- Geweke, J. F. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association* 77, 304–313.
- Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association* 79, 907–915.
- Giannone, D., J. Henry, M. Lalik, and M. Modugno (2012). An area-wide real-time database for the euro area. *Review of Economics and Statistics* 94, 1000–1013.
- Hosoya, Y. (1991). The decomposition and measurement of the interdependency between second-order stationary processes. *Probability theory and related fields* 88, 429–444.
- Hosoya, Y. (2001). Elimination of third-series effect and defining partial measures of causality. *Journal of time series analysis* 22, 537–554.
- Pierce, D. A. (1979). R 2 measures for time series. *Journal of the American Statistical Association* 74, 901–910.
- Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical association* 89, 1303–1313.
- Roebroeck, A., E. Formisano, and R. Goebel (2005). Mapping directed influence over the brain using granger causality and fmri. *Neuroimage* 25, 230–242.
- Wen, X., G. Rangarajan, and M. Ding (2013). Multivariate granger causality: an estimation framework based on factorization of the spectral density matrix. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 371.

An implied rating software system

Ventsislav Nikolov

Abstract. The paper presents a mathematical approach to create credit rating scale and to classify financial institutions by it which is implemented as a model in a software system. The presented model is based on competitive trained neural network working in model building and classification stages. Thus an individual point of view can be provided for any institution according to their available data.

Keywords: Implied rating, Classification, Self-organizing map.

1 INTRODUCTION

The credit ratings evaluating the credit worthiness of different obligators [2] are important data for the business and governments. When there is a financial and economic crisis the importance of the ratings produced by the rating agencies even raises because they influence the investors' decisions and corporate operations toward a given direction. The ratings determine the interest rate for the borrower which leads to different prices of loaning money. Here a method for automatic building of a credit ratings scale based on specific corporate and government data as well as determination of the credit rating of a given borrower is described. The method is implemented as a software module which can be integrated in a variety of software products. The need of new methods for credit rating determination emerges because of the following reasons.

- The reaction of the rating agencies often is too slow and the market is dynamic. Sometimes even a default occurs of a company or institution while their rating is still classified as high one.
- The ratings are determined in long time intervals. Ratings available on daily basis could be a significant advantage.
- Sometimes the rating agencies are criticized of conflict of interests. They analyze the political environment, regulations, the ability to return already borrowed loans, etc. If the ratings are determined based extensively on the statistics they would be more accurate.
- Every market participant could produce its own rating scale to classify the other participants. Thus independent credit rating estimation could use its own data and the distributed overall assessment could increase the quality of the taken decisions.

2 SOLUTION

The stages of the proposed system are shown in fig. 1. The first stage is choosing the grades of the rating scale. Different agencies work with different scales which contain upper and lower alphabetical letters combined with positive or negative signs and sometimes digits. Some rating agencies use different scales for the long and short term rating. In our system only one scale is used. The information for the scale is the first needed information because it determines the number and order of the rating grades which define the model building step.

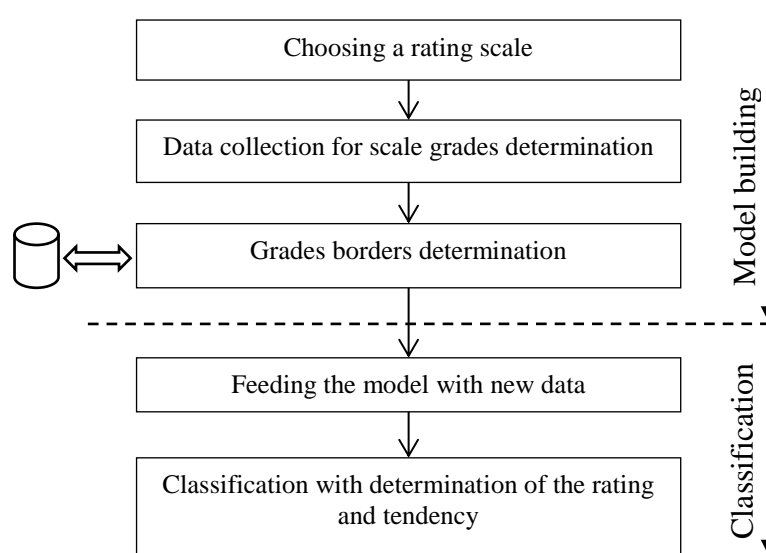


Fig. 1. The stages of implied rating system working

In the next step the available data should be collected in order to determine the centers and boundaries of the grades. In our approach this data comprises of the levels of Credit Default Swap (CDS) spread curves, Credit Default Swap Index (CDX) spread indices [1], share prices and bond prices which themselves contain information reflecting the current worthiness of the participants represented by such data. The working of the automated rating determination system relies on the reverse task of the upper data determination.

Thus by given data the credit rating of a participant should be calculated by mathematical approaches. The spread curves and indices could be considered similar to assurance which means that the higher their values are the higher the risk of the default is for the participant they represent.

The data used for the scale building are in fact a set of time series for a given historical time period. Every time series represent a market participant. The time horizon should be the same for all series. If the values do not coincide by dates then interpola-

tion is performed to make them for equal dates. The more data collected the more accurate rating scale will be build.

The grades centers and bounds determination is performed by a mathematical model composed by a self-organizing features map [7] with one-dimensional output lattice which is described in more details below. Once the scale is built it could be saved in a database for use after that. Periodically the scale must be built and saved again. Thus the whole scale with all grades fluctuates over time according to the state of all participants which series are used in the scale building. This means that if for example a crisis happens then the series of the most participants would move upward because the interest rates increase but this occurs correlated for them and thus most participants could also preserve their rating.

The scale building is based on ranges determination from the given data. In fact the series may overlap one another in some time periods but the grades boundaries are finally precisely specified. The degrees are determined by their centers. Having these centers the boundaries between them are calculated as their averages. When a new series is classified it could crosses some of the boundaries but this should not hinder from finding the nearest degree center to the series. The simple or decayed Euclidean distance is used in this case as it is shown to be d_1 and d_2 in fig. 2 where the new series center is shown in yellow and it is between the grades AA and A.

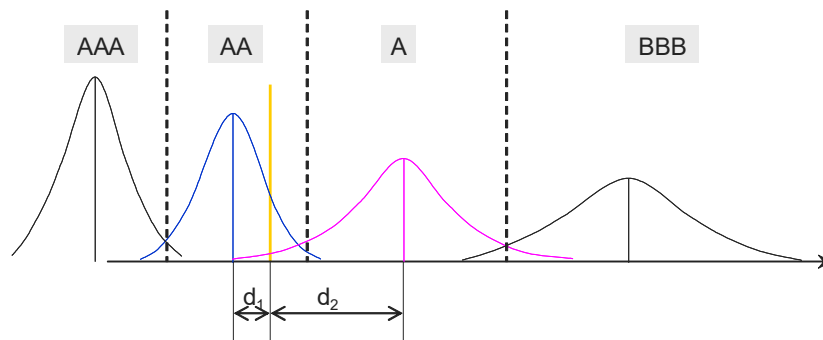


Fig. 2. Classification of a new obligator series by measuring the distance

All data values in every degree are considered as a set and their probability distribution are built supposing normal distribution. To do this first the normal distribution parameters μ (mean) and σ (standard deviation) are estimated for every degree and then the theoretical histogram is graphically shown [3] [4]. The new series together with its degree center and boundaries are shown in fig. 3.

3 THE MATHEMATICAL MODEL

3.1 Model building

The mathematical model used in our approach is based on self-organizing map used for clustering and classification of time series. Its input and output layers are shown in fig. 4. This self-learning mathematical model determines the parameters of its internal

structure based on the input data. The topological ordering of the output nodes is one-dimensional in our case because the grades are ordered one-dimensionally. The spatial shape of an output unit is chosen to be a square. Every output node corresponds to a group which represents a rating scale degree. Each input node corresponds to a single historical time series value. Thus in the learning stage of the self-organizing map the time series are classified into groups based on both their magnitude and historical behaviour. The learning is an iterative process in which all time series are subsequently used as input and the output nodes prototypes change their positions. In the beginning of the learning the changes are greater and with the time it decays non-linearly doing fine correction until the end.

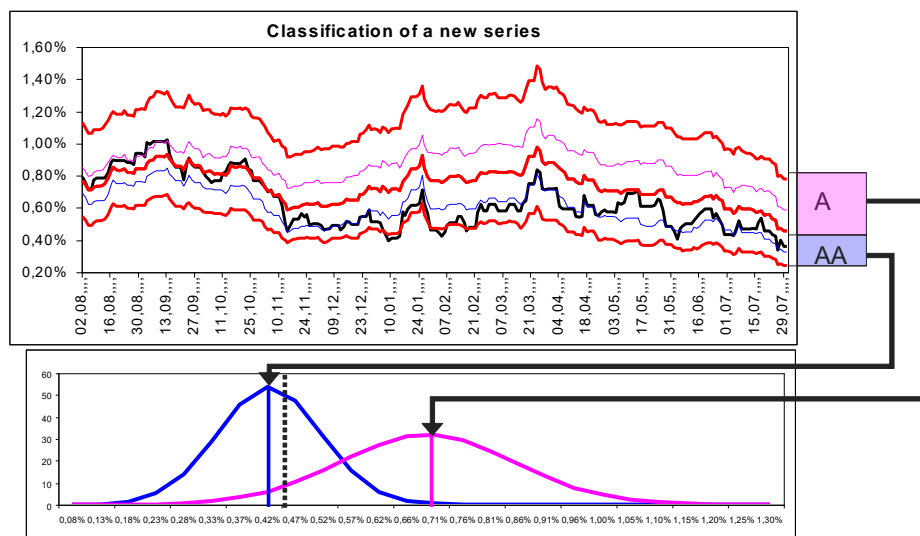


Fig. 3. A new series (in black color) shown together with its degree and the second nearest degree

When the groups and their centers are determined the grades bounds are calculated based on the groups' centers which correspond to the output nodes prototypes. The greater the rating is the smaller the prototype magnitude is. The prototypes are shown in figure 4 in the output layer nodes as small curves in the squares. The prototypes are considered as grades centers so groups' bounds are calculated to be the average of the neighbors' prototypes which can be seen in figure 3 where the center of the group representing the grade A is shown in pink and the center of AA in blue. Thus if the width of two neighbor grades is not the same then the prototype will not be in the center of the grade after its bound determination.

The grades determination is followed by their reordering according to the mean of their prototypes. This is needed because the self-organizing map weights are initialized with random values and the grades must always be sorted in the same way. When the group determination should be repeatable the random generation is set appropriately. Thus the groups are sorted descending after the learning stage.

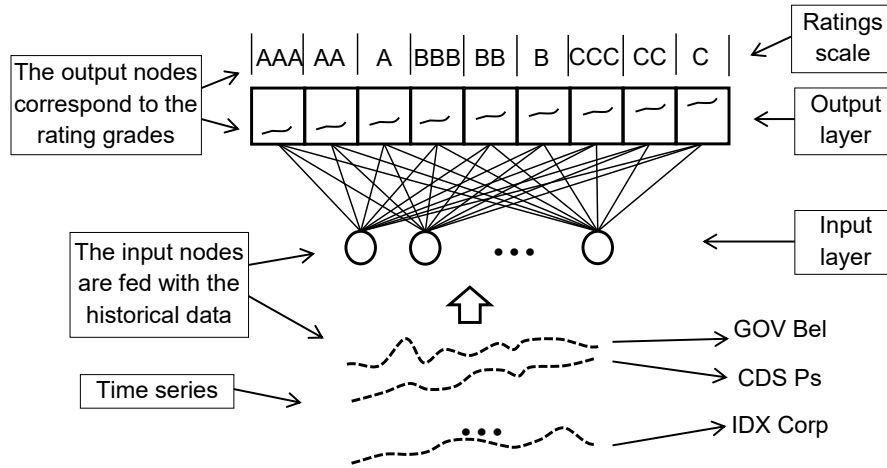


Fig. 4. The self-organizing map based model

3.2 Using the model for classification of a new series

The prototypes of the output nodes are used not only for the determination of the grades but also for classification in the next stage. When the rating scale is determined the next stage is to use it in order to classify a new series. In this way the rating category of a participant is determined. This is performed by comparing the new series with each prototype of output node and finding the nearest one which is chosen to be the rating grade. In the classification stage not only the best matching grade is found but also the second nearest one which is considered as the tendency grade with a given confidence. The confidence is calculated as closeness between the new input series and the prototypes of the rating (nearest) and tendency (second nearest) grades using (1) – (3).

$$s = r + t \quad (1)$$

$$v_r = 100 * \left(1 - \frac{r}{s}\right) \quad (2)$$

$$v_t = 100 * \left(1 - \frac{t}{s}\right) \quad (3)$$

where

r – distance between the time series and the nearest (rating) prototype;

t – distance between the time series and the second nearest (tendency) prototype;

v_r – confidence of the rating;

v_t – confidence of the tendency;

3.3 Using of decay factor

An important fact is that the model is built not only giving an account of the data magnitude but also of their historical behavior. The more recent data however should be considered as more important in both the grades determination and for the classification stage. That is why a decay factor is used when the distance is calculated between a data time series and a grade prototype using (4).

$$d = \sqrt{\frac{1}{\sum_{i=1}^N \lambda^{i-1}} \sum_{i=1}^N \lambda^{N-i} (x_i - y_i)^2} \quad (4)$$

where

λ – a decay factor;

N – time series size;

x – time series;

y – grade prototype;

The decay factor values vary from 0 to 1. When a decay factor is used in the grades determination stage the entire set of time series in the grade tend to be within the grades bounds at their end. In fig. 5 in the left a grade is shown with time series without using a decay factor and in the right the grade is shown using it. It can be seen that at the end of the X axis the series tend to be between the grade bounds only if the decay is used. Otherwise the series scatter also outside of the grade bounds through all their length.

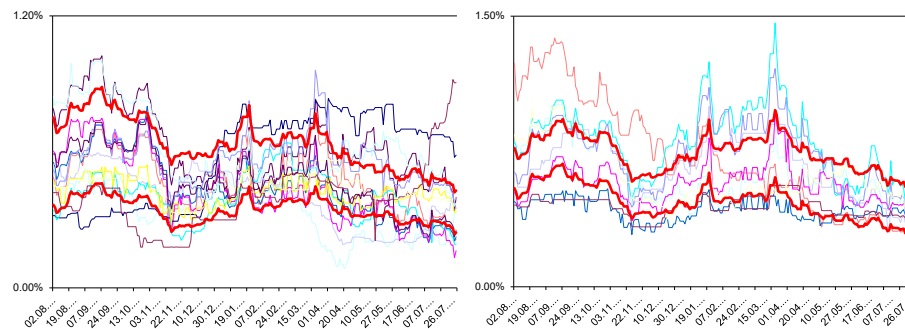


Fig. 5. The effect of the using of the decay factor during the grades determination stage

The decay factor could also be used for the mean calculation of a series. The effect can be seen in fig. 5 where the red dashed line is the mean with decay giving more weight to the last values in the series and with blue dashed line the mean is shown calculated without a decay factor.

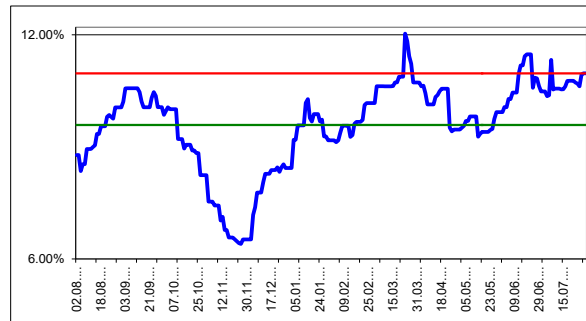


Fig. 6. Calculation of mean of a series with decay factor (dashed red line) and without it (dashed green line)

4 CONCLUSIONS AND FUTURE WORK

The prototype of the software system realizing the described approach is developed in Java. Its computational part is implemented as a JAR library which can be used either as a directly used module in business layer logic of other software systems or as service operations.

Below some examples of classification of new series are shown. The grades are ordered from AAA that is the best credit rating to D that is default. In fig. 7 the new series is classified as CC and despite of the fact that its trend is increasing its tendency is shown to be CCC because some part of the history has been in that grade.

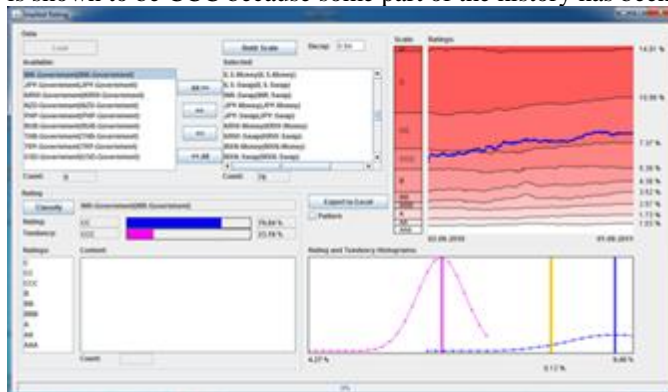


Fig. 7. The implemented prototype

In fig. 8 the series is classified as the best rating AAA with tendency to be AA. There could not be other tendency grade because there is no better credit rating than AAA. That is why in such cases the most important information is to what extend the tendency is to be AA.

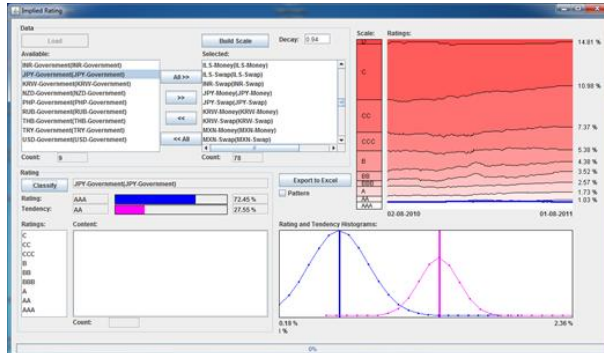


Fig. 8. The implemented prototype

In fig. 9 the series is classified to be BB but is moving too near to the board between BB and BBB. In the history it has been in the neighbor BBB grade.

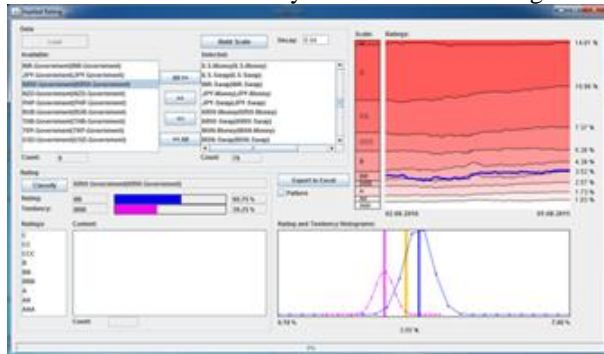


Fig. 9. The implemented prototype

An interesting series is shown in fig. 10 where the series starts from B moves in BB and finishes in BBB. The decay factor here shows the importance of the last values in such cases. Moreover the tendency is shown to be A even though the series has been from the other side of the scale and never in A.

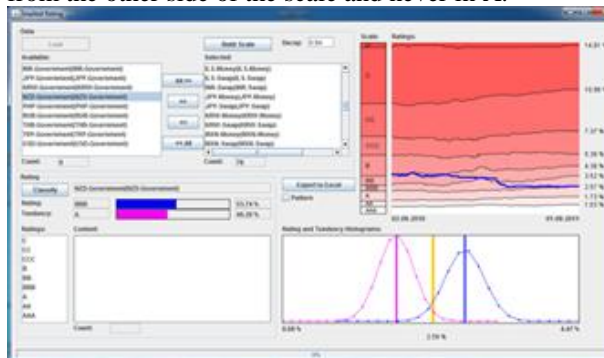


Fig. 10. The implemented prototype

Fig 11 also shows a case when the decay is important for the right classification. The series is determined to be in BBB thought in his historical movement it changes from B to A.

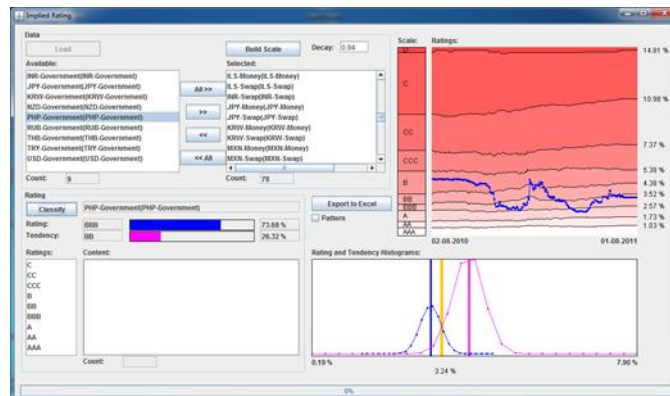


Fig. 11. The implemented prototype

The experiments show that the system is robust especially regarding the ability to classify according to the more actual data. Thus the historical values are taken into account but not so important than the last ones. And such a system could be used on daily basis and with individual settings that are good advantages not only for experimental but also for practical uses.

5 REFERENCES

1. Beinstein, E. et al. Credit Derivatives Handbook Detailing credit default swap products, markets and trading strategies. Corporate Quantitative Research New York, London December, JPMorgan. 2006.
2. Standard & Poor's. Global Credit Portal – Standard and Poor's Rating Definitions. June 22, 2012.
3. Bury, K. Statistical Distributions in engineering. Cambridge University Press, 1999.
4. Krishnamoorthy, K. Handbook of statistical distributions with applications. Chapman & Hall, 2006.
5. Mun, J. Modeling Risk. Applying Monte Carlo Simulation, Real Options Analysis, Forecasting, and Optimization Techniques. Wiley, 2006.
6. Vallentin, M. Probability and Statistics Cookbook, 2011 (<http://matthias.vallentin.net/probability-and-statistics-cookbook/>).
7. vallentin.net/probability-and-statistics-cookbook/.
8. Kohonen, T. Self-organizing maps. Springer, 2001.

6 ABOUT THE AUTHOR

Assist. Prof. Ventsislav Nikolov, PhD, Department of Computer Science and Engineering, Technical University of Varna, Phone: +359 52 383 424, E-mail: v.nikolov@tu-varna.bg.

Estimating Fiscal Policy Rules

Diederik Kumps

Department of Applied Economics
Vrije Universiteit Brussel

Peter Claeys

Department of Applied Economics
Vrije Universiteit Brussel

April 2016

Abstract

The 2008–2009 financial crisis has shaken the beliefs about how macroeconomic policy should be conducted. Central banks in G7 countries shifted to unconventional policy measures in the aftermath of the Financial Crisis, when faced with economic slack, financial instability and fiscal trouble. Governments were swift in saving the banking system, and the economy as a whole, from collapse. This shift ended a spell of rules-based time consistent policy that started in the mid-1980s in many industrialised economies. Changes in policy regimes occur in response to economic or political events. These changes are often modelled by estimating the respective policy rules with Markov Switching (MS) techniques. One important pitfall in these estimations is the endogeneity of explanatory variables. When the explanatory variables are endogenous, MS estimates are inconsistent. We account for this endogeneity with a novel MS test of a fiscal rule. Results show that the stable mix of policies during the Great Moderation has given way to a mix of 'passive' monetary and 'active' fiscal policies since the mid 2000s.

Keywords: fiscal regimes, Markov Switching, endogenous variables, IV

JEL: E62, E65, H11, H62

1 Introduction

The 2008–2009 financial crisis has shaken the beliefs about how macroeconomic policy should be conducted. Governments have been swift in saving the banking system, and the economy as a whole, from collapse. These bail-outs have come at the cost of a substantial increase in government debt. Central banks too have implemented unconventional monetary policies by providing unlimited credit at near zero interest rates. This policy mix of fiscal and monetary stimuli has prevented an economic meltdown. However, a return to stability oriented policies for the long term — characterised by stable government debt and inflation — may require a set of unorthodox policies in the short term. Central bankers may announce changes in policy regime to mold beliefs by agents on future policy actions (Davig & Leeper, 2006). Bianchi (2013) and Bianchi and Melosi (2014) show that governments can also determine the monetary/fiscal policy mix by guiding beliefs through announcements of future actions. In these studies, empirical characterisation of past changes in regimes is based on the properties of the macro-economic series included in the model by estimating policy rules with Markov Switching techniques.

In this study, we start with developing a simple and static fiscal policy rule for the US during the period 1966–2014, that describes how a fiscal variable — the primary deficit — responds to the current state of government debt and the business cycle (Favero & Monacelli, 2005; Davig & Leeper, 2006; Afonso, Claey's, & Sousa, 2010). One important issue that to the best of our knowledge has not been addressed properly, is the possible problem of endogenous explanatory variables in the specification of the fiscal policy rule. We test for endogeneity with a simple Instrumental Variables (IV) based regression, using a common set of instruments suggested in the macroeconomic literature. Not surprisingly, statistical tests indicate the existence of endogenous variables.

This paper is structured as follows. In section 2 we overview the literature on fiscal policy rules, build the baseline model and discuss the methodology of regime switching models where we account for the endogeneity of policy action. Section 3 describes the data. Section 4 discusses results for models with and without endogeneity. A final section concludes.

2 Methodology

2.1 Defining a fiscal rule

A vast amount of macroeconomic research has been dedicated to the specification of so called policy rules. At the base of this literature are the seminal articles of Kydland and Prescott (1977) and Barro and Gordon (1983). Perhaps the most well-known example of these policy rules is the so-called Taylor-rule, named after James Taylor's specification of a monetary policy rule for the U.S. (Taylor, 1993). In this study we focus our analysis on the specification of a fiscal policy rule. This type of policy rule typically relates a fiscal indicator of choice to a multitude of macroeconomic indicators that might influence

the behaviour of fiscal policy makers. Fatás and Mihov (2003) stress the importance of these rules as an effective way to restrict discretionary fiscal policy.

A large part of the literature on fiscal policy rules is dedicated to the assessment of the cyclical behaviour of fiscal policy. Gali and Perotti (2003) decompose fiscal policy in (i) a cyclical or non-discretionary component and (ii) a structural or discretionary component. The first is considered to be out of direct control of the policy makers and often described as the automatic stabilisers in fiscal policy theory (e.g. changes in tax revenues for given tax rates or changes in unemployment benefits). The latter can be viewed as the value of the fiscal indicator if output were at its potential level. A typical rule to test the cyclicity of fiscal policy regresses the fiscal indicator on an indicator of the current state of the business cycle (see for example Fatás and Mihov (2001)).

A second driver of the behaviour of the fiscal policy maker is the willingness to pursue a debt-stabilisation motive (a.o. Bohn!).

2.2 Specification of the model

We summarise the behaviour of fiscal policy makers with a reaction function that describes how a fiscal indicator f_t changes in response to government debt and to the business cycle. Assume the government has some long-term fiscal target f^* and that it decides to adjust this target at time t to control the deviation of debt b_t from some target level b^* . Given the structure of spending and taxation, the fiscal target will also fluctuate in response to expected deviations of output from some desired target output level y^* . The output response of the budget includes two components. In an economic boom, as output rises above its long-term level, unemployment benefits and transfer payments fall or tax receipts rise. In addition to these automatic stabilisers, the elasticity of these budget items with respect to output captures also systematic discretionary interventions of the government to steer the economy. The government may wish to lean against the wind during an economic crisis by cutting taxes or raising expenses. The fiscal reaction function for this time-varying target surplus \hat{f}_t can then be summarised as follows:

$$\hat{f}_t = f^* + \gamma(y_t - y^*) + \theta(b_t - b^*) \quad (1)$$

Given that the budgeting process is typically characterised by long implementation lags, we include a smoothing component to allow for a gradual adjustment towards the target. In contrast to previous studies (Favero & Monacelli, 2005; Afonso et al., 2010) and because the budgeting process is only executed once a year, we argue that a lag of four quarters is more suitable than a single quarter lag. Equation (2) presents this simple feedback rule:

$$f_t = \rho f_{t-4} + (1 - \rho)\hat{f}_t + v_t \quad (2)$$

Substitution of (1) into (2) gives the following non-linear relation between the fiscal indicator and public debt, and is the baseline fiscal rule we test:

$$f_t = \rho f_{t-4} + (1 - \rho)[\kappa + \gamma x_t + \theta b_t] + v_t \quad (3)$$

In (3), the output gap is given by x_t . The constant term κ can be interpreted as a long-term fiscal indicator: it adjusts the target surplus for the deviation between the government's output target and long-term potential output, and for the government debt target. Deviations from the rule, which are captured by the residual term, are discretionary changes in systematic fiscal policy. We allow the reaction coefficients γ and θ in (3) to vary over time. We moreover follow Favero and Monacelli (2005) and substitute debt b_t for the debt-stabilising primary deficit d_t in (3). This non-linear fiscal rule implicitly controls for the time-varying effects of interest rates and growth on the debt service component of the deficit that are not under direct control of the government itself. The interpretation of our specification of the fiscal rule (3) is quite straightforward and is compatible with the distinction that has typically been made in the literature between policies that pursue a debt-stabilising motive and those that do not (Sims, 1994). Leeper (1991) classifies a policy that pursues debt-stabilisation as 'passive' and as 'active' when it does not. In this case, fiscal policy is passive when the coefficient associated to the debt-stabilising primary deficit d_t is not statistically different from one. In addition, the constant term κ should not be statistically different from zero. A non-zero surplus would imply trend growth in debt. In contrast, fiscal policy is active if $\theta = 0$ and $\kappa \neq 0$. We assume that the reaction coefficients in (3) can change between different policy regimes. We estimate the fiscal rule with a Markov Switching (MS) model in which the probability of each different regime — indicated by the state m_t — can vary endogenously over time:

$$f_t = \rho(m_t)f_{t-4} + (1 - \rho(m_t))[\kappa(m_t) + \gamma(m_t)x_t + \theta(m_t)d_t] + v_t(m_t) \quad (4)$$

3 Data

We compile a new dataset covering the period between 1966Q1 and 2014Q4. The data are retrieved from NIPA Table 3.2 as calculated by the Bureau of Economic Analysis and from the FREDII database from the Federal Reserve Bank of Saint Louis. The primary deficit to GDP ratio f_t is calculated as the difference between Federal Government Current Receipts and the Federal Government Current Expenditure — net of Interest Payments — divided by the GDP. The output gap x_t is calculated as the percentage difference of real GDP and potential real GDP — as estimated by the Congressional Budget Office. The debt-stabilising primary deficit d_t is calculated following Favero and Monacelli (2005) and is depicted in figure 1.

4 Results

4.1 Testing for endogeneity

Before we estimate the MS model outlined above, we address the possible existence of endogenous explanatory variables. A large part of the literature suggests that, at least in the short run, fiscal policy does have an effect on output growth (Jaimovich & Panizza,

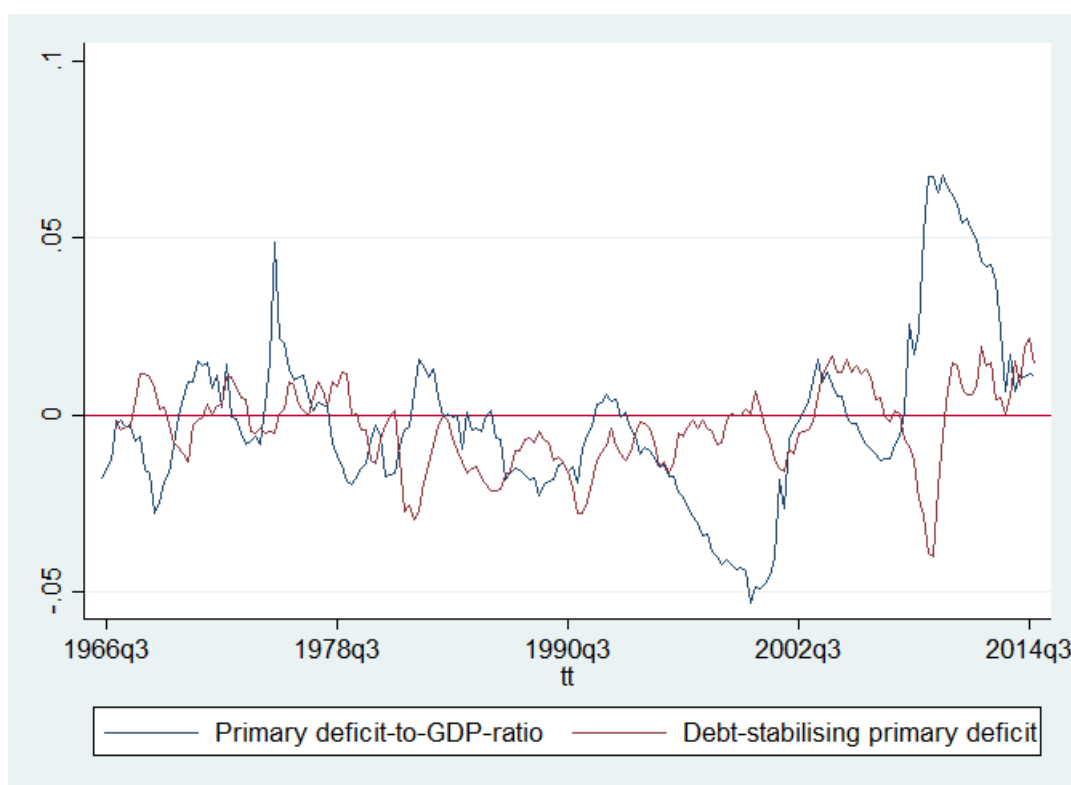


Figure 1: Debt-stabilising deficit vs. primary deficit in the US

2007). Under the existence of endogenous explanatory variables, it is not possible to consistently estimate MS models by applying the standard Hamilton (1989) filter (Bae, Kim, & Kim, 2012).

We first test for the hypothesis of endogeneity by using the standard Durbin and Wu-Hausman test after instrumenting x_t and d_t with their four-period lags. We apply the test on x_t and d_t individually as well as on the combination the two variables. Table 1 presents the p-values for the three possibilities as well as the results of a test for weak instruments.

Test statistic	x_t	d_t	x_t and d_t
Durbin score	.0000	.0000	.0000
Wu-Hausman	.0000	.0000	.0000
Minimum eigenvalue statistic	73.8143	48.8165	21.9101

Table 1: Durbin and Wu-Hausman tests of endogeneity

In either specification, we reject the null hypothesis of exogenous variables, which

confirms our a priori expectations about the existence of endogenous explanatory variables. The minimum eigenvalue statistic indicates that the chosen instruments should not be considered as weak.

4.2 Markov Switching estimates of the fiscal rule

In contrast to papers that only test the change in the debt response, or the symmetry of the cyclical response, we test for stochastic changes over time in all coefficients of (4). Additionally, we allow the variance of the shocks to switch between regimes. We start by estimating a two-regime MS model of (4) by maximum likelihood, using the Expectation-Maximization (EM) algorithm. Table 2 exhibits the results for estimation of the fiscal rule. First of all, and in contrast to previous work, our results indicate that for none of the regimes we find a parameter estimate for d_t that is not statistically different from one. According to the specification of the fiscal rule, we can not conclude that the US government pursued a debt-stabilising fiscal policy for the period covered. The transition probabilities for the basic model are given by figure 2.

State m_t	Var	Coef.	Std. Err.	t-ratio	95% Conf. Interval	
State 1	f_{t-4}	.665	.043	15.33	.580	.750
	x_t	-.204	.040	-5.06	-.283	-.125
	d_t	-.357	.084	-4.23	-.523	-.192
	κ	-.012	.001	-10.81	-.015	-.010
State 2	f_{t-4}	.629	.050	15.51	.530	.727
	x_t	-.326	.044	-7.39	-.413	-.240
	d_t	-.583	.097	-6.01	-.773	-.393
	κ	.004	.001	2.82	.001	.007

Table 2: Estimation results: Standard MS model

4.3 Accounting for endogeneity in MS models

However, as mentioned before the previous results should be evaluated with care because the endogeneity of the regressors leads to inconsistent estimators. To account for the endogenous character of the explanatory variables in this MS framework, we propose the same two-step MLE procedure as pioneered in Kim (2004) and further developed by Psaradakis, Sola, and Spagnolo (2006), Kim (2009) and Bae et al. (2012).

To be completed

5 Conclusion

The 2008–2009 financial crisis has shaken the beliefs about how macroeconomic policy should be conducted Central banks in G7 countries shifted to unconventional policy

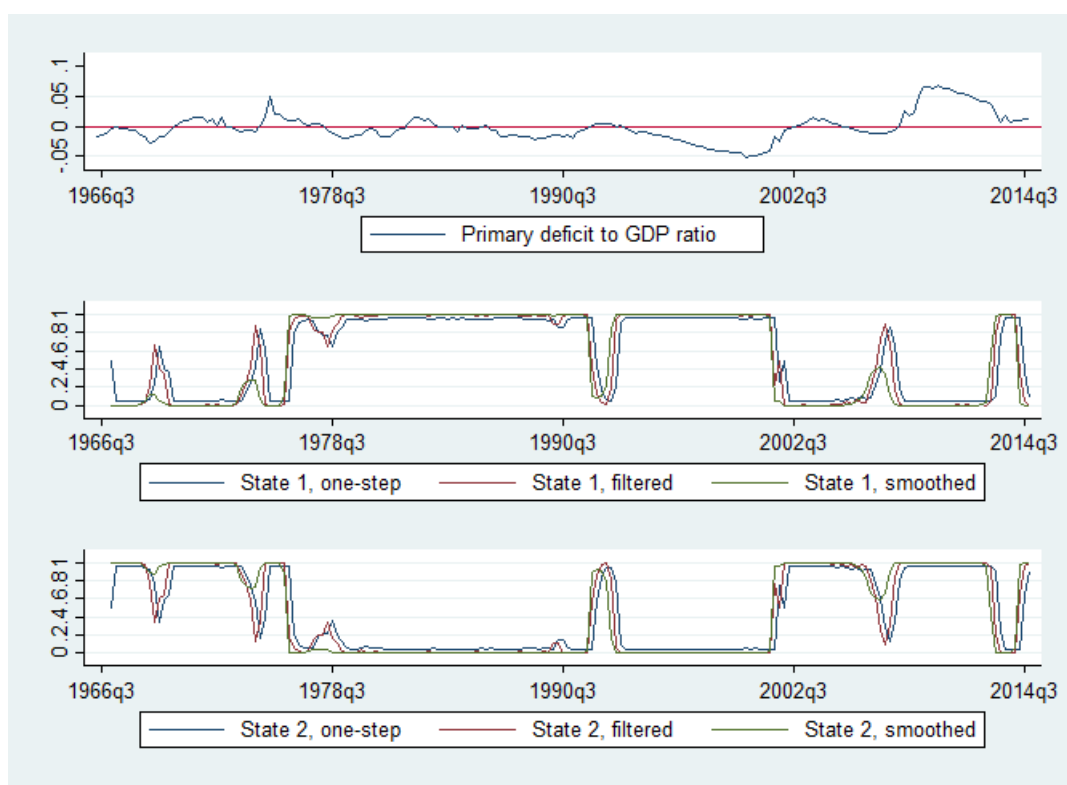


Figure 2: Transition probabilities of standard MS model

measures in the aftermath of the Financial Crisis, when faced with economic slack, financial instability and fiscal trouble. Governments were swift in saving the banking system, and the economy as a whole, from collapse. This shift ended a spell of rules-based time consistent policy that started in the mid-1980s and in many industrialised economies. Changes in policy regimes occur in response to economic or political events. We show that expectations on policy regimes cause the effects of policy actions to be anticipated by households. Their subjective expectations are modified by policy actions that make policy likely to move in one direction. We account for this endogeneity with a novel Markov Switching test of a fiscal rule. Results show that the stable mix of policies during the Great Moderation gave way to a mix of ‘passive’ monetary and ‘active’ fiscal policy since the mid-2000s.

References

- Afonso, A., Claey's, P., & Sousa, R. M. (2010). Fiscal regime shifts in Portugal. *Portuguese Economic Journal*, 10(2), 83–108.
- Bae, J., Kim, C.-J., & Kim, D. H. (2012). The evolution of the monetary policy regimes in the U.S. *Empirical Economics*, 43(2), 617–649.
- Barro, R. J., & Gordon, D. B. (1983). Rules, discretion and reputation in a model of monetary policy. *Journal of Monetary Economics*, 12(1), 101–121.
- Bianchi, F. (2013). Regime Switches, Agents' Beliefs, and Post-World War II U.S. Macroeconomic Dynamics. *The Review of Economic Studies*, 80(2), 463–490.
- Bianchi, F., & Melosi, L. (2014). Dormant Shocks and Fiscal Virtue. *NBER Macroeconomics Annual*, 28, 1–46.
- Davig, T., & Leeper, E. M. (2006). Fluctuating Macro Policies and the Fiscal Theory. *NBER Macroeconomics Annual*, 21, 247–305.
- Fatás, A., & Mihov, I. (2001). Fiscal Policy and Business Cycles: An Empirical Investigation. *Moneda y Credito*, 212.
- Fatás, A., & Mihov, I. (2003). On Constraining Fiscal Policy Discretion in EMU. *Oxford Review of Economic Policy*, 19(1), 112–131.
- Favero, C. A., & Monacelli, T. (2005). *Fiscal Policy Rules and Regime (In)Stability: Evidence from the U.S.* Retrieved from <http://papers.ssrn.com/abstract=665506>
- Gali, J., & Perotti, R. (2003). Fiscal policy and monetary integration in Europe. *Economic Policy*, 18(37), 533–572.
- Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle on JSTOR. *Econometrica*, 57(2), 357–384.
- Jaimovich, D., & Panizza, U. G. (2007). *Procyclicality or Reverse Causality?* Retrieved from <http://papers.ssrn.com/abstract=1820869>
- Kim, C.-J. (2004). Markov-switching models with endogenous explanatory variables. *Journal of Econometrics*, 122(1), 127–136.
- Kim, C.-J. (2009). Markov-switching models with endogenous explanatory variables II: A two-step MLE procedure. *Journal of Econometrics*, 148(1), 46–55.
- Kydland, F. E., & Prescott, E. C. (1977). Rules Rather than Discretion: The Inconsistency of Optimal Plans on JSTOR. *Journal of Political Economy*, 85(3), 473–492.
- Leeper, E. M. (1991). Equilibria under active' and passive' monetary and fiscal policies. *Journal of Monetary Economics*, 27(1), 129–147.
- Psaradakis, Z., Sola, M., & Spagnolo, F. (2006). Instrumental-variables estimation in Markov switching models with endogenous explanatory variables: An application to the term structure of interest rates. *Studies in Nonlinear Dynamics & Econometrics*, 10(2).
- Sims, C. A. (1994). A simple model for study of the determination of the price level and the interaction of monetary and fiscal policy. *Economic Theory*, 4(3), 381–399.

Taylor, J. B. (1993, dec). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39, 195–214.

Robust autocovariance estimation from the frequency domain

Higor Cotta^{1,2}, Valdério Reisen^{1,2}, and Pascal Bondon²

¹NuMEs - DEST/PGGEA - Federal University of Espírito Santo - Brazil

²L2S-CNRS - CentraleSupélec - France

`higor.cotta@l2s.centralesupelec.fr`

`valderio.reisen@ufes.br`

`pascal.bondon@l2s.centralesupelec.fr`

Abstract. This paper proposes a robust estimation method of the sample autocovariance and autocorrelation functions in the presence of additive outliers. The robustness property is achieved by replacing the standard Fourier transform by its robustified version obtained by substituting the least square procedure in the harmonic regression by the non-linear M-regression. Simulation experiments are conducted to assess the performance of the estimators under contaminated and non-contaminated scenarios.

Keywords: Autocovariance Function; Outliers; M-periodogram; Estimation

Atypical observations (outliers) are present in time series of diversified origins. It is well known that outliers significantly destroy the correlation structure of a time series even when only one atypical observation is present, see, for example, [1–3] and the references therein. As a possible approach for solving this problem, [4] proposed a highly robust estimator of the autocovariance function (ACOVF) and autocorrelation function (ACF), denoted by $\hat{\gamma}_Q(\cdot)$ and $\hat{\rho}_Q(\cdot)$, respectively. These estimators are based on the $Q_n(\cdot)$ scale estimator proposed by [5], whose asymptotic properties were studied by [6] for univariate time series.

As noticed by [4], their robust ACOFV estimator does not provide a non-negative definite sample covariance matrix. Although this is an undesirable property for an autocovariance function estimator, the highly robust performance of $\hat{\gamma}_Q(\cdot)$ motivated its adoption by [3] to obtain an estimator of the spectral density function which is robust against additive outliers.

In addition time series analysis in the frequency domain is based on the study of the spectral density function from which the periodogram is an estimator. As demonstrated by [3], the periodogram lacks robustness properties against outliers. Therefore, robust methods to minimize the effect of outliers on the estimation of periodogram have to be considered. In this direction, different robust periodogram methods have been proposed by the literature, see, for instance, [7–11], among others.

It is known that the periodogram can be obtained directly from a least squares estimates of the Fourier coefficients and is hence sensitive to outliers in data. Thus, to mitigate this problem, one may consider the use of robust regression methods, e.g., a robust M-regression, instead of the standard approach. Recently, this approach has been considered by [10] and [11] providing good results in the estimation of the coefficients of PARMA models and the fractional parameter of ARFIMA models, respectively.

In addition, the sample autocovariance function and periodogram are related by means of the Fourier transform. Thus, this work considers the estimation of the sample autocovariance and autocorrelation functions from the robust M-periodogram. The approach consists in fitting a robust harmonic regression to obtain a robustified version of the discrete Fourier transform. That is, at each Fourier frequency, a sine and cosine coefficients are fitted using M-regression. Then, the ACOVF and ACF are obtained by taking the inverse of the squared robust Fourier transform.

References

1. Chan, W.: A note on time series model specification in the presence of outliers. *Journal of Applied Statistics* **19**(1) (1992) 117–124
2. Chan, W.: Outliers and financial time series modelling: a cautionary note. *Mathematics and Computers in Simulation* **39**(3) (1995) 425–430
3. Molinares, F.F., Reisen, V.A., Cribari-Neto, F.: Robust estimation in long-memory processes under additive outliers. *Journal of Statistical Planning and Inference* **139**(8) (2009) 2511–2525
4. Ma, Y., Genton, M.G.: Highly robust estimation of the autocovariance function. *Journal of Time Series Analysis* **21** (2000) 663–684
5. Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**(424) (1993) 1273–1283
6. Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M.S., Reisen, V.A.: Robust estimation of the scale and of the autocovariance function of Gaussian short-and long-range dependent processes. *Journal of Time Series Analysis* **32**(2) (2011) 135–156
7. Zhang, Z., Chan, S.C.: Robust adaptive lomb periodogram for time-frequency analysis of signals with sinusoidal and transient components. In: *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. Volume 4., IEEE (2005) iv–493
8. Li, T.H.: Laplace periodogram for time series analysis. *Journal of the American Statistical Association* **103**(482) (2008) 757–768
9. Li, T.H.: A nonlinear method for robust spectral analysis. *IEEE Transactions on Signal Processing* **58**(5) (2010) 2466–2474
10. Sarnaglia, A.J.Q., Reisen, V.A., Bondon, P., Lévy-Leduc, C.: A robust estimation approach for fitting a PARMA model to real data. In: *2016 IEEE Statistical Signal Processing Workshop (SSP)*. (2016)
11. Reisen, V., Lévy-Leduc, C., Taqqu, M.: An m-estimator for the long-memory parameter. *Journal of Statistical Planning and Inference* **187** (2017) 44 – 55

Event Related Causality analysis of electrocorticographic (ECoG) time series as diagnostic tool for epileptic surgery.

Anna Korzeniewska^{1*}(0000-0002-9453-7751), Piotr J. Franaszczuk²(0000-0002-5166-4224),

Nathan E. Crone¹(0000-0001-7950-2617)

¹Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, USA
akorzen@jhmi.edu, ncrone@jhmi.edu

²Human Research and Engineering Directorate, US Army Research Laboratory, Aberdeen Proving Ground, USA
pfranasz@gmail.com

Abstract. Event Related Causality (ERC), a model-free method based on the concept of Granger causality, employs multivariate autoregressive model (MVAR) to investigate dynamics of directed interactions among neural networks. ERC is designed to determine the direction and intensity of neural activity propagation, and to selectively indicate only direct propagations. Short-time event-related changes are followed using an adaptive approach to stochastic non-stationary signals analysis. The statistical method accounts for a non-stationary baseline as well. Application of ERC to human electrocorticographic recordings (ECoG) shows the dynamics of high gamma activity flow among cognitive networks, as well as high frequency oscillations (HFOs) propagation in epileptic pathology. Moreover, ERC identifies networks' nodes, which may help to discriminate cognitive vs. epileptogenic neural networks when planning for epilepsy surgery.

Keywords: Multivariate autoregressive model • Granger causality • EEG • Stochastic non-stationary signals • Neural networks • Event Related Causality (ERC)

1 Introduction

Oscillations of frequency range 60-200 Hz, called high gamma, have been proposed to play a role in the dynamic organization of neuronal assemblies in large-scale brain networks responsible for cognition and other brain functions [1-3]. Oscillatory activity arising from one subnetwork is hypothesized to have a causal influence on activity of another subnetwork. These causal influences are expected to occur with different strengths and directionalities, reflecting the changing functional demands on cortical networks during task performance and thereby identify network nodes crucial for function.

On the other hand, high frequency oscillations (HFOs) of same frequency range and higher, have been observed within seizure onset zone [4-5]. Seizures are understood to arise from epileptogenic networks across

which ictal activity is propagated and sustained. The pattern by which high frequency activity is propagated may help elucidate epileptogenic networks, and thereby identify network nodes relevant for surgical planning [6].

To capture the dynamics of such functional and pathological brain interactions, their directions, intensities, spectral characteristics, as well as networks' nodes, of both functional and pathological activity propagation, event-related causality (ERC) method is proposed.

The rest of this paper is structured as follows. Section 2 introduces the Event Related Causality (ERC) method and presents its effectiveness in identifying patterns of activity flows in simulated models. Section 3 shows examples of ERC applications to human cognitive and epileptogenic neural networks. Section 4 concludes advantages of ERC.

2 Event Related Causality

ERC method is a multichannel extension of the Granger causality concept [7], which states that an observed time series $y(t)$ causes another time series $x(t)$, if knowledge of $y(t)$'s past significantly improves prediction of $x(t)$ (Fig. 1).

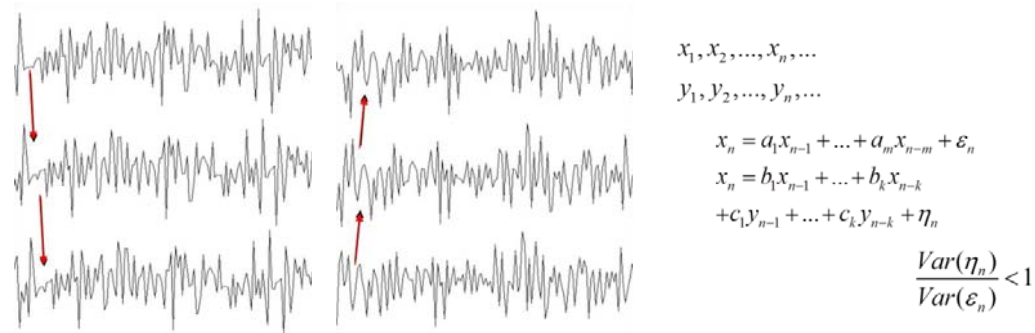


Fig. 1. Granger causality

Unfortunately, nonzero values of multichannel Granger causality between two signals do not necessarily imply that the causal influences between counterpart recording sites are direct. Multichannel Granger causality represents a linear combination of causal influences along all causal pathways – direct and indirect – originating from one signal-site and terminating at another. The influence may be mediated by another site or by several sites. To overcome this limitation, a previous multichannel extension of Granger causality was combined with partial coherence.

In a multivariate autoregressive (MVAR) model, recorded signals can be expressed:

$$x(t) = -\sum_{j=1}^p A_j x(t-j) + e(t) \quad (1)$$

where A_j is a MVAR coefficients matrix, $e(t)$ is a zero-mean uncorrelated residual noise vector, and p is the model order. The model order can be obtained from Akaike information criterion [8].

After transformation to the frequency domain, a transfer matrix H of the multivariate process can be obtained [9]:

$$H(f) = \left(\sum_{j=0}^p A_j e^{-i2\pi j f \Delta t} \right)^{-1} \quad (2)$$

By combining elements h_{kl} of transfer matrix H with the partial coherence we have [10]:

$$X_{kl}(f) = \frac{c_{kl}(f)}{\sqrt{c_{kk}(f)c_{ll}(f)}} \quad (3)$$

the short-time direct directed transfer function (SdDTF) is defined in the form [11]:

$$\zeta_{kl}(f) = \frac{|h_{kl}(f)| |X_{kl}(f)|}{\sqrt{\sum_f \sum_{kl} |h_{kl}(f)|^2 |X_{kl}(f)|^2}} \quad (4)$$

SdDTF (Fig.2 far right, and Fig. 3 - far right) gives an estimate of the intensity and direction of activity propagation between recording sites as a function of frequency. This measure is designed to selectively index only direct relationships. A newly introduced standardization procedure makes it possible to compare estimates of information flow across subjects [11].

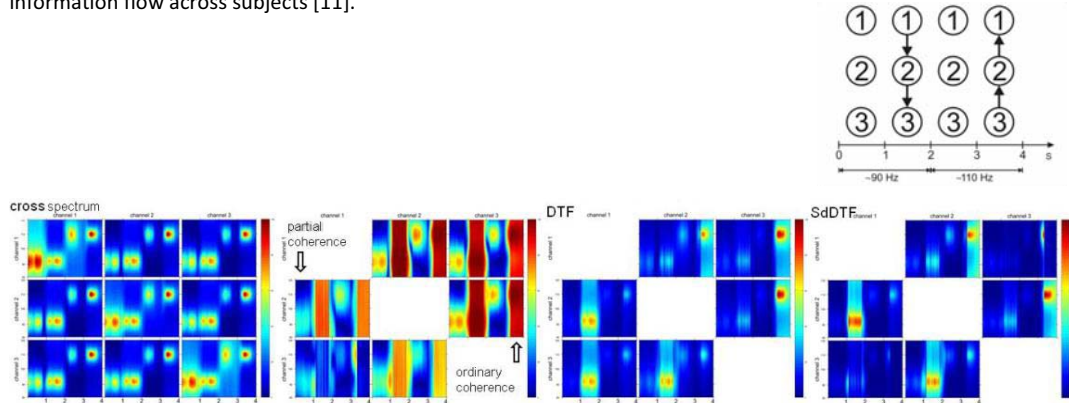


Fig. 2. Top: Schematic of simulated model of activity flows. 1st time interval; all signals contained the same spectral components ~90 Hz, and different white noise, but no flow (no causal relations) between channels was simulated. 2nd interval; ~90 Hz flows from channel 1 to 2 (1→2), and from 2 to 3 (2→3). 3rd interval; ~110 Hz components, but no flow. 4th interval; ~110 Hz flows 3→2 and 2→1 were simulated.

Bottom: Cross-spectra, coherences, SdDTFs, and SdDTFs of multivariate MVAR model of the simulated signals. In each plot, horizontal axis - time, vertical axis - frequency, colorscale - value of calculated functions (blue - minimum, red - maximum). SdTF and SdDTF matrices are not symmetric; each plot shows flows from the channel labeled above the plot to the channel labeled to the left of the plot. SdDTF measures only direct activity flows.

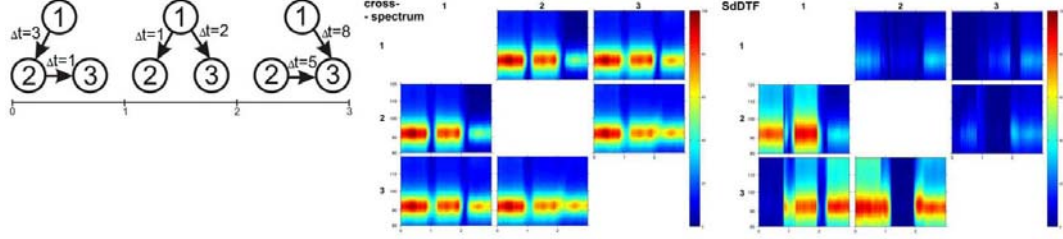


Fig. 3. Left: schematic of another simulated model of activity flows. Center: cross-spectra. Right: SdDTF.

To follow the temporal course of brief changes in neural activity propagation between different brain regions, an algorithm introduced by Ding [12] was used for MVAR coefficient estimation in multichannel signals recorded during multiple repetitions (trials) of the same process. This enables the analysis of nonstationary signals such as ECoG activity accompanying cognitive processes.

To evaluate the statistical significance of event-related changes in SdDTF, *i.e.* event-related causality (ERC), a new statistical methodology was developed for comparing pre-stimulus (baseline) with post-stimulus SdDTF values. Both the baseline and post-stimulus epochs are treated as non-stationary. Bivariate smoothing was applied, using a penalized thinplate spline model, to construct a joint 95% confidence interval [13], while the Family-Wise Error Rate (FWER) was controlled using the Bonferroni correction. The implicit null hypothesis was that:

$$H_O^T: \mu(F_1) - \mu(G_T) = 0 \text{ or ... or } \mu(F_{t_0}) - \mu(G_T) = 0 \quad (5)$$

with the corresponding alternative

$$H_A^T: \mu(F_1) - \mu(G_T) \neq 0 \text{ and ... and } \mu(F_{t_0}) - \mu(G_T) \neq 0 \quad (6)$$

F_t denotes the SdDTF baseline probability distribution at time t , with $1 \leq t \leq t_0$ where t_0 is the start of the last time-window within the baseline. Similarly, G_T denotes the SdDTF probability distribution at time T after stimulus, with $1 \leq T \leq T_0$ where T_0 is the start of the last post-stimulus time-window. The means of the probability distributions F_t and G_T , are respectively denoted by $\mu(F_t)$ and $\mu(G_T)$. This test rejects H_O^T if zero is not contained in one of the confidence intervals.

3 Neural networks

3.1. ERC reveals dynamics of high gamma activity propagation during cognitive processes

The utility of the ERC method has been demonstrated through its application to human electrocorticographic signals (ECoG) recorded during simple language tasks [11]. ERC analysis revealed frequency-dependent

interactions, particularly in the high gamma frequency range (80-100 Hz), between brain regions known to participate in the language processes. The temporal evolution of these interactions is consistent with the putative processing stages of this task (Fig. 4).

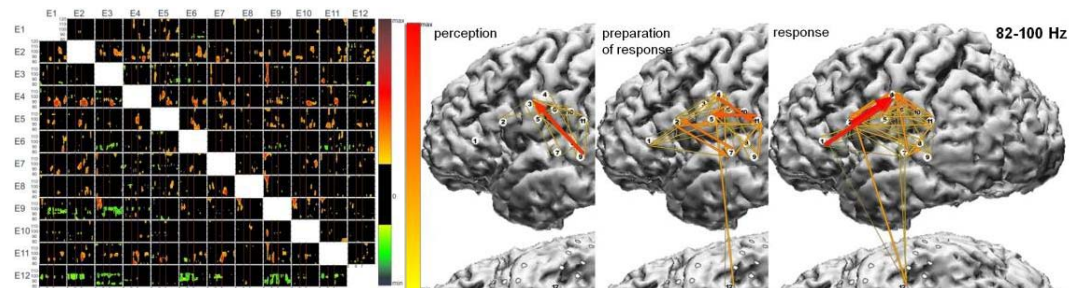


Fig. 4. Left: ERC matrix of ECoG recorded during auditory word repetition. Colorscale (min to max) - to right of array. Black indicates time-frequency points with no significant difference between SdDTF after stimulus and SdDTFs for baseline. Yellow-green-blue show event-related decreases of poststimulus SdDTF. Orange-red-brown show event-related increases. Right: Integrals of ERC for high gamma frequency 82–100 Hz calculated for three stages of auditory word repetition task: auditory perception, response preparation, and verbal response. Arrows indicate directionality of ERC, width and color represent the value of ERC integral. Colorscale at the left. For clarity, only integrals for event-related flow increases are shown [11].

3.2. Divergence of ERC flows identifies nodes of language network, characterized also by prominent changes in signal energy

ERC analyses of ECoG signals recorded during a bimodal word production task, when responses were spoken or were gestured in American Signed Language (ASL), revealed that the language cortex interacts with different areas of the sensorimotor cortex during spoken vs. signed responses (mouth/tongue areas vs. hand and arm areas, Fig 5).

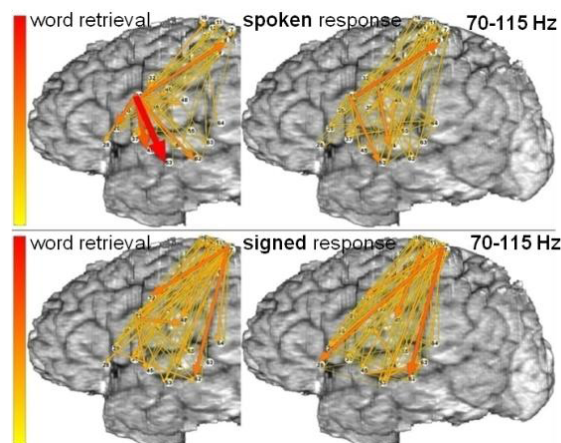


Fig. 5. Integrals of ERC flows at 70–115 Hz calculated for two sequential time intervals of picture naming task with spoken (top panels) vs. signed responses (bottom panels).

Furthermore, it has been shown that the sites from which the most numerous and prominent causal interactions originated, i.e. sites with a pattern of ERC “divergence”, were also sites where high gamma power increases were most prominent and where electrocortical stimulation mapping interfered with word production. (Fig 6).

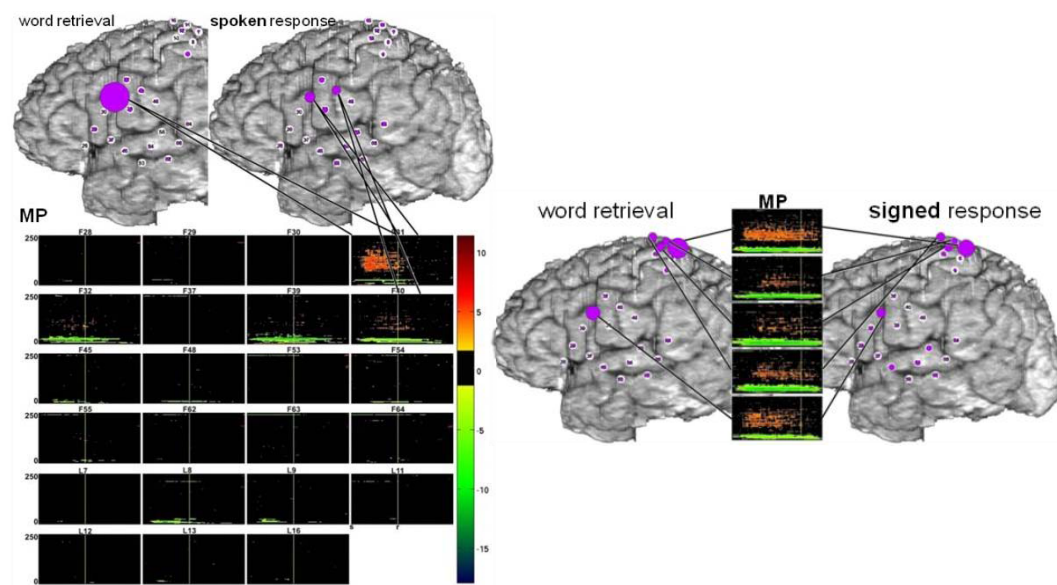


Fig. 6. Comparison of event-related causal interactions vs. functional activation during picture naming with spoken and signed responses. Top-left - relative magnitudes of ERC outflows from each site during picture naming with spoken responses. The radius of each circle is proportional to normalized sum of statistically significant event-related increases in causal interactions directed outwardly from the site. Bottom left - power changes shown by matching pursuit (MP) in the time-frequency plane for the same task. Right: Comparison of event-related causal interactions (top-right) vs. functional activation (bottom-right) during picture naming with signed responses [14].

These findings suggest that the number, strength and directionality of event-related causal interactions may help identify network nodes that are not only activated by a task but are critical to its performance [14].

3.3. Divergence/convergence of high-frequency epileptic activity identifies nodes of epileptogenic networks. Locations of these nodes correspond to locations of seizure foci identified by epileptologists

Analyses of ECoG ictal recordings (i.e. during epileptic seizures), as well as interictal recordings, revealed prominent divergence and convergence of high frequency activity propagation at sites identified by epileptologists as part of the ictal onset zone (Fig. 7). In contrast, relatively little propagation of this activity was observed among the other analyzed sites. This pattern was observed in both subdural and depth electrode recordings of patients with focal ictal onset, but not in patients with a widely distributed ictal onset.

These patterns elucidated epileptogenic networks, and thereby identified network nodes relevant for surgical planning [15].

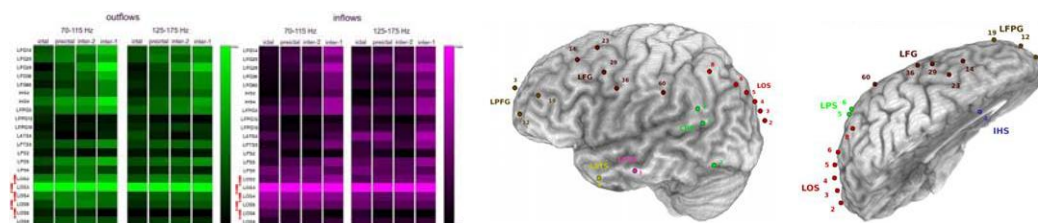


Fig. 7. SdDTF for propagations between the site of ictal onset and all other recording sites, for 70-115 Hz and 125-175 Hz, during ictal and preictal periods. Each column shows average of SdDTFs for propagations *from* the site indicated at the left to all other recording sites (outflowing, green-black), and *to* the site indicated at the left from all other recording sites (inflowing, purple-black). The highest magnitudes indicate the largest outflows (the brightest green), or the largest inflows (brightest purple) [15].

ERC analyses have been successfully applied to EEG recordings from human scalp [16], as well as in group analyses of several patients [16-17].

4 Conclusions

- ERC is a sensitive method for determining dynamics of causal interactions among cognitive and epileptic networks.
- ERC identifies nodes of functional and pathological networks.
- ERC may help to discriminate cognitive vs. epileptogenic networks when planning for epilepsy surgery.

References

1. Crone, N.E., Korzeniewska, A., Franaszczuk, P.J.: Cortical gamma responses: searching high and low. *Int J Psychophysiol.* 79(1), 9–15 (2011). doi: 10.1016/j.ijpsycho.2010.10.013
2. Lachaux, J-P., Axmacher, N., Mormann, F., Halgren, E., Crone, N.E.: High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research *Progress in Neurobiology* 98, 279–301 (2012). doi: 10.1016/j.pneurobio.2012.06.008
3. Bressler, S.L., Menon, V.: Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn Sci* 14,277-290. (2010). doi: 10.1016/j.tics.2010.04.004
4. Wood H.: Epilepsy: High-frequency oscillations pinpoint the seizure-onset zone. *Nature Reviews Neurology* 7, 475 (2011). doi:10.1038/nrneurol.2011.127
5. Jiruska, P., Alvarado-Rojas, C., Schevon, C.A., Staba, R., Stacey, W., Wendling, F., Avoli, M.: Update on the mechanisms and roles of high-frequency oscillations in seizures and epileptic disorders. *Epilepsia*. In Print (2017) Jul 6. doi: 10.1111/epi.13830
6. Wilke, C., Worrell, G., He, B.: Graph analysis of epileptogenic networks in human partial epilepsy. *Epilepsia* 52(1), 84-93 (2011). doi: 10.1111/j.1528-1167.2010.02785.x.
7. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424-438 (1969). doi:10.2307/1912791
8. Akaike H.: New Look at Statistical-Model Identification. *IEEE Transactions on Automatic Control* AC19(6):716-723(1974). doi:10.1109/TAC.1974.1100705
9. Kaminski, M.J., Blinowska, K.J.: A new method of the description of the information flow in the brain structures. *Biol Cybern* 65, 203–210 (1991). doi:10.1007/BF00198091

10. Korzeniewska, A., Manczak, M., Kaminski, M., Blinowska, K.J., Kasicki, S.: Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method. *J Neurosci Methods* 125, 195–207 (2003). doi:10.1016/S0165-0270(03)00052-9
11. Korzeniewska, A., Crainiceanu, C.M., Kus, R., Franaszczuk, P.J., Crone, N.E.: Dynamics of event-related causality in brain electrical activity. *Hum Brain Mapp* 29, 1170–1192(2008). doi:10.1002/hbm.20458
12. Ding M, Bressler SL, Yang W, Liang H.: Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biol Cybern* 83(1):35–45(2000). doi:10.1007/s004229900137
13. Ruppert D., Wand M.P., Carroll R.J.: *Semiparametric Regression*. Cambridge: Cambridge University Press. xvi, 386 pp (2003). doi:10.1017/CBO9780511755453
14. Korzeniewska, A., Franaszczuk, P.J., Crainiceanu, C.M., Kus, R., Crone, N.E.: Dynamics of large-scale cortical interactions at high gamma frequencies during word production: event related causality (ERC) analysis of human electrocorticography (ECoG). *Neuroimage* 56, 2218–2237(2011). doi:10.1016/j.neuroimage.2011.03.030
15. Korzeniewska, A., Cervenka, M.C., Jouny, C.C., Perilla, J.R., Harezlak, J., Bergey, G.K., Franaszczuk, P.J., Crone, N.E.: Ictal propagation of high frequency activity is recapitulated in interictal recordings: effective connectivity of epileptogenic networks recorded with intracranial EEG. *Neuroimage* 101, 96–113(2014). doi:10.1016/j.neuroimage.2014.06.078
16. Ewen, J.B., Lakshmanan, B.M., Hallett, M., Mostofsky, S.H., Crone, N.E., Korzeniewska, A.: Dynamics of functional and effective connectivity within human cortical motor control networks. *Clin Neurophysiol.* 126(5), 987–96 (2015). doi: 10.1016/j.clinph.2014.09.006
17. Nishida, M., Korzeniewska, A., Crone, N.E., Toyoda, G., Nakai, Y., Ofen, N., Brown, E.C., Asano, E.: Brain network dynamics in the human articulatory loop. *Clin Neurophysiol.* 128(8), 1473–1487 (2017). doi: 10.1016/j.clinph.2017.05.002.

Sieves Estimators and Predictors for Functional Autoregressive Processes

Nesrine Kara-Terki and Tahar Mourid

Laboratoire de Statistiques et Modélisations Aléatoires
Université Abou Bekr Belkaid -Tlemcen- 13000 - Algeria

Abstract We study sieves estimators for a class of functional autoregressive processes when the parameter operator belongs to classes of Hilbert-Schmidt operators. We then show the almost sure convergence, almost complete convergence, exponential bounds and obtain rates of convergence of the sieves estimators in each class. We also present the same results of convergence for the sieves predictors. The rates of convergence are deeply depend on smoothness of the sieves and decay rate of the eigenvalues of the parameter operator. Numerical simulations illustrate the behavior of the sieves predictors.

Keywords : Functional Autoregressive Processes; Sieves Estimators; Exponential Bounds; Sieves Predictors; Rates of Convergence.

1 Notations and Results

Let (H, \mathcal{H}) be a real separable Hilbert with inner product $\langle \cdot, \cdot \rangle$, the norm $\|\cdot\|$. Let $\mathcal{L}(H)$ be the space of linear bounded operators on H , $\|\cdot\|_{\mathcal{L}}$ the norm of linear bounded operators. A H -valued Gaussian white noise $(\varepsilon_n, n \in \mathbb{Z})$ with zero mean and covariance operator C_ε , is a Gaussian sequence of i.i.d. H -valued rv's. A H -valued process $X = (X_n, n \in \mathbb{Z})$ is said to be an Hilbertian autoregressive process if

$$X_n = \rho(X_{n-1}) + \varepsilon_n \quad (1)$$

where the parameter $\rho \in \mathcal{L}(H)$ such that $\|\rho\|_{\mathcal{L}} < 1$.

The parameter $\rho \in \Theta$ a parameter space which will be an Hilbert space of the space $\mathcal{S}^2(H)$ of Hilbert-Schmidt operators on H and will be specified later. Let $X^n = (X_0, \dots, X_n)$ be observations of (1), $P_{n,\rho} = P_{(X_1, \dots, X_n, \rho)}^{X_0}$ is the conditional probability law of (X_1, \dots, X_n) given X_0 and $P_{n,\varepsilon}$ is the probability law of $(\varepsilon_1, \dots, \varepsilon_n)$. We denote $f(X^n, \rho) := \frac{dP_{n,\rho}}{dP_{n,\varepsilon}}$ the derivative of the absolute continuous part of $P_{n,\rho}$ with respect to $P_{n,\varepsilon}$. The probability law of $X = (X_n, n \in \mathbb{Z})$ is denoted by P_ρ . We consider the sieves estimators of Grenander (or projection method). A sequence $\{S_k, k \geq 1\}$ of subsets of Θ is called a sieves if S_k is a compact set, $\{S_k\}$ is increasing sequence and $\bigcup S_k$ is a dense set in Θ . A sequence of estimators $\{\hat{\rho}_{n,k}\}$ is called a sieves estimators corresponding to $\{S_k\}$ if it satisfies

$$f(X^n, \hat{\rho}_{n,k}) = \sup_{\rho \in S_k} f(X^n, \rho) \quad (2)$$

For a sieves $\{S_k, k \geq 1\}$ of Θ , we introduce the following notations :

1. $\underline{\rho}_k$ is the orthogonal projection of ρ on S_k .
2. Let $\beta > 0$, $U_{n,k} = n^\beta(S_k - \underline{\rho}_k)$ and for $u \in U_{n,k}$, let $Z_{n,k}(u) = \log[f(X^n, \underline{\rho}_k + u/n^\beta)/f(X^n, \underline{\rho}_k)]$ be the likelihood process.
3. For $x \in H$, $a_j(x) = \langle x, e_j \rangle$ denotes the Fourier coefficient of x with respect to e_j and $[a]$ is the largest integer less than or equal to $a \in \mathbb{R}$.

Let $\{e_j, \sigma_j^2, j \geq 0\}$ be the eigen-elements of the covariance operator C_ε of ε_0 such that $\sigma_j^2 > 0$. We set the following condition on the eigenvalues $\{\sigma_j^2, j \geq 0\}$.

A0. $\min(\inf_{r \geq 0} \frac{\sigma_r^2}{\sigma_{r-1}^2}, \inf_{r \geq 0} \frac{\sigma_r^2}{\sigma_{r+1}^2}) > 0$, $\max(\sup_{r \geq 0} \frac{\sigma_r^2}{\sigma_{r-1}^2}, \sup_{r \geq 0} \frac{\sigma_r^2}{\sigma_{r+1}^2}) < \infty$ with $\sigma_{-1} = \sigma_0$

We introduce two classes of parameters.

First Class : the parameter space $\tilde{\Theta}$ is defined by

$$\tilde{\Theta} = \{\rho \in \mathcal{S}^2(H) / \rho \text{ is symmetric and commutes with } C_\varepsilon\}.$$

The space $\tilde{\Theta}$ is a separable Hilbert space of $\mathcal{S}^2(H)$ with inner product $\langle \rho_1, \rho_2 \rangle_2 = \sum_{j \geq 0} \lambda_j(\rho_1) \lambda_j(\rho_2)$ where $\{\lambda_j(\rho), j \geq 0\}$ are the eigenvalues of ρ and norm $\|\cdot\|_2$.

We define the sieves $\{S_k, k \geq 1\}$ in the parameter space $\tilde{\Theta}$: for $q > 0$

$$S_k = \{\rho \in \tilde{\Theta} / \rho = \sum_{j=0}^k \langle \rho, s_j \rangle_2 s_j, \sum_{j=0}^k \langle \rho, s_j \rangle_2^2 \leq k^{2q}\}.$$

Second Class : the parameter space $\tilde{\Theta}^*$ is the kernel operators space acting on the functions space $L^2_{[0,1]}$. An operator ρ in $\tilde{\Theta}^*$ is defined for $\theta \in L^2 = L^2_{[-1,1]}$ by

$$\rho f(t) := \rho_\theta f(t) = \int_0^1 \theta(t-s) f(s) ds. \quad (3)$$

The parameter is now the kernel θ . We suppose that the eigenfunctions of C_ε are the trigonometric base : $e_0 = 1$, $e_{2j}(t) = \sqrt{2} \cos(2\pi j t)$, $j \geq 1$, $e_{2j+1}(t) = \sqrt{2} \sin(2\pi(j+1)t)$, $j \geq 0$, $t \in [-1, 1]$ and eigenvalues $\{\sigma_j^2, j \geq 0\} \in l^1$.

We make the following conditions on the kernel θ :

A1. θ is a continuous periodic function, of period 1 and $\|\theta\|_{L^2} < 1$.

A2. θ is an even function.

A3. θ is an odd function.

We introduce the sieves $\{\Theta_k, k \geq 1\}$ and $\{\Theta_k^*, k \geq 1\}$ of $L^2_{[-1,1]}$: for $q > 0$,

$$\Theta_k = \{\theta \text{ satisfies } \mathbf{A1} \text{ and } \mathbf{A2} / \theta = \sum_{j=0}^k \langle \theta, e_j \rangle e_j, \sum_{j=0}^k a_j^2(\theta) \leq k^{2q}\}.$$

$$\Theta_k^* = \{\theta \text{ satisfies } \mathbf{A1} \text{ and } \mathbf{A3} / \theta = \sum_{j=1}^k \langle \theta, e_j \rangle e_j, \sum_{j=1}^k a_j^2(\theta) \leq k^{2q}\}.$$

and their respective sieves estimators $\hat{\theta}_{n,k}$, $\hat{\theta}_{n,k}^*$ of θ .

For the two classes we have the expressions of the sieves estimators.

Theorem 1.1 .

1. Let $\rho \in \tilde{\Theta}$ and $q > 0$. Then the sieves estimators are given by : for all n, s ,

$$\hat{\rho}_{n,[n^s]}x = \sum_{j=0}^{[n^s]} \hat{\lambda}_j \langle x, e_j \rangle e_j, \quad x \in H \quad (4)$$

where for $0 \leq j \leq [n^s]$, $\hat{\lambda}_j = \frac{\sum_{i=1}^n a_j(X_i) a_j(X_{i-1})}{2\mu\sigma_j^2 + \sum_{i=1}^n a_j^2(X_{i-1})}$ and $\mu > 0$ is the Lagrange multiplier such that $\sum_{j=0}^{[n^s]} \hat{\lambda}_j^2 = [n^s]^{2q}$.

Theorem 1.2 .

Let $\rho \in \tilde{\Theta}^*$ and defined by (3) and $m_n = 2[n^s/2]$.

1. Suppose **A1** and **A2**. Then for all $t \in [0, 1]$,

$$\hat{\theta}_{n,m_n}(t) = \sqrt{2} \sum_{j=1, j, \text{ even}}^{m_n} \hat{r}_j e_j(t) + \hat{r}_0 e_0(t)$$

where $\hat{r}_0 = \frac{\sum_{i=1}^n a_0(X_i) a_0(X_{i-1})}{2\mu\sigma_0^2 + \sum_{i=1}^n a_0^2(X_{i-1})}$ and for $2 \leq j \leq m_n$, j even,

$$\hat{r}_j = \frac{\sum_{i=1}^n \frac{a_j(X_i) a_j(X_{i-1})}{\sigma_j^2} + \sum_{i=1}^n \frac{a_{j-1}(X_i) a_{j-1}(X_{i-1})}{\sigma_{j-1}^2}}{2\mu + \sum_{i=1}^n \frac{a_j^2(X_{i-1})}{\sigma_j^2} + \sum_{i=1}^n \frac{a_{j-1}^2(X_{i-1})}{\sigma_{j-1}^2}}$$

and $\mu > 0$ is the Lagrange multiplier such that $\sum_{j=0, j \text{ even}}^{m_n} \hat{r}_j^2 = m_n^{2q}$.

2. Suppose **A1** and **A3**. Then for $0 \leq t \leq 1$,

$$\hat{\theta}_{n,m_n}^*(t) = \sum_{j=1, j, \text{ odd}}^{m_n} \sqrt{2} \hat{r}_j^* e_j(t)$$

where for $1 \leq j \leq m_n$, j odd,

$$\hat{r}_j^* = \frac{\sum_{i=1}^n \frac{a_j(X_i) b_j(X_{i-1})}{\sigma_j^2} - \sum_{i=1}^n \frac{a_{j+1}(X_i) b_{j+1}(X_{i-1})}{\sigma_{j+1}^2}}{2\mu + \sum_{i=1}^n \frac{b_j^2(X_{i-1})}{\sigma_j^2} + \sum_{i=1}^n \frac{b_{j+1}^2(X_{i-1})}{\sigma_{j+1}^2}}$$

μ is the Lagrange multiplier such that $\sum_{j=1, j, \text{ odd}}^{m_n} (\hat{r}_j^*)^2 = m_n^{2q}$ and $b_j(x) = \langle x, \psi_j \rangle$ where the base (ψ_j) is a base in H .

For the first class $\tilde{\Theta}$, we have a.s. convergence, exponential bounds and rates for the sieves estimators.

Theorem 1.3 . Let $\rho \in \tilde{\Theta}$. Then the sieves estimators $\{\hat{\rho}_{n,k}\}$ defined by (2) satisfy :

i) $\hat{\rho}_{n,[n^s]} \rightarrow_{n \rightarrow \infty} \rho, \quad P_\rho - \text{a.s.}$

ii) $\forall h > 0, \exists N = N(\rho, h) > 0$ such that $\forall n > N$,

$$P_\rho\{\|\hat{\rho}_{n,[n^s]} - \rho\|_2 > h\} \leq \exp(-Cn^{1-s}h^2)$$

where $0 < s < 1/2$ and the constant $C > 0$ depends only on ρ .

Theorem 1.4 . Let $\rho \in \tilde{\Theta}$, $0 < \beta < 1/2$ and $0 < s < \frac{1-2\beta}{2}$.

If $n^\beta (\sum_{j=[n^s]+1}^{+\infty} \lambda_j^2(\rho))^{1/2} \rightarrow_{n \rightarrow \infty} 0$, then we have :

- i) $n^\beta \|\hat{\rho}_{n,[n^s]} - \rho\|_2 \rightarrow_{n \rightarrow \infty} 0$, P_ρ -a.s.
- ii) $\forall h > 0$, $\exists N = N(\rho, h)$ such that $\forall n > N$,

$$P_\rho\{n^\beta \|\hat{\rho}_{n,[n^s]} - \rho\|_2 > h\} \leq \exp(-Cn^b h^2)$$

where $b = 1 - 2\beta - s$ and the constant $C > 0$ depends only on ρ .

For the second class $\tilde{\Theta}^*$, the following results give the a.s. convergence, exponential bounds and rates for the sieves estimators.

Theorem 1.5 . Let $\beta > 0$, $0 < s < 1/2$ and $m_n = 2[n^s/2]$.

a. Under **A1** and **A2** we have :

- i) $\rho_\theta \in \tilde{\Theta}$.
- ii) $\|\hat{\theta}_{n,m_n(s)} - \theta\|_{L^2} \rightarrow_{n \rightarrow \infty} 0$, P_θ -a.s.
- iii) $\forall h > 0$, $\exists N = N(\theta, h) > 0$ such that $\forall n > N$,

$$P_\theta\{\|\hat{\theta}_{n,m_n(s)} - \theta\|_{L^2} > h\} \leq \exp(-Cn^b h^2)$$

where $b = 1 - s$ the constant $C > 0$ depends only on θ .

b. Under **A0**- **A1** and **A3**, we have

- i) $\rho_\theta \in \tilde{\Theta}^*$.
- ii) $\|\hat{\theta}_{n,m_n(s)}^* - \theta\|_{L^2} \rightarrow_{n \rightarrow \infty} 0$, P_θ -a.s.
- iii) $\forall h > 0$, $\exists N = N(\theta, h) > 0$ such that $\forall n > N$,

$$P_\theta\{\|\hat{\theta}_{n,m_n(s)}^* - \theta\|_{L^2} > h\} \leq \exp(-Cn^b h^2)$$

where $b = 1 - s$ and the constant $C > 0$ depends only on θ .

Theorem 1.6 . Let $0 < \beta < 1/2$ and $0 < s < \frac{1-2\beta}{2}$.

a. Under **A1** and **A2** and if $n^\beta (\sum_{j=m_n(s)+1}^{+\infty} \lambda_j^2(\rho))^{1/2} \rightarrow_{n \rightarrow \infty} 0$, then the results **a.ii)** and **a.iii)** of Theorem 1.5 hold with the rate of convergence n^β and $b = 1 - 2\beta - s$.

b. Under **A0**- **A1** and **A3** if $n^\beta (\sum_{j=m_n(s)+1}^{+\infty} \lambda_j^2(\rho))^{1/2} \rightarrow_{n \rightarrow \infty} 0$, then the results **b.ii)** and **b.iii)** of Theorem 1.5 hold with the rate of convergence n^β and $b = 1 - 2\beta - s$.

Remark . The condition $n^\beta (\sum_{j=[n^s]+1}^{+\infty} \lambda_j^2(\rho))^{1/2} \rightarrow_{n \rightarrow \infty} 0$ of Theorem 4, gives the rate of convergence n^β of the sieves estimators which is determined by the smoothness s of the sieves and the decay rate of the eigenvalues of the parameter operator. Examples satisfying this condition are in section simulations.

The best probabilistic predictor of the random variable X_{n+1} is $\rho(X_n)$. We may define a statistical predictor $\hat{\rho}_{n,[n^s]}(X_n)$ (or $\hat{\rho}_{n,[n^s]}^*(X_n)$) corresponding to the two sieves estimators. We have the following results for the two sieves predictors.

Theorem 1.7 Under the conditions of Theorem 1.3 and Theorem 1.5 we have respectively for the two sieves predictors, under P_ρ :

$$\|\hat{\rho}_{n,[n^s]}(X_n) - \rho(X_n)\|_2 \rightarrow_{n \rightarrow +\infty} 0 \quad \text{and} \quad \|\hat{\rho}_{n,[n^s]}^*(X_n) - \rho(X_n)\|_2 \rightarrow_{n \rightarrow +\infty} 0$$

2 Numerical Simulations

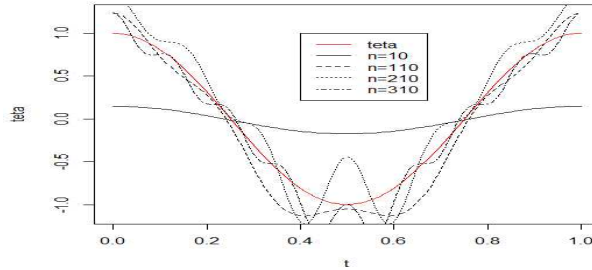
In this section, we carry out numerical simulations to illustrate the behavior of sieves estimators when the operator ρ belongs to the second Class and when $H = L^2[0, 1]$. A strong white noise is generated by its Karhunen-Loeve expansion where $\sigma_k^2 = 1/(k+1)^2, k \geq 0$ and $\{e_k(t), k \geq 0\}$ is the trigonometric base of $L^2([0, 1])$ developed by J. Damon and S. Guillas available at *R-package far*. Then we generate 310 observations of the ARH(1) process, each trajectory is calculated at $m=20$ values of the corresponding interval. To show the behavior of the sieves estimator we divided $[0, 1]$ into 100 parts, we take $s = (1/2) - 0.01$ and the Lagrange multiplier $\mu = 5$. We consider two cases : even kernel and odd kernel.

Example 1. We take a kernel integral operator $\rho_\theta f(s) = \int_0^1 \theta(s-t)f(t)dt$ where θ is an even continuous function, periodic on \mathbb{R} with period 1.

The true kernel is defined on $[0, 1]$ by

$$\theta(t) = \cos(2\pi t)$$

The following graphic shows the oscillations of the Sieves estimator (blue curve) for one realization around the true function θ (red curve) for different sample values n .

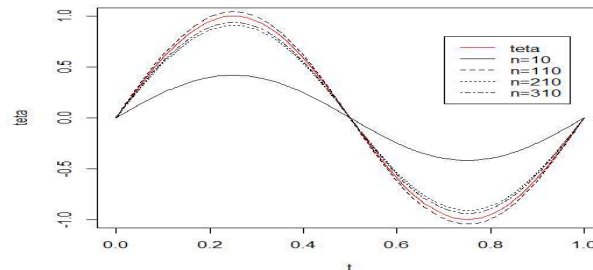


Example 2. We take a kernel integral operator $\rho_\theta f(s) = \int_0^1 \theta(s-t)f(t)dt$ where θ is an odd continuous function, periodic on \mathbb{R} with period 1.

The true kernel is defined on $[0, 1]$ by

$$\theta(t) = \sin(2\pi t)$$

The following graphic shows the oscillations of the Sieves estimator (blue curve) for one realization around the true kernel θ (red curve) for different values of n .



We may observe that the graphics show a good behavior of the sieves estimators in the two cases as n increase.

Références

- [1] D. Bosq. *Linear Processes in Function Spaces*. Springer, New York, 2000.
- [2] F. Ferraty, P. Vieu. *Nonparametric Functional Data Analysis :Theory and Practice*. Springer-Verlag, New York, 2006.
- [3] U. Grenander, *Abstract Inference*. Wiley, New York, 1981.
- [4] J. Hajek, Local asymptotic Minimax and Admissibility in Estimation. *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob (Univ. of Calif. Press)*.1 (1972) 175-194.
- [5] I.A. Ibragimov, R.Z. Khas'Minskii, Asymptotically Normal Families of Distributions and Efficient Estimation. *Ann. Statist.* 19 (1991) 1681-1724.
- [6] G. Kallianpur, R.S. Selukar. *Estimation of Hilbert Space Valued Parameters by the Method of Sieves*, Wiley Eastern Limited, Publishers (1993) 325-347.
- [7] N. Kara Terki, T. Mourid, On Local Asymptotic Normality For Functional Autoregressive Processes, *J. Multivariate Anal*, 148, 120-140 (2016) .
- [8] T. Mourid, N. Bensmain. Sieves estimator of the operator of a functional autoregressive process. *Statist. Probab. Lett.* 76 (2006) 93-108.
- [9] S. Omatu, H. Nagamine, T. Soeda. Optimal filter for a discret-time distributed parameter system and its application to environmental data processing. *In Applications of Information and Control Systems, Vol.III.2nd International. Conf. Inform.Sci. Systems, Patras,Greece. 1979*.
- [10] J.O. Ramsay, B.W. Silverman. *Functional Data Analysis*. Second ed., Springer, New-York, 2005.
- [11] Ju.A. Rozanov. *Infinite-Dimensional Gaussian Distributions*. American Mathematical Society, Providence, Rhode Island, 1971.

Modeling of p -order persistent time series by the modified Langevin equation

Zbigniew Czechowski

Institute of Geophysics, Polish Academy of Sciences, Ks. Janusza 64, 01-452 Warsaw, Poland
zczech@igf.edu.pl

Keywords: nonlinear time series modeling; stochastic processes; Langevin equation; persistent processes; reconstruction procedure.

1 Introduction

Modeling lies at the heart of time series analysis. The registered data contain hidden information about the process under investigation. They reflect complexity of the phenomenon and its important features, e.g.: power law distributions, nonlinear dynamics, multifractal structure or long range correlations. Non-regularity of many natural data induces us to accept the assumption about stochastic basis of time series, so the registered data can be treated as realizations of stochastic processes. In order to describe mathematically these processes different stochastic models were applied and tested. Procedures of reconstruction of a chosen model from time series are necessary for modeling in practice. In the case of linear ARMA models very good reconstruction procedures were elaborated. However, a proper description of many processes requires nonlinear models. The Langevin equation introduces nonlinearity in drift and diffusion terms and leads to a wide class of distributions; from Gaussian to inverse-power. Moreover, following the well-known correspondence of the Langevin and the Fokker-Planck equation, Siegert et al. [1] introduced the procedure of reconstruction of the standard Langevin equation, which was based on numerical estimations of the joint distribution function (i.e., histograms). The method was then developed in other papers (e.g., [2 – 6]). Both the models, linear ARMA and nonlinear Langevin, have their merits and demerits. ARMA models can describe Markov time series of order m , but the linearity is limiting their usefulness. On the other hand, the Langevin model is nonlinear, but it can describe only Markov processes (of order 1).

In this work we introduce the generalized discrete Langevin equation for some class of non-Markov processes, namely for persistent time series of order p . The standard reconstruction procedures fail in this case, therefore, we propose a new method of reconstruction of the generalized Langevin equation from data. This work is a significant extension of our previous approach [7] in which persistent processes of order $p = 1$ were taken into account.

2 The generalized discrete Langevin equation

For persistent processes non-local effects must be considered. We assume that the next state $y(t + \Delta t)$ of the process is dependent not only on the present state $y(t)$ but also on signs s_{tk} of p previous jumps $\Delta y = y(t - (k - 1)\Delta t) - y(t - k\Delta t)$, where $k = 1, 2, \dots, p$. To this aim, the standard discrete Langevin equation is modified by introducing a new random function $c(s_t, r_t; \mathbf{d})$ which determines the sign of the diffusion term, i.e.,

$$y(t + \Delta t) = y(t) + a(y(t))\Delta t + c(s_t, r_t; \mathbf{d})\sqrt{b(y(t))}\sqrt{\Delta t} |\xi_t| \quad (1)$$

The function $c(s_t, r_t; \mathbf{d})$ depends on vector random variable $s_t = [s_{t1}, s_{t2}, \dots, s_{tp}]$, random scalar variable r_t and on vector persistence parameter $\mathbf{d} = [d_1, d_2, \dots, d_{2^p}]$. The function can be equal 1 or -1 randomly, according to rules which describe a complex persistence of time series. For persistent processes of order p the function $c(s_t, r_t; \mathbf{d})$ is keeping the tendency of increase/decrease of $y(t)$ in the next step according to given probabilities $p_i = 1 - d_i$, where $i = 1, 2, \dots, 2^p$. When all $p_i = 1/2$ then Eq. (1) reduces to the standard Langevin equation without the modification.

3 Reconstruction procedure

The standard procedure [1] of reconstruction of the Langevin equation from time series leads to the proper estimation of the diffusion function $b(y)$ but to the wrong reconstruction of the drift function $a(y)$ in the case of generalized equation (1). To estimate the deviation in the drift we propose the modified reconstruction procedure. The algorithm can be summarized in three steps as follows:

Step 1. First reconstruction of $a(y)$.

The first using of the standard procedure to the input persistent time series leads to the first reconstruction $a_1(y)$ of function $a(y)$ and $b_1(y)$ of function $b(y)$:

$$\begin{aligned} a_1(y) &= a(y) + \alpha_1(y), \\ b_1(y) &\equiv b_R(y) \approx b(y), \end{aligned} \quad (2)$$

where α is the deviation.

Step 2. Estimation of parameter \mathbf{d} .

A direct method of estimation of parameters p_i from data is based on histograms P^H of the joint probability $P(s_{t2^p}, \dots, s_{t3}, s_{t2}, s_{t1}, c)$.

Step 3. Final reconstruction of $a(y)$.

Time series generated by the modified Langevin equation (1), with parameter \mathbf{d} estimated in Step 2 and reconstructed functions $a_1(y)$ and $b_R(y)$, is treated as the input to the second use of the standard procedure. At the result, function $a_2(y)$ is reconstructed, where

$$a_2(y) = a_1(y) + \alpha_2(y). \quad (3)$$

Assuming that deviation $|\alpha_1(y)| \ll |a(y)|$ and then $\alpha_1(y) \approx \alpha_2(y)$ we obtain

$$\begin{aligned} a(y) &= a_1(y) - \alpha_1(y) \approx a_1(y) - \alpha_2(y) = a_1(y) - [a_2(y) - a_1(y)] \\ &= 2a_1(y) - a_2(y) \equiv a_R(y). \end{aligned} \quad (4)$$

The reconstructed modified Langevin equation has a form of Eq. (1) with $a(y) = a_R(y)$, $b(y) = b_R(y)$ and parameter d estimated in Step 2.

It should be underlined that for the case $p = 1$ the following correct analytical formula was derived for the deviation in reconstruction of the drift function [7]:

$$\alpha_1 = -\frac{1-2d}{\pi} [2a_1(y) - b_1'(y)] \quad (5)$$

The expression depends on the estimated persistence parameter d and on first reconstructions $a_1(y)$ and $b_1(y)$ given by the standard procedure.

4 Testing of the modified reconstruction procedure

In order to test an efficiency of the procedure we generate time series by using the modified Langevin equation (1) with different functions $a(y)$ and $b(y)$, different values of the parameter d , different time increments Δt and considering different time series lengths N . This enables to compare the input parameters and functions to the reconstructed ones.

Figure 1 and 2 show how the procedure is working for the case: $a(y) = -(1 - 2y)/2$, $b(y) = y^2$, $d_1 = d_2 = 0.65$, $d_3 = d_4 = 0.60$, $d_5 = d_6 = 0.55$, $d_7 = d_8 = 0.5$, $N = 1000000$, $\Delta t = 0.01$. The Step 1 of the procedure leads to first reconstructions, $a_1(y)$ and $b_1(y)$. They are represented by least-square fits (long-dashed lines) to a cloud of points in Fig. 1 and Fig. 2, respectively. We note that $b_1(y)$ is a good estimation of the input diffusion function, it coincides with the input $b(y) = y^2$. In Step 2 we estimate the parameters d and we find: $d_1 = d_2 = 0.646$, $d_3 = d_4 = 0.599$, $d_5 = d_6 = 0.547$, $d_7 = d_8 = 0.499$. Then, according to Step 3, we generate a new time series by using the modified Langevin equation (1) with reconstructed functions $a_1(y)$ and $b_R(y) = b_1(y)$ and estimated parameter d . We apply again the standard procedure to obtain $a_2(y)$ (represented by the dashed line fitted to a cloud of '+' signs, see Fig. 1). Relation (4) leads to the final reconstruction $a_R(y)$ (continuous line in Fig. 1). We can note that the reconstruction is very close to the input $a(y)$ (thick continuous line).

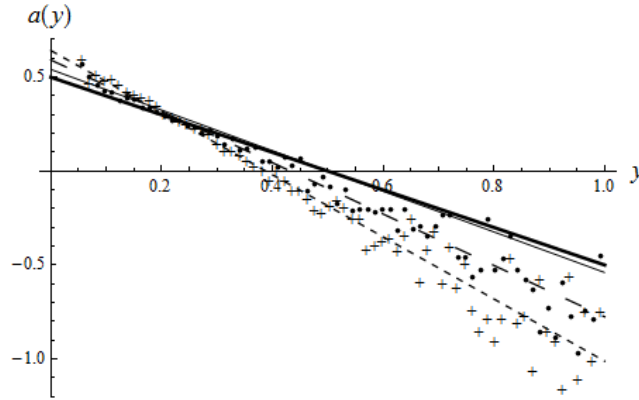


Fig. 1. Reconstruction of the drift function. The final reconstruction $a_R(y)$ (continuous line) is close to the input function (thick continuous line). Long-dashed line and dashed line represent intermediate reconstructions of drift function, $a_1(y)$ and $a_2(y)$, respectively.

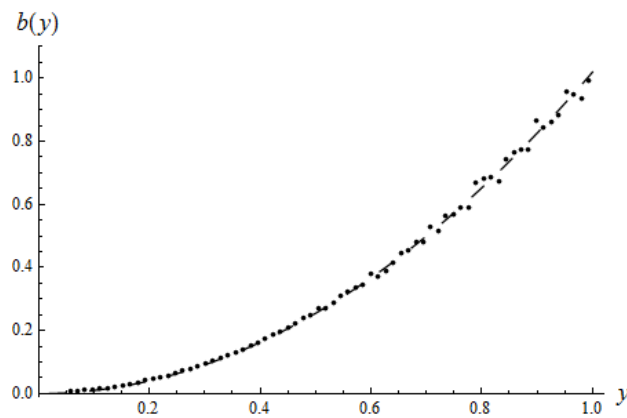


Fig. 2. Reconstruction of the diffusion function. The final reconstruction $b_R(y) = b_1(y)$ (long-dashed line) fits very well to the input function $b(y) = y^2$.

Acknowledgements

This work has been financed by the project of the National Science Centre (contract No. 2016/21/B/ST10/02998).

1. Siegert S., R. Friedrich, and J. Peinke, 1998, Analysis of data sets of stochastic systems, Phys. Lett. A **243**, 275-280.
2. Anteneodo C. and R. Riera, 2009, Arbitrary-order corrections for finite-time drift and diffusion coefficients, Phys. Rev. E **80**, 031103.
3. Gottschall J. and J. Peinke, 2008, On the definition and handling of different drift and diffusion estimates, New J. Phys. **10**, 083034.

4. Hindriks R., Jansen R., Bijama F., Mansvelder H. D., de Gunst M. C. M., van der Vaart A. W., 2011. Unbiased estimation of Langevin dynamics from time series with application to hippocampal field potentials *in vitro*, Phys. Rev. E **84**, 021133-1-13.
5. Kleinhaus D., Friedrich R., 2007. Maximum likelihood estimation of drift and diffusion functions, Phys. Lett. A **368**, 194-198.
6. Lamaurox D. and K. Lehnertz, 2009, Kernel-based regression of drift and diffusion coefficients of stochastic processes, Phys. Lett. A **373**, 3507-3512.
7. Czechowski Z., 2016, Reconstruction of the modified discrete Langevin equation from persistent time series, CHAOS **26**, 053109.

Bootstrap confidence intervals for conditional density function in Markov processes

Inés Barbeito, Ricardo Cao, and Dimitris Politis

University of A Coruña, University of California, San Diego
`{ines.barbeito,rcao}@udc.es`
`politis@math.ucsd.edu`

Abstract. Several bootstrap algorithms for prediction intervals have been considered to construct confidence intervals for the conditional density function in Markov processes. These methods are: Model-Free Bootstrap, Predictive Model-Free, Limit Model-Free, Nonparametric Autoregression with fitted and predictive residuals, Local Bootstrap and Bootstrap based on estimates of the transition density (see [3], [5], [4], [6], [7] and [8], respectively). However, to achieve good coverage of confidence intervals, it is of utmost importance to use an appropriate bandwidth selector to estimate the conditional density function. In this sense, rule of thumb and cross validation smoothing parameters have been considered. Furthermore, smoothed stationary bootstrap (see [1]) has been used to obtain a new bootstrap bandwidth selector for conditional density estimation via working out a closed expression for the mean integrated squared error. This is very useful since Monte Carlo approximation is no longer needed. Finally, a simulation study has been carried out to compare empirically the performance of these methods considering the aforementioned smoothing parameters.

Keywords: conditional density, kernel method, model-free bootstrap, confidence interval, mean integrated squared error, smoothed stationary bootstrap, markov processes

1 Introduction and bootstrap confidence intervals

Bootstrap methods for time series have been extensively studied during the last three decades. To a deeper insight on the topic, a recent review of the state of the art of literature was given by Kreiss and Paparoditis (see [2]).

In particular, our aim is to construct confidence intervals for the conditional density function in the setting of Markov processes. In this context, two well-known methods have been already proposed:

1. The bootstrap method based on kernel estimates of the transition density of the Markov processes, proposed by Rajarshi (see [8]).

2. The Local Bootstrap for Markov processes of Paparoditis and Politis (see [6], [7]).

Moreover, in terms of establishing prediction intervals, Pan and Politis (see [3], [5]) proposed several bootstrap algorithms: the model-free bootstrap for Markov processes, the predictive model-free bootstrap and the limit model-free bootstrap. Furthermore, the model-free bootstrap algorithm has been discussed by Politis (see [5]) to construct confidence intervals for the conditional expectation function as well.

In addition, considering a nonparametric autoregression model, Pan and Politis (see [4]) proposed two bootstrap methods fitting the model via kernel smoothing, assuming iid errors:

1. Forward bootstrap with fitted residuals.
2. Forward bootstrap with predictive residuals.

In this work, the aforementioned bootstrap algorithms are empirically compared when constructing confidence intervals for the conditional density function for Markov processes.

2 Bandwidth selection for nonparametric conditional density estimation

Consider a probability density K on \mathbb{R}^2 , a 2-dimensional random sample coming from a Markov process of order 1, $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, and two positive bandwidths h_1, h_2 to construct the kernel conditional density estimator, given by

$$\hat{f}(y|x) = \frac{\hat{f}_{h_1, h_2}(x, y)}{\hat{f}_{h_1}(x)}, \quad (1)$$

where $\hat{f}_{h_1, h_2}(x, y) = \frac{1}{(n-1)h_1h_2} \sum_{i=2}^n K\left(\frac{x-X_{i-1}}{h_1}, \frac{y-Y_i}{h_2}\right)$, $x, y \in \mathbb{R}$ and $\hat{f}_h(x) = \int \hat{f}_{h_1, h_2}(x, y) dy$.

As can be seen in (1), the estimator strongly depends on the choice of the smoothing parameters h_1, h_2 . Consequently, different bandwidths selectors have been considered when carrying out the simulation study. Rule of thumb and cross validation procedures have been considered.

On the other hand, let us consider a 2×2 , symmetric and positive-definite matrix of bandwidths, H , given by

$$H = \begin{pmatrix} h_1^2 & h_1 h_2 \\ h_2 h_1 & h_2^2 \end{pmatrix},$$

and the 2-dimensional kernel joint density estimator:

$$\hat{f}_H(x, y) = \frac{1}{n} \sum_{i=1}^n K_H(x - X_i, y - Y_i),$$

being $K_H(x, y) = |H|^{-1/2} K(H^{-1/2}(x, y)^T)$. Thus, in this case, the kernel conditional density estimator is given by:

$$\hat{f}(y|x) = \frac{\hat{f}_H(x, y)}{\hat{f}_{h_1}(x)}. \quad (2)$$

Considering an approximation of the estimator given in (2), our goal is to work out a closed expression for the smoothed stationary bootstrap, namely SSB, (see [1]) version of the mean integrated squared error, so as to obtain a bootstrap bandwidth selector by minimizing it. As a matter of fact, this smoothing parameter is really useful since Monte Carlo approximation is not needed to compute it. The new bootstrap smoothing parameter is also used in the simulation study.

Acknowledgments. The first two authors acknowledge partial support by the MINECO grant MTM2014-52876-R (ERDF support included). Additionally, financial support from the Xunta de Galicia (Centro Singular de Investigación de Galicia accreditation ED431G/01 2016-2019 and Grupos de Referencia Competitiva ED431C2016-015) and the European Union (ERDF), is gratefully acknowledged. The work of the first author has been carried out during a visit at the University of California, San Diego, financed by INDITEX, with reference INDITEX-UDC 2017.

References

1. Barbeito, I., Cao, R.: Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data. *Comput. Statist. & Data Anal.* 104, 130–147. (2016)
2. Kreiss J.P., Paparoditis E.: Bootstrap methods for dependent data: a review. *J. Korean Stat. Soc.* 40(4), 357–378. (2011)
3. Pan, L., Politis, D.N.: Bootstrap prediction intervals for Markov processes. *Comp. Statist. & Data Anal.* 100, 467–494. (2016)
4. Pan, L., Politis, D.N.: Bootstrap prediction intervals for linear, nonlinear, and non-parametric autoregressions. *J. Statist. Plan. Infer.* 177, 1–27. (2016)
5. Politis, D. N.: Model-free prediction and regression: a transformation-based approach to inference. Springer. (2015)
6. Paparoditis, E., Politis, D.N.: The local bootstrap for Markov processes. *J. Statist. Plan. Infer.* 108(1-2), 301–328. (2001)

7. Paparoditis, E., Politis, D.N.: A Markovian local resampling scheme for nonparametric estimators in time series analysis. *Econometr. Theor.* 17, 540–566. (2002)
8. Rajarshi, M.B.: Bootstrap in Markov-sequences based on estimates of transition density. *Ann. Inst. Stat. Math.* 42(2), 253–268. (1990)

Forecasting with Functional Time Series

Sara Leulmi and Fatiha Messaci

Laboratoire LAMASD, département de Mathématiques
Université frères Mentouri, route d'Ain EL Bey
25017, Constantine, Algeria

E-mail: math17sara@yahoo.fr and f_messaci@yahoo.fr

Abstract. We introduce a conditional median estimator of a scalar response given a random variable taking values in a semi metric space. We establish its strong consistency, with rate, when the sample is an α -mixing sequence. Then, a real data set study illustrates the performance of our methodology with respect to other known estimators.

Keywords: Functional data, Local linear estimation, Strong consistency, α -mixing.

1 Introduction

Since the pioneer works in [5], several studies dealt with the nonparametric functional estimation. This research field is motivated by the fact that several data collected in practice, are given in the form of curves and that the progress of the digital computing tools allows the treatment of such observations.

In the previous reference, only the kernel method has been considered. Later, the local linear method has been extended to the functional framework, for the first time, in [2]. Then, other local linear nonparametric estimators has been investigated in some papers as [4] and [7].

Moreover, observed data can exhibit a dependence form. A large studied example in Time Series is the case of the α -mixing dependence. We cite [1] and [6] for papers dealing with such functional dependent data.

This work takes place within this field. We establish the almost complete convergence (stronger than the almost sure one) of a local linear nonparametric estimator of the conditional distribution function of a scalar response variable given a random variable taking values in a semi metric space (the functional variable) when the collected observations are α -mixing. Then, we derive the consistency of a conditional median estimator which is a prediction tool. Finally, a real data study shows that our estimator performs well with respect to other known conditional median estimators.

2 Estimation and hypotheses

Let us consider n pairs of random variables $(X_i, Y_i)_{i=1, \dots, n}$ identically distributed as the pair (X, Y) which is valued in $\mathcal{F} \times \mathbb{R}$, where \mathcal{F} is a infinite-dimensional

space equipped with a semi-metric d .

We first estimate the conditional cumulative distribution function $F^x(y) = P(Y \leq y \mid X = x)$, from which we derive a new tool in order to make forecasting.

Following [2], we propose a local linear estimate $\hat{F}^x(y)$ of $F^x(y)$ given by

$$\hat{F}^x(y) = \frac{\sum_{i,j=1}^n W_{ij}(x) 1_{]-\infty, y]}(Y_j)}{\sum_{i,j=1}^n W_{ij}(x)} \quad \left(\frac{0}{0} := 0 \right), \quad (1)$$

with

$$W_{ij}(x) = \beta(X_i, x) (\beta(X_i, x) - \beta(X_j, x)) K(h^{-1}d(X_i, x)) K(h^{-1}d(X_j, x)),$$

where $\beta(\cdot, \cdot)$ is a known operator from $\mathcal{F} \times \mathcal{F}$ into \mathbb{R} such that, $\forall x \in \mathcal{F}$, $\beta(x, x) = 0$, the function K is a kernel and $h := h_n$ is a sequence of strictly positive real numbers which plays a smoothing parameter role.

Remark that a double kernel local linear estimator is introduced in [7] and studied for independent data.

As the conditional quantile of order α ($\alpha \in (0, 1)$) is $t_\alpha(x) = \inf\{y \in \mathbb{R}, F^x(y) \geq \alpha\}$, we deduce from \hat{F}^x a natural conditional quantile estimator as,

$$\hat{t}_\alpha(x) = \inf\{y \in \mathbb{R}, \hat{F}^x(y) \geq \alpha\}. \quad (2)$$

Recall that $t_{1/2}(x)$ is the so called conditional median.

For easy reference, we recall the following definitions.

Definition 1 Let $\{Z_i, i = 1, 2, \dots\}$ be a strictly stationary sequence of random variables, $F_i^k(Z)$ denotes the σ -algebra generated by $\{Z_j, i \leq j \leq k\}$. Given a positive integer n , set

$$\alpha(n) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in F_1^k(Z) \text{ and } B \in F_{k+n}^\infty(Z), k \in \mathbb{N}^*\}.$$

The sequence is said to be α -mixing (strongly mixing) if the mixing coefficient $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty$.

Many processes do satisfy the strong mixing property, see for example [3] for more details.

Definition 2 Let $(Z_n)_{n \in \mathbb{N}^*}$ be a sequence of real random variables (r.r.v.). We say that $(Z_n)_{n \in \mathbb{N}^*}$ converges almost completely to some r.r.v. Z , and we note $Z_n \xrightarrow{a.co.} Z$, if and only if

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} P(|Z_n - Z| > \varepsilon) < \infty.$$

Moreover, let $(u_n)_{n \in \mathbb{N}^*}$ be a sequence of positive real numbers going to zero; we say that the rate of the almost complete convergence of $(Z_n)_{n \in \mathbb{N}^*}$ to Z is of order (u_n) and we note $Z_n - Z = O_{a.co.}(u_n)$, if and only if

$$\exists \varepsilon_0 > 0, \sum_{n=1}^{\infty} P(|Z_n - Z| > \varepsilon_0 u_n) < \infty.$$

It is clear, from Borel Cantelli lemma, that this convergence is stronger than the almost-sure one.

Let x be a fixed point in \mathcal{F} . For any positive real h , $B(x, h) := \{y \in \mathcal{F} / d(x, y) \leq h\}$ is a closed ball in \mathcal{F} of center x and radius h . We also denote $\Phi_x(r_1, r_2) := P(r_1 \leq d(X, x) \leq r_2)$, where r_1 and r_2 are two real numbers and \mathcal{N}_x stands for a neighbourhood of x .

To study the asymptotic behaviour of the local linear estimator \hat{F}^x , we need the following assumptions.

- (H1) For any $h > 0$, $\Phi_x(h) := \Phi_x(0, h) > 0$.
- (H2) There exist $\delta > 0$, $C' > 0$, $b > 0$ such that: $\forall x' \in \mathcal{N}_x$, $\forall y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]$, $|F^x(y) - F^{x'}(y)| \leq C(d^b(x, x'))$.
- (H3) The function $\beta(., .)$ is such that: $\exists 0 < M_1 < M_2, \forall x' \in \mathcal{F}$,

$$M_1 d(x, x') \leq |\beta(x, x')| \leq M_2 d(x, x').$$

- (H4) The kernel K is a positive and differentiable function on its support $[0, 1]$ and $\exists C, C'$ such that

$$0 < C1_{[0,1]}(t) \leq K(t) \leq C'1_{[0,1]}(t) < \infty.$$

- (H5) The sequence (X_i, Y_i) is a stationary α -mixing sequence with coefficient $\alpha(n)$, moreover (H5a) and (H5b) are satisfied, where
 - (H5a): $\exists C > 0, \exists a > 3, \forall n \in \mathbb{N}; \alpha(n) \leq Cn^{-a}$,
 - (H5b): $\exists C, C' > 0$ such that: $C' [\Phi_x(h)]^{a/(a-1)} < \psi_x(h) \leq C [\Phi_x(h)]^{a/(a-1)}$, with $\psi_x(h) := \psi_x(0, h)$ and $\psi_x(h_1, h_2) := P(h_1 \leq d(X_1, x) \leq h_2, 0 \leq d(X_2, x) \leq h_2)$.
- (H6) The bandwidth h satisfies

$$\exists n_0 \in \mathbb{N}, \forall n > n_0, \frac{1}{\psi_x(h)} \int_0^1 \psi_x(zh, h) \frac{d}{dz} (z^2 K(z)) dz > C > 0$$

and

$$\begin{aligned} & h^2 \int_{B(x, h)} \int_{B(x, h)} \beta(u, x) \beta(t, x) dP_{(X_1, X_2)}(u, t) \\ &= o \left(\int_{B(x, h)} \int_{B(x, h)} \beta^2(u, x) \beta^2(t, x) dP_{(X_1, X_2)}(u, t) \right), \end{aligned}$$

where $dP_{(X_1, X_2)}$ is the joint distribution of (X_1, X_2) .

- (H7) $\lim_{n \rightarrow \infty} h = 0$ and $\exists 0 < \eta_0 < \frac{a-3}{a+1}, \exists C_1 > 0$ such that $C_1 n^{\frac{3-a}{a+1} + \eta_0} \leq \Phi_x(h)$.

Hypotheses (H1) and (H3) have been used in the independent case in [2]. (H2) is a standard regularity condition allowing to deal with the bias. (H4) is a technical condition. (H5a) means that (X_i, Y_i) is arithmetically mixing and is extensively used in the literature as in [5] and in [6]. (H6) is of the same kind as (H6) together with (H7) in [2]. The choice of bandwidth is given by (H7), in particular it implies that $\ln n / n\Phi_x(h) \rightarrow 0$ as $n \rightarrow \infty$.

3 Results

Our first result concerns the asymptotic behaviour of $\hat{F}^x(y)$.

Proposition 1 *Under assumptions (H1)–(H7), we have*

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |\hat{F}^x(y) - F^x(y)| = O(h^b) + O_{a.co.} \left(\sqrt{\frac{\ln n}{n\Phi(h)}} \right).$$

It is easy to see that the proof of Proposition 1 is a direct consequence of the standard decomposition given, for all x by

$$\hat{F}^x(y) - F^x(y) = \frac{1}{\hat{F}_D^x} \left[\left(\hat{F}_N^x(y) - E\hat{F}_N^x(y) \right) - \left(F^x(y) - E\hat{F}_N^x(y) \right) \right] - \frac{F^x(y)}{\hat{F}_D^x} (\hat{F}_D^x - 1) \quad (3)$$

where,

$$\hat{F}_N^x(y) = \frac{1}{n(n-1)EW_{12}(x)} \sum_{i \neq j} W_{ij}(x) 1_{\{Y_j \leq y\}}, \quad \hat{F}_D^x = \frac{1}{n(n-1)EW_{12}(x)} \sum_{i \neq j} W_{ij}(x),$$

and of the following lemmas whose proofs are relegated to the Appendix.

Lemma 1 *Assume that hypotheses (H1)–(H6) hold, then*

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} \left| F^x(y) - E\hat{F}_N^x(y) \right| = O(h^b).$$

Lemma 2 *Under assumptions of Theorem 1, we obtain that*

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} \left| \hat{F}_N^x(y) - E\hat{F}_N^x(y) \right| = O_{a.co.} \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right).$$

Lemma 3 *If assumptions (H1), (H3)–(H7) are satisfied, we get*

$$\left| \hat{F}_D^x - 1 \right| = O_{a.co.} \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right) \text{ and } \sum_{n=1}^{\infty} P \left(\hat{F}_D^x < \frac{1}{2} \right) < \infty.$$

To obtain the consistency of the conditional quantile estimator, we add the following assumption.

(H8): F^x is differentiable with a continuous density f^x satisfying $f^x(t_\alpha(x)) > 0$. A known method can be applied to derive the following result from Proposition 1, see for example the proof of Theorem 3.1 in [6].

Theorem 1 *Under the hypotheses of Proposition 1 and if (H8) is satisfied, we obtain*

$$|\hat{t}_\alpha(x) - t_\alpha(x)| = O(h^b) + O_{a.co.} \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right).$$

4 Real data application

In this section, a real data set will permit us to illustrate the efficacy of our studied estimator $\hat{t}_{1/2}$ with respect to other conditional median estimators: The kernel one (denoted KM) studied in [5] and the local linear estimator (denoted LLM) introduced in [7].

The KM (resp. LLM) estimator is computed with the same parameters as at subsection 12.4 in [5] (resp. at section 4 in [7]). For the computation of the estimator $\hat{t}_{1/2}$, we use the kernel $K(x) = [\frac{3}{2}(1-x^2) + 0,001] 1_{[0,1]}(x)$ (close to the quadratic kernel), the bandwidth h is chosen by the cross-validation method and the semimetric d is the PCA one described in [5] (see routines "semimetric.pca" in the website <http://www.lsp.ups-tlse.fr/staph/npfda> with $q = 4$) and $\beta := d$. Our aim is to study the US monthly electricity consumption observed during 338 months (from January 1973 up to February 2001) which can be found at <http://www.econmagic.com>. As pointed out in [5], this time series can be viewed as dependent functional data.

The consumption of a year is the explanatory variable and the consumption of each month of the following year is the response one. We eliminate the 337 and 338 months and we retain the remaining 28 years.

Fix $s \in \{1, 2, \dots, 12\}$, in order to predict the electricity consumption of the s^{th} month of the last year (the 28^{th}) by each cited method, we use the 27 first years to define the training sample $(X_i, Y_i^s)_{(i=1, \dots, 26)}$ used to build the estimators under investigation, where X_i stands for the consumption of the whole i^{th} year and Y_i^s is the consumption of the s^{th} month of the $(i+1)^{th}$ year. Then, for all $s \in \{1, 2, \dots, 12\}$, we predict Y_{27}^s , which is the consumption of the s^{th} month of the 28^{th} year, given X_{27} .

The criteria allowing us to compare between the three estimators is the empirical Mean Square Error (MSE), defined by

$$MSE := \frac{1}{12} \sum_{i=1}^{12} \left(Y_i - \hat{Y}_i \right)^2,$$

where Y_i (resp. \hat{Y}_i) is the real (resp. the estimated) value of the i^{th} month of the last year.

The obtained results are:

$MSE(\hat{t}_{1/2})=0.00235$, $MSE(LLM)=0.00333$ and $MSE(KM)=0.00253$.

Based on this data set, we see that our estimator provides an acceptable performance.

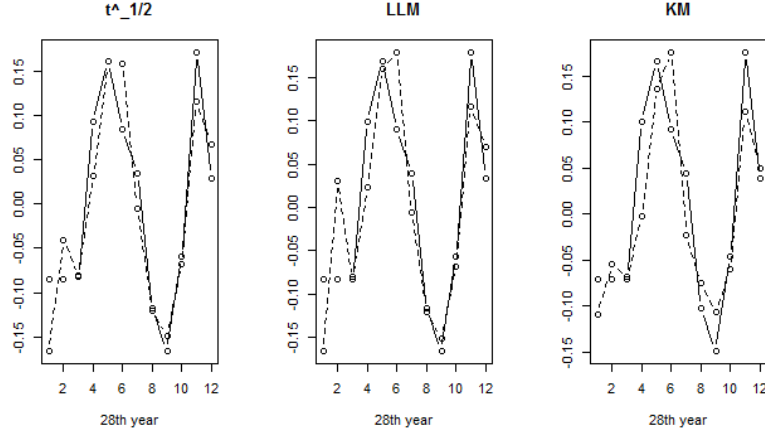


Fig. 1. Performance of the three methods for the Electricity data.

In Figure 1 and for each mentioned method, the dotted (resp. solid) lines stand for the true (resp. forecasted) values.

5 Appendix

In what follows, let C be some strictly positive generic constant and for any $x \in \mathcal{F}$, and for all $i = 1, \dots, n$, we set

$$K_i(x) := K(h^{-1}d(X_i, x)) \text{ and } \beta_i(x) := \beta(X_i, x).$$

To treat the almost-complete convergence of $\hat{F}^x(y)$, we need the following preliminary technical lemma.

Lemma 4 *Under assumptions (H1), (H3), (H4), (H5b) and (H6), we obtain*

- i) $\forall (p, l) \in \mathbb{N}^* \times \mathbb{N}$, $E(K_1^p(x)|\beta_1^l(x)) \leq Ch^l\Phi_x(h)$.
- ii) $\forall (p_1, p_2, l_1, l_2) \in \mathbb{N}^* \times \mathbb{N}^* \times \mathbb{N} \times \mathbb{N}$,
 $E[K_1^{p_1}(x)K_2^{p_2}(x)|\beta_1^{l_1}(x)|\beta_2^{l_2}(x)] \leq Ch^{(l_1+l_2)}[\Phi_x(h)]^{a/(a-1)}$.
- iii) $E[K_1(x)K_2(x)\beta_1^2(x)] > Ch^2[\Phi_x(h)]^{a/(a-1)}$ for n sufficiently large.

Proof 1 i) (see Lemma A.1-i in [2]).

ii) In view of hypotheses (H3) and (H4), we get

$$\begin{aligned} & E\left(K_1^{p_1}(x)K_2^{p_2}(x)|\beta_1^{l_1}(x)|\beta_2^{l_2}(x)\right) \\ & \leq Ch^{(l_1+l_2)}E\left[1_{[0,1]}(h^{-1}d(X_1, x))1_{[0,1]}(h^{-1}d(X_2, x))\right] \\ & \leq Ch^{(l_1+l_2)}P[(X_1, X_2) \in B(x, h) \times B(x, h)]. \end{aligned}$$

So, we derive the claimed result by using (H5b).

iii) Applying (H3), it is easy to see that

$$E [K_1(x)K_2(x)\beta_1^2(x)] > CE [K_1(x)d^2(X_1, x)K_2(x)].$$

Combining hypothesis (H4) with Fubini's theorem, we obtain

$$\begin{aligned} E [K_1(x)d^2(X_1, x)K_2(x)] &= h^2 \int_0^1 \int_0^1 t^2 K(t)K(u) dP_{(h^{-1}d(X_1, x), h^{-1}d(X_2, x))}(t, u) \\ &> Ch^2 \int_0^1 \left(\int_0^1 \int_0^1 1_{[z, 1]}(t) dP_{(h^{-1}d(X_1, x), h^{-1}d(X_2, x))}(t, u) \right) \frac{d}{dz} (z^2 K(z)) dz. \end{aligned}$$

Moreover, we have

$$\int_0^1 \int_0^1 1_{[z, 1]}(t) dP_{(h^{-1}d(X_1, x), h^{-1}d(X_2, x))}(t, u) = P(zh \leq d(X_1, x) \leq h, 0 \leq d(X_2, x) \leq h) = \psi_x(zh, h).$$

Finally (H6) permits us to end the proof.

As the dependence assumption reveals covariance terms, let us define for $p \in \{2, 3, 4\}$ and $l \in \{0, 1\}$

$$(S^x)_{n,l,p}^2(y) = \sum_{i=1}^n \sum_{j=1}^n |Cov(\Gamma_{i,p,l}^x(y), \Gamma_{j,p,l}^x(y))|, \quad (4)$$

where, for $i \in \{1, \dots, n\}$

$$\Gamma_{i,p,l}^x(y) = \frac{1}{h^{p-2}} \left\{ K_i(x) \beta_i^{p-2}(x) 1_{\{Y_i \leq y\}}^l - E[K_i(x) \beta_i^{p-2}(x) 1_{\{Y_i \leq y\}}^l] \right\}. \quad (5)$$

Following the same lines as for proving relation (6.9) in [6], along with the application of Lemma 4 i) and ii), we get for all y

$$(S^x)_{n,l,k}^2(y) = O(n\Phi_x(h)). \quad (6)$$

Proof of Lemma 1 We have

$$E\hat{F}_N^x(y) = \frac{1}{EW_{12}(x)} E[W_{12}(x) 1_{\{Y_2 \leq y\}}]$$

and $E\hat{F}_N^x(y)$ can also be written as

$$E\hat{F}_N^x(y) = E[E(\hat{F}_N^x(y)|X_2)] = \frac{1}{EW_{12}(x)} E[W_{12}(x) E(1_{\{Y_2 \leq y\}}|X_2)].$$

So, we get under assumption (U4)

$$\left| F^x(y) - E\hat{F}_N^x(y) \right| = \frac{1}{|EW_{12}(x)|} \left| E\{W_{12}(x) [F^x(y) - F^{X_2}(y)]\} \right| \leq \sup_{x|t \in B(x, h)} \left| F^x(y) - F^{x'}(y) \right|.$$

It suffices to take into account hypothesis (U2) to obtain the result.

Proof of Lemma 2 Inspired by the decomposition given in the proof of Lemma 4.4 in [2], we set

$$\hat{F}_N^x(y) = Q(x) [S_{2,1}^x(y)S_{4,0}^x(y) - S_{3,1}^x(y)S_{3,0}^x(y)],$$

where

$$S_{p,l}^x(y) = \frac{1}{n\Phi_x(h)} \sum_{i=1}^n \frac{K_i(x)\beta_i^{p-2}(x)1_{\{Y_j \leq y\}}^l}{h^{p-2}}$$

and

$$Q(x) = \frac{n^2 h^2 \Phi_x^2(h)}{n(n-1)EW_{12}(x)}.$$

So, it suffices to show that, that for $p \in \{2, 3, 4\}$ and $l \in \{0, 1\}$, we have

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |ES_{p,l}^x(y)| = O(1) \text{ and } Q(x) = O(1),$$

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |S_{p,l}^x(y) - ES_{p,l}^x(y)| = O_{a.co.} \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right),$$

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |Cov[S_{2,1}^x(y), S_{4,0}^x(y)]| = O \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right)$$

$$\text{and } \sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |Cov[S_{3,1}^x(y), S_{3,0}^x(y)]| = O \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right).$$

- Applying Lemma 4 i), we readily obtain

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |ES_p^x(y)| = O(1). \quad (7)$$

- Treatment of the term $Q(x)$

On one hand, we have

$$h^2 E[\beta_1(x)\beta_2(x)K_1(x)K_2(x)] \leq Ch^2 \int_{B(x,h)} \int_{B(x,h)} \beta(u,x)\beta(t,x)dP_{(X_1,X_2)}(u,t).$$

On the other hand and in view of (H3) and (H5b), we obtain

$$E[\beta_1(x)\beta_2(x)K_1(x)K_2(x)] = o\left(h^2 [\Phi_x(h)]^{a/(a-1)}\right).$$

Now, Lemma 4-(iii) and the last result allow to write, for n sufficiently large

$$Q(x) = \frac{n^2 h^2 \Phi_x^2(h)}{n(n-1)EW_{12}(x)} \leq C \frac{[\Phi_x(h)]^2}{[\Phi_x(h)]^{a/(a-1)}} \leq C.$$

• Treatment of the term $\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |S_{p,l}^x(y) - ES_{p,l}^x(y)|$, for $p \in \{2, 3, 4\}$ and $l \in \{0, 1\}$.

We have for any $y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]$,

$$S_{p,l}^x(y) - ES_{p,l}^x(y) = \frac{1}{n\Phi_x(h)} \sum_{i=1}^n \Gamma_{i,p,l}^x(y),$$

where $\Gamma_{i,p,l}^x(y)$ is defined in relation (5).

By applying Proposition A.11-ii in [5], we get for any $\varepsilon > 0$, $r \geq 1$ and for some $0 < C < \infty$

$$P(|S_{p,l}^x(y) - ES_{p,l}^x(y)| > \varepsilon) \leq P\left(\left|\sum_{i=1}^n \Gamma_{i,p,l}^x(y)\right| > n\varepsilon\Phi_x(h)\right) \leq C[A_1(x) + A_2(x)], \quad (8)$$

where

$$A_1(x) = \left(1 + \frac{\varepsilon^2 n^2 [\Phi_x(h)]^2}{r(S_{n,l,k}^x(y))^2}\right)^{-r/2} \quad \text{and} \quad A_2(x) = nr^{-1} \left(\frac{r}{\varepsilon n\Phi_x(h)}\right)^{a+1}.$$

Now, taking for $\eta > 0$

$$\varepsilon = \eta \sqrt{\frac{\ln n}{n\Phi_x(h)}} \quad \text{and} \quad r = (\ln n)^2,$$

we obtain

$$A_2(x) \leq Cn^{1-(a+1)/2} (\ln n)^{2a - \frac{(a+1)}{2}} [\Phi_x(h)]^{-(a+1)/2},$$

and using (H7), one gets

$$A_2(x) \leq Cn^{-1-\eta_0(a+1)/2} (\ln n)^{2a - \frac{(a+1)}{2}}. \quad (9)$$

Moreover, in view of equation (6) and the fact that $\ln(x+1) = x - x^2/2 + o(x^2/2)$ where x tends to zero, we can write

$$A_1(x) \leq Cn^{-\eta^2/2}, \quad (10)$$

which shows that $A_1(x)$ is the general term of a convergent series for an appropriate choice of η .

Hence, by combining relations (8), (9) and (10), we derive

$$|S_{p,l}^x(y) - ES_{p,l}^x(y)| = O_{a.co.} \left(\sqrt{\frac{\ln n}{n\Phi_x(h)}} \right).$$

From this last result, it is easy to obtain the uniformity on the compact $[t_\alpha(x) - \delta, t_\alpha(x) + \delta]$. We omit the details because they are well known, we can see for

instance the second part of the proof of Lemme 2.4 in [7].

- Finally, by following similar arguments used to prove (6), we obtain

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |Cov[S_{2,1}^x(y), S_{4,0}^x(y)]| = O\left(\frac{1}{n\Phi_x(h)}\right)$$

and

$$\sup_{y \in [t_\alpha(x) - \delta, t_\alpha(x) + \delta]} |Cov[S_{3,1}^x(y), S_{3,0}^x(y)]| = O\left(\frac{1}{n\Phi_x(h)}\right).$$

In view of (H7), this last rate is negligible with respect to $O\left(\sqrt{\frac{\ln n}{n\Phi_x(h)}}\right)$.

Proof of Lemma 3 The first part of the claimed results can be directly deduced from the proof of Lemma 2 by taking $l = 0$ in all its proof and this easily yields to the second part.

References

1. Attaoui, S.: On the Nonparametric Conditional Density and Mode Estimates in the Single Functional Index Model with Strongly Mixing Data. *Sankhya: The Indian Journal of Statistics*. 76-A(Part 2), 356–378 (2014)
2. Barrientos-Marin, J., Ferraty, F., Vieu, P.: Locally modelled regression and functional data. *Journal of Nonparametric Statistics* 22, 617–632 (2010)
3. Bradley, R.C.: Introduction to strong mixing conditions. Vol I-III. Kendrick Press, Utah (2007)
4. Demongeot, J., Laksaci, A., Madani, F., Rachdi, M.: Functional data analysis: conditional density estimation and its application. *Statistics*. 47, No. 1, 26-44 (2013)
5. Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Theory and Practice. Springer Series in Statistics, New York (2006).
6. Laksaci, A., Lemdani, M., Ould Said, E.: Asymptotic Results for an L1-norm Kernel Estimator of the Conditional Quantile for Functional Dependent Data with Application to Climatology. *Sankhya : The Indian Journal of Statistics*. 73-A (Part 1), 125-141 (2011)
7. Messaci, F., Nemouchi, N., Ouassou, I., Rachdi, M.: Local polynomial modelling of the conditional quantile for functional data. *Statistics Methods and Applications*. 24, No. 4, 597-622 (2015)

Time Series predictor based on deterministic and stochastic assumptions

Pedro Cadahia¹, José M. Bravo¹, Manuel E. Gegundez¹, and Antonio Golpe^{2*}

¹Escuela Técnica Superior de Ingeniería, Universidad de Huelva,
Carretera Huelva-Palos de La Frontera s/n. 21819. La Rábida - Palos de la Frontera.
Huelva. Spain

pedro.cadahia@outlook.es, {caro, gegundez}@uhu.es

²Facultad de Ciencias Empresariales y Turismo, Universidad de Huelva,
Plaza de la Merced, 11. 21002 Huelva, Spain
antonio.golpe@dehie.uhu.es

Abstract. In this article we reconsider the prediction problem in time series by using a new nonparametric approach. The prediction is obtained by a weighted sum of past observed data. These weights are obtained solving a constrained linear optimization problem that minimizes an outer bound of the prediction error. The main novelty of the proposed predictor is to consider deterministic and stochastic assumptions in order to obtain the upper bound of the prediction error. A tuning parameter is used to balance these deterministic-stochastic assumptions to improve the predictor performance. An example is included to illustrate that the proposed predictor can obtain suitable results in a prediction scheme and can be an interesting alternative method to classical nonparametric methods.

Keywords: time series forecasting; nonparametric regression; optimization

1 Introduction

The aim of this paper is to provide a new predictor for time series based on the observed past values of the time series, by means of a nonparametric approach. It is well-known fact that in parametric time series analysis the relationship between observed past values of the time series and the prediction is defined by specifying a functional form and a fixed finite number of parameters. Widely studied parametric options are autoregressive (AR) models, moving average (MA) models, and different combinations as ARMA or ARIMA models [2, 9]. In nonlinear time series, some common parametric structures has been studied, the threshold autoregressive (TAR) models [12], the exponential autoregressive (EXPAR) model [10] and smooth-transition autoregressive (STAR)

* This research has been supported by DPI2016-76493-C3-2-R of Ministerio de Economía y Competitividad (Spain).

models [11] are some examples. The performance of the parametric predictor is a consequence of the a priori function form chosen.

By contrast, in nonparametric approaches a more flexible class of functions is considered. Nonparametric methods avoid the choosing of a specific functional form. Collected data provides the information to obtain a new prediction. The price to pay is the 'curse of dimensionality', that is, a possible poor performance in high dimensions prediction problems. Local conditional mean or median method provides a prediction using the mean or the median of a neighborhood of the interest point [8]. The Nadaraya-Watson estimator [7] averages past observations by a kernel function to obtain a prediction. Local linear or polynomial functions of past observations can be used to approximate a nonlinear relationships [6, 4]. Semiparametric models as nonlinear additive autoregressive (NAAR) models [3] or functional coefficient autoregressive (FAR) models have been proposed too [13]. An extensive review of nonparametric method applied to time series prediction can be found in [5, 1].

In this paper a new nonparametric prediction method is proposed. The prediction is obtained by a weighted sum of past observations. An upper bound of the prediction error is computed under some deterministic and stochastic assumptions. A constrained optimization problem is formulated to minimize the upper bound of the prediction error and to obtain the set of optimal weights used to compute the prediction. The optimization problem includes a parameter to balance the deterministic-stochastic assumptions. This is the main novelty of the proposed method. This parameter can be tuned with training data and a cross-validation scheme to improve the predictor performance. The proposed predictor provides a general framework that encompasses some relevant nonparametrics predictors as the Nadaraya-Watson predictor [7] or predictors based on local linear regression [4].

The paper is organized as follows. In Section 2, the problem formulation is addressed. The deterministic and stochastic assumptions are presented in Section 3. The new predictor is proposed in Section 4. An example is illustrated in Section 5. Finally, Section 6 reports some conclusions.

2 FORMULATION

Let us consider a discrete¹ time series process $\{z_t\}$ with $t \in \{0, \pm 1, \pm 2, \dots\}$. At time instant k we assume that past data $\{z_t\}$ with $t \in \{k, k-1, k-2, \dots\}$ has been observed and we are interested in providing a forecast for predicting z_{k+1} . Once we detrend,² the time series is now the series $\{y_t\}$ with $t \in \{k, k-1, \dots\}$, where $z_t = y_t + \mu_t$, being μ_t the trend component and y_{k+1} the detrended future time series value.

¹ We assume a discrete version of data.

² It should be noted that in coherence with the prediction system and in order to estimate μ_{k+1} , only the past observations can be used, independently of the detrending method used.

Let us also denote by $\{x_i\}$ with $i = 0, 1, \dots, k$ the set of the vectors consisting of the observed past values of the time series, that is $x_i = [y_i, y_{i-1}, \dots, y_{i-p}]^T$. Henceforth this p -dimensional vector set will be called *embedding vector*. This set of data is used to forecast future values for the time series. We should make this point clearer in order to precise the sense of the parametric and nonparametric models used in this article. A parametric approach is characterized by the use of the training set for estimating the parameters of the model and once this inference is done the data set is not used again. A nonparametric approach it is a local approach in which each forecast is obtained by using all the available data set but selecting a neighborhood of the interest point. In this sense, we assume that the time series can be generated by an unknown local linear model.

Assumption 1 Consider that we can model the forecast of y as:

$$y_{k+1} = r(x_k)^T \theta_k + e_k \quad (1)$$

where we are assuming the existence of an unknown vector of parameters $\theta_k \in \mathcal{R}^n$, a known function $r(\cdot)$ valued at the embedding set and an unknown error term e_k .³

In order to complete the presentation of the model we should discuss in more detail the so called *regressor generator function* $r(\cdot)$. This function allows transform the original values into vectors of dimension n_r by means of the vectors belonging to the embedding set. A formal definition of this regressor generation function is as follows.

Definition 1 (Regressor generator function). The function $r(\cdot) : \mathcal{R}^p \rightarrow \mathcal{R}^{n_r}$ defines the components of the regressor vector. This function admits any kind of autoregressive representation, nonlinear expression of past components and different functional forms for decomposing the different components of the time series.⁴

The aim of this paper is to provide a new predictor for time series that are based on past measurement. An estimation of the output y_{k+1} is obtained by a linear combination of outputs y_i , with $i = 1, 2, \dots, k$ (see [17]).

Definition 2 (Linear Prediction). At time instant k , a prediction of $y_{k+1} \in \mathbb{R}$ can be obtained by a linear combination of the past system outputs, that is,

$$\begin{aligned} \hat{y}_{k+1}(\lambda) &= b_Y^T \lambda \\ &= \sum_{j=1}^k \lambda_j y_j \end{aligned} \quad (2)$$

³ This modelling is flexible enough to admit alternative assumptions about the error term. As we will discuss later, we will present the model by using both deterministic and stochastic bounds for the error term e_k .

⁴ For instance suppose a set $x_k = [y_k, y_{k-1}, y_{k-2}]$. Then $r(x_k)$ could be the function $r(x_k) = x_k$ that is, an autoregressive model. We can also use alternative configurations such as a nonlinear autoregressive model $r(x_k) = [y_k^2, y_{k-1}, y_{k-2}, y_k \cdot y_{k-2}]$ or any possible combination.

where $\lambda \in \mathbb{R}^k$ is a weights vector and $b_Y = [y_1, \dots, y_k]^T$.

Now it is possible to define the prediction error as the difference of y_{k+1} and the linear prediction $\hat{y}_{k+1}(\lambda)$.

Definition 3 (Prediction error). *At time instant k , the prediction error $\hat{e}_k(\lambda)$ is defined by*

$$\hat{e}_k(\lambda) = y_{k+1} - \hat{y}_{k+1}(\lambda). \quad (3)$$

Therefore, the key issue is how to obtain the weights vector λ and an outer bound of the prediction error. This outer bound is estimated using the assumed relationship between x_{i-1} and y_i , with $i = 1, 2, \dots, k$ in expression (1). Then, a set of past components x_i with $i = 0, 1, \dots, k$ must be available. Next section formulates these key ideas.

3 Assumptions based on local affine approximations

The proposed predictor is based on the approximation error. This approximation error is defined as the error resulting from the use of vectors $r(x_{j-1})$ and θ_k to infer the output y_j .⁵

Definition 4 (Approximation error). *Given vector θ_k , the approximation error e_{j-1} associated to the pair (x_{j-1}, y_j) with $j = 1, 2, \dots, k$ is defined by*

$$e_{j-1} = e_{j-1}(\theta_k) = y_j - r(x_{j-1})^T \theta_k. \quad (4)$$

Henceforth, and with a slight abuse of notation, the explicit dependence of $e_{j-1}(\theta_k)$ with respect to θ_k is omitted. Note that the exact value of θ_k is unknown. On the other hand, it is clear that the prediction error $\hat{e}_k(\lambda)$ can be influenced by the chosen vector λ . Next theorem proposes an expression to characterize the prediction error $\hat{e}_k(\lambda)$ as a function of vector λ and approximation errors e_j previously defined.

Theorem 1. *For any vector $\lambda \in \mathbb{R}^N$ such that*

$$\sum_{j=1}^k \lambda_j r(x_{j-1}) = r(x_k) \quad (5)$$

then prediction error $\hat{e}_k(\lambda) = y_{k+1} - \hat{y}_{k+1}(\lambda)$ is a linear combination of approximation errors e_j , that is

$$\hat{e}_k(\lambda) = - \sum_{j=1}^k \lambda_j e_{j-1} + e_k.$$

⁵ The reader should note that the point is to relate the k -th prediction error e_k and the prediction errors generated by using the k -th vector of unknown parameters θ_k with the i -th regressors $r(x_i)$, with $i = 0, \dots, k-1$.

Note that λ_i denotes the i -th element of vector λ .

Proof. In matrix notation, expression (5) is equivalent to $\lambda \in \{\lambda : A^T \lambda = r(x_k)\}$ where matrix A is defined by

$$A^T = [r(x_0) \ r(x_1) \ \dots \ r(x_{k-1})] \quad (6)$$

Taking into account Assumption 1 and Definitions 3 and 4 the following equalities can be inferred.

$$\begin{aligned} \hat{e}_k(\lambda) &= y_{k+1} - \hat{y}_{k+1}(\lambda) \\ &= y_{k+1} - \lambda^T b_Y \\ &= r(x_k)^T \theta_k - \lambda^T b_Y + e_k \\ &= (A^T \lambda)^T \theta_k - \lambda^T b_Y + e_k \\ &= \lambda^T (A \theta_k - b_Y) + e_k \\ &= \sum_{j=1}^k \lambda_j (r(x_{j-1})^T \theta_k - y_j) + e_k \\ &= - \sum_{j=1}^k \lambda_j e_{j-1} + e_k. \end{aligned}$$

QED

In order to obtain a value of error e_{j-1} , the vector θ_k must be known, but an exact value of θ_k is not available. However, other properties of e_{j-1} can be assumed. Deterministic and stochastic options are present in the literature. From a deterministic point of view, an upper bound of $|e_{j-1}|$ could be considered.

3.1 Deterministic error assumption

In bounded-error methods (see [15]), an unknown-but-bounded error is considered and an upper bound of this error is assumed. A similar consideration is assumed here.

Assumption 2 *These are constants $\sigma, L \geq 0$ such that approximation error e_{j-1} and e_k are bounded by expressions*

$$|e_{j-1}| \leq \sigma + L \|x_{j-1} - x_k\| \quad (7)$$

with $j = 1, \dots, k$ and

$$|e_k| \leq \sigma \quad (8)$$

where $\|\cdot\|$ is a norm.

Notice that the error term is bounded by $|e_k| \leq \sigma$. Assumption 2 has been widely used in the context of bounded-error system identification (see [15]). Note that constant σ defines the minimum level of noise considered and L the uncertainty due to the local affine approximation.

Remark 1. If no previous knowledge about constants σ and L is available, a set of historical data can be used to estimate an approximated value of σ and L . In [16] a method based on bounded-error and non-falsified data is proposed.

Lemma 1. *Taking into account Assumptions 1 and 2, for any λ such that $A^T \lambda = r(x_k)$, prediction error $\hat{e}_k(\lambda) = y_{k+1} - \hat{y}_{k+1}(\lambda)$ is bounded by expression*

$$|\hat{e}_k(\lambda)| \leq \sum_{j=1}^k |\lambda_j|(\sigma + L||x_{j-1} - x_k||) + \sigma. \quad (9)$$

Proof. Expression (9) is obtained by a direct application of Theorem 1 and bound $|e_i| \leq \sigma + L||x_i - x_k||$. QED

Now it is possible to propose a suitable option to obtain vector λ . A reasonable choice is to consider the vector that minimizes an upper bound of $|\hat{e}_k(\lambda)|$ using expression (9).

Definition 5 (Deterministic predictor). *The deterministic prediction $\hat{y}_{k+1}(\lambda^D)$ is defined by expression*

$$\hat{y}_{k+1}(\lambda^D) = \sum_{j=1}^k \lambda_j^D y_j$$

where vector λ^D solves the following constrained linear optimization problem

$$\begin{aligned} \lambda^D = \arg \min_{\lambda} \quad & ||W_k \lambda||_1 \\ \text{s.t.} \quad & A^T \lambda = r(x_k) \end{aligned} \quad (10)$$

where W_k is a diagonal matrix with central elements $w_{k,j-1} = \sigma + L||x_{j-1} - x_k||$ with $j = 1, \dots, k$. Therefore, vector λ^D minimizes an upper bound of the absolute value of the prediction error.

Note that notation λ^D emphasizes the deterministic nature of the estimation. Expression (10) use L_1 -norm to obtain the vector solution λ^D . In this case, λ^D is sparse, that is, the number of components λ_i^D of vector λ^D that are different from zero remains small. As λ^D is sparse and taking into account Definition 5 then it is inferred that $\hat{y}_{k+1}(\lambda^D)$ use a relative small number of measurements y_i .

3.2 Stochastic error assumption

From a stochastic point of view, a second option is to consider the approximation error e_i as a random variable. In this case some assumptions about mean and variance of e_i can be considered.

Assumption 3 *Approximation error e_{j-1} and error term e_k are independent random variables of zero mean and variances bounded by $\text{var}(e_{j-1}) \leq \sigma + L||x_{j-1} - x_k||$ and $\text{var}(e_k) \leq \sigma$ respectively. Again, the value of positive constants σ and L is the assumed prior knowledge.*

As noted in Remark 1, if no previous knowledge about constants σ and L is available, a set of historical data can be used to obtain an estimation. The variance of error e_{j-1} consists of a minimum value defined by σ and a term depending of the local approximation, i.e. $\|x_{j-1} - x_k\|$. As e_{j-1} and e_k are random variables then $\hat{e}_k(\lambda)$ is a random variable too and some properties can be derived.

Assumption 4 *Taking into account Assumptions 1 and 3, for any λ such that $A^T \lambda = r(x_k)$, prediction error $\hat{e}_k(\lambda) = y_{k+1} - \hat{y}_{k+1}(\lambda)$ is a random variable with zero mean and variance that can be approximated by expression*

$$\begin{aligned} \text{var}(\hat{e}_k(\lambda)) &= \sum_{j=1}^k \lambda_j^2 \text{var}(e_{j-1}) + \sigma \\ &\leq \sum_{j=1}^k \lambda_j^2 (\sigma + L \|x_{j-1} - x_k\|) + \sigma. \end{aligned} \quad (11)$$

Now, a predictor that minimize the outer bound of the variance prediction error can be formulated.

Definition 6 (Stochastic central prediction). *The stochastic prediction $\hat{y}_{k+1}(\lambda^S)$ is defined by expression*

$$\hat{y}_{k+1}(\lambda^S) = \sum_{j=1}^k \lambda_j^S y_j$$

where vector λ^S solves the following constrained linear optimization problem

$$\begin{aligned} \lambda^S &= \arg \min_{\lambda} \quad \lambda^T W_k \lambda \\ \text{s.t.} \quad &A^T \lambda = r(x_k) \end{aligned} \quad (12)$$

An explicit solution of this optimization problem is obtained by

$$\lambda^S = W_k^{-1} A (A^T W_k^{-1} A)^{-1} r(x_k) \quad (13)$$

As before, notation λ^S emphasizes the stochastic assumptions considered to obtain the estimation. The following equality is fulfilled

$$\hat{y}_{k+1}(\lambda^S) = b_Y^T \lambda^S = r(x_k)^T \theta^*$$

where $\theta^* = (A^T W_k^{-1} A)^{-1} A^T W_k^{-1} b_Y$ is the minimizing argument of an optimization problem which minimizes the following quadratic prediction-error functional cost.

$$\begin{aligned} J(\theta) &= (b_Y - A\theta)^T W_k^{-1} (b_Y - A\theta) \\ &= \sum_{j=1}^k \frac{(y_j - r(x_{j-1})^T \theta)^2}{(\sigma + L \|x_{j-1} - x_k\|)}. \end{aligned} \quad (14)$$

Therefore, the stochastic prediction is equivalent to solve a weighted least-squares problem where the weights are defined by the elements of the diagonal of W_k squared. Usually, vector λ^S is not sparse. Then a great number of measurements y_i are used to predict y_{k+1} .

Vectors λ^D and λ^S provide two predictions based on different assumptions. The aim of this paper is to provide a predictor mixing both predictors. Next section presents the main idea of this paper.

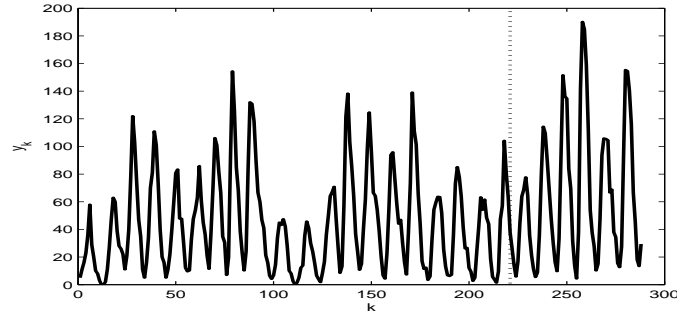


Fig. 1. SunSpot serie.

Table 1. Kernel functions

Epanechnikov	Gaussian	Tricube
$\lambda_i = \begin{cases} 1 - v_i^2 & \text{si } v_i \leq 1 \\ 0 & \text{si } v_i > 1 \end{cases}$	$e^{-\frac{1}{2}v_i^2}$	$\lambda_i = \begin{cases} (1 - v_i ^3)^3 & \text{si } v_i \leq 1 \\ 0 & \text{si } v_i > 1 \end{cases}$
$v_i = \frac{\ x_i - x_k\ }{\gamma \frac{1}{N} \sum_{j=1}^N \ x_i - x_k\ }$		

4 Proposed predictor

Next, a formal definition of the proposed predictor is provided. This definition use a constant $\gamma \geq 0$ to balance the deterministic or stochastic nature of the prediction.

Definition 7. Given a constant $\gamma \geq 0$, the predictor $\hat{y}_{k+1}(\lambda^*)$ is defined by $\hat{y}_{k+1}(\lambda^*) = \sum_{j=1}^k \lambda_j^* y_j$ where λ^* is the optimal solution of

$$\begin{aligned} \lambda^*(\gamma) = \arg \min_{\lambda} \quad & \|W_k \lambda\|_1 \\ \text{s.t.} \quad & A^T \lambda = r(x_k) \\ & \|\lambda - \lambda^S\|_1 \leq \gamma \end{aligned} \quad (15)$$

and vector λ^S is defined in (13).

Some qualitative properties of the proposed predictor can be clarified. Note that, expression (15) is a constrained linear convex optimization problem and can be solved in an efficient way [14]. Assuming that (15) has a bounded solution, there is a constant $\bar{\gamma}$ such that if $\gamma \geq \bar{\gamma}$ then equality $\lambda^* = \lambda^D$ is obtained. Term $\|\lambda - \lambda^S\|_1$ of expression (15) takes into account the stochastic Assumption 3 to obtain the optimal solutions λ^* . If $\gamma = 0$ then $\lambda^* = \lambda^S$. So, constant γ can be seen as a tuning parameter to balance the deterministic or stochastic nature of the considered approximation error.

Remark 2. It is important to remark that the proposed predictor encompasses some relevant nonparametrics predictors. If $\gamma = 0$ and $r(x_k) = 1$ the proposed predictor is equivalent to the Nadaraya-Watson predictor [7]. On the other hand if $\gamma = 0$ and $r(x_k) = [x_k^T \ 1]$ a predictor based on Local Linear Regression is obtained.

Table 2. Obtained MAE results

	γ	MAE	γ	MAE
$AR(10)$				12.76
NW	0.13	18.47	0.15	18.38
LLR_1	1	12.33	1.31	11.71
LLR_2	0.59	11.90	0.76	11.68
LLR_3	1.17	12.21	1.47	11.74
CP	1.1	11.44	1.0	11.34

Table 3. Obtained MSE results

	γ	MSE	γ	MSE
$AR(10)$				299.46
NW	0.12	697.54	0.12	697.54
LLR_1	1.08	260.01	1.15	254.28
LLR_2	0.6	255.20	0.65	254.38
LLR_3	1.23	257.29	1.32	254.53
CP	1.1	251.08	1.1	251.08

Table 4. Combined results

	γ	MAE	γ	MSE
$LLR_1 - LLR_2$	1,0.59	12.11	1.08,0.6	256.19
$LLR_1 - LLR_3$	1,1.17	12.27	1.08,1.23	258.52
$LLR_2 - LLR_3$	0.59,1.17	12.04	0.6,1.23	255.39
$CP - LLR_1$	1.1,1	11.71	1.1,1.15	249.08
$CP - LLR_2$	1.1,0.59	11.58	1.1,0.65	249.23
$CP - LLR_3$	1.1,1.17	11.65	1.1,1.32	248.13

Table 5. Obtained results

	γ	MAE	γ	MAE
$AR(10)$				13.09
NW	0.13	17.06	0.13	17.06
LLR_1	1	11.19	1.15	11.07
LLR_2	0.59	11.11	0.56	11.09
LLR_3	1.17	11.18	1.28	11.08
CP	1.1	10.88	2.0	10.80

Table 6. Obtained results

	γ	MSE	γ	MSE
$AR(10)$				309.71
NW	0.12	582.62	0.12	582.62
LLR_1	1.08	223.15	1.04	222.86
LLR_2	0.6	229.52	0.55	228.61
LLR_3	1.23	224.60	1.24	224.58
CP	1.1	231.03	0.13	230.42

5 Example

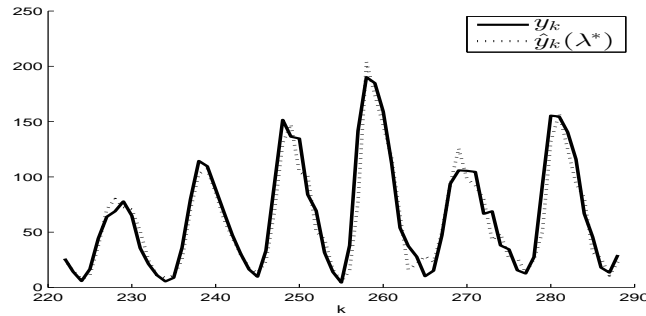
The Sunspot Numbers is an interesting benchmark for predictions methods because it is a real world example, the prediction of these data is relevant in many applications fields and these data are public. The time series consisting in 288 data, from year 1700 to 1987 (see Figure 1). The first 221 data have been used as training set. The last 67 data have been included in a validation set. A one-step ahead predictor has been considered where $r(x_k) = x_k = [y_{k-1} \ y_{k-2} \ \dots \ y_{k-10}]^T$.

The proposed predictor (CP) with $\sigma = 0$ and $L = 1$ is compared to an Autoregressive model $AR(10)$, a Narayada-Watson predictor (denoted NW) with a Gaussian kernel function and a γ -bandwidth and three local linear regression models LL_1 , LL_2 and LL_3 using Epanechnikov, Gaussian and Tricube kernel functions respectively. Table 1 shows the expression of weights λ_i with $i = 1, \dots, N$ for these kernel functions. Mean absolute error (MAE) and mean square error (MSE) are used as indexes to compare.

Tables 2, 3 and 4 show the obtained results using only the training set to infer the prediction. Table 2 shows the performance obtained by these prediction methods in the validation set by a Mean Absolute Error (MAE) index. The first column of Table 2 shows parameters γ selected by a leave-one-out scheme in the training set. The second column shows the MAE obtained with this value of γ in the validation set. The third column gives the value γ that provides the best MAE in the validation set that is displayed in the fourth column. A grid-search in the validation set is used in this case. Note that the third and fourth columns are included to estimate the best possible performance, but can not be used as reference because the optimal value of γ is not known a priori.

Table 7. Combined results

	γ	MAE	MSE
$LL1 - LL2$	1,0.59	11.12	225.54
$LL1 - LL3$	1,1.17	11.19	223.78
$LL2 - LL3$	0.59,1.17	11.12	226.56
$CP - LL1$	1.1,1	10.85	223.08
$CP - LL2$	1.1,0.59	10.91	227.24
$CP - LL3$	1.1,1.17	10.88	223.67

**Fig. 2.** Proposed predictor with $\gamma = 1.1$

As can be seen in the Table 2 the best results are obtained by the proposed predictor (CP). The same information with Mean Square Error (MSE) is showed in Table 3. Again the best results are obtained by the proposed predictor. Finally, new predictors obtained by the combination of two nonparametric predictors are proven. For example, $LLR_1 - LLR_2$ is a predictor obtained by the mean of predictors LLR_1 and LLR_2 . Table 4 shows the results obtained using these combined predictors. In all cases, a combination with CP outperforms the results obtained by a single Local Linear Regression or combinations of Local Linear Regressions.

If all past information available at time instant k is used to infer the prediction, Tables 5, 6 and 7 show the new results. In general the fresh data improve the performance of predictors. As can be seen, the proposed predictor CP obtain the best results using the MAE index, or can be combined with other predictors to obtain the best MSEs. Figure 2 shows the obtained prediction with $\gamma = 1.1$.

6 Conclusions

A new nonparametric Time Series forecasting method has been proposed. The prediction is obtained by a weighted sum of past observations. A combination of deterministic and stochastic assumptions are used to obtain an expression of the outer bound of the prediction error. The weights are obtained solving a

convex optimization problem that minimizes the upper bound of the prediction error. The method includes a tuning parameter. This parameter may balance the deterministic and stochastic considered assumptions. By a cross-validation scheme, a suitable parameter can be obtained. The performance of the proposed predictor is illustrated by an example.

References

1. Gao, J.: Nonlinear Time Series: Semiparametric and Nonparametric Methods. Chapman and Hall/CRC, Australia (2007)
2. Box, G. E. P., Jenkins, G. M.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1976)
3. Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman and Hall (1990)
4. Fan, J., Gijbels, I.: Local polynomial modelling and its applications. Chapman and Hall, London (1996)
5. Fan, J., Yao, Q.: Nonlinear Time Series: Nonparametric Methods and Parametric Methods. Springer, New York (2003)
6. Härdle, W.: Applied nonparametric regression. Cambridge University Pr, Cambridge u. a. (1990)
7. Nadaraya, E. A.: On Estimating Regression. Theory of Probability & Its Applications. 9(1), 141-142 (1964)
8. Truong, Y. K.: A nonparametric framework for time series analysis. Springer, New York (1993)
9. Hamilton, J. D.: Time series analysis. Princeton Univ. Press, New Jersey (1994)
10. Haggan, V., Ozaki, T.: Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. Biometrika. 68, 186-196 (1981)
11. Chang, K. S., Tong, H.: On estimating thresholds in autoregressive models. Journal of Time Series Analysis. 7, 179-190 (1986)
12. Tong, H.: Threshold Models in Nonlinear Time Series Analysis (Vol. 21). Springer, Heidelberg (1983)
13. Härdle, W., Lükpohl, H., Chen, R.: A Review of Nonparametric Time Series Analysis. International Statistical Review. 65(1), 49-73 (1997)
14. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press. (2004)
15. Milanese, M., Norton, J., Piet-Lahanier, H., Walter, E.: Bounding Approaches to System Identification. Plenum Press, New York (1996)
16. Milanese, M., Novara, C.: Set Membership identification of nonlinear systems. Automatica. 40(6), 957-975 (2004)
17. Roll, J., Nazin, A., Ljung, L.: Nonlinear system identification via direct weight optimization. Automatica. 41(3), 475-490 (2005)

Functional Data Classification by Discriminative Interpolation with Features

Rana Haber¹, Anand Rangarajan², Nenad Mijatovic¹,
Anthony O. Smith¹, and Adrian M. Peter¹

¹Florida Institute of Technology, Melbourne, FL, 32901, USA

²University of Florida, Gainesville, FL, 32611, USA **

rhaber2010@my.fit.edu, anand@cise.ufl.edu, nmijatov2005@my.fit.edu
anthonymsmith@fit.edu, apeter@fit.edu

Abstract. In this article, we present a novel approach for the categorization of functional data, i.e. curves that are inherently varying over a continuum such as time or space. The method proposes a new vector-valued functional representation of input time series data that inherently encode the time evolution of localized features. This vector-valued feature function is subsequently represented in a wavelet basis. During training, the wavelet representation of these multivariate time series in the same class are warped to become more similar to each other, while moving away from functions in different classes. This process—termed *discriminative interpolation*—leads to a k -nearest neighbor (k -NN) style supervised learner that induces discriminating warps through adaptation of the wavelet basis coefficients. We detail the improvement gains from adopting a generalized vector-valued feature representation for the functional data, illustrating consistent performance improvement over previous formulations. We term the overall approach *Classification by Discriminative Interpolation with Features* (CDIF). Optimization of the new objective function is accomplished via an alternating procedure which has empirically shown to achieve convergence to good approximate solutions. The utility of the proposed CDIF method is experimentally validated on several UCR time series data sets, demonstrating its competitiveness against related functional techniques as well as contemporary and state-of-the-art feature-based methods.

Keywords: Functional Data Classification, Time Series Classification, k -Nearest Neighbor, Discriminative Interpolation, Wavelets

1 Introduction

In today's data rich environment, a multitude of sensor-acquired data come in the form of time series data. Examples include weather (daily temperature measured over time), medical data (ECG readings), economics (stock market), and astronomy (starlight measurements). These measurements are often referred to as functional data, i.e. mathematical representation of real-valued functions

** The authors acknowledge support from NSF grant No. 1560345. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

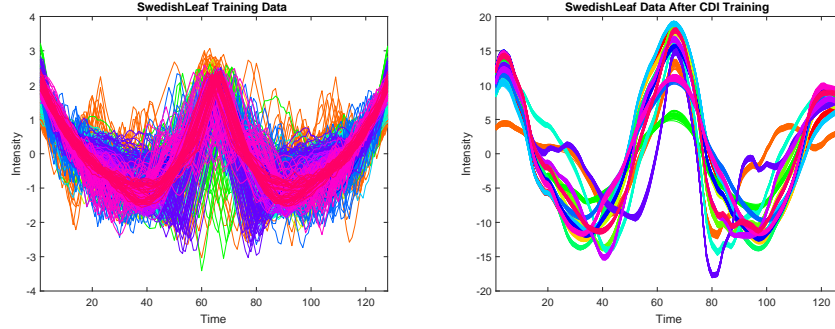


Fig. 1. Fifteen-class *classification by discriminative interpolation* with features (CDIF). (a) Original training functions from fifteen different classes. (b) Post training via the proposed CDIF method, functions in each class morphed to resemble k -nearest neighbors. (Note: The colors of the curves represent the different classes. For visualization purposes, only the curves were used instead of multivariate features.)

collected over a continuum like time or space. The need to analyze such data has lead to the growth of the Functional Data Analysis (FDA) [19] field. Most contemporary machine learning approaches employ a feature extraction process that strips the functional nature of these data, transforming them instead to feature vectors in \mathbb{R}^n . Rather than adopt this informal methodology, here we adopt the FDA approach and work rigorously in functions spaces to develop our proposed classification framework. As we detail, this has yielded a unique signal representation and classification framework that shows significant promise to tackle difficult functional data categorization problems.

The present work focuses on one-dimensional functional data analysis—where the data are time-series measurements (and throughout this article we interchangeably use the following terms to refer to functional data: time series, signals, curves, or waveforms). We leverage the interpolation property of functions to formulate a new classification model. It is worth noting that such an approach is not rigorously possible with the feature-vector approach due to the absence of a continuum linking the dimensions. In our new method, functional data in the same class are adaptively reconstructed to be more similar to each other, while simultaneously repelling nearest neighbor functional data in other classes. Akin to other recent nearest-neighbor metric learning paradigms like stochastic k -neighborhood selection [20] and large margin nearest neighbors [21], our technique uses class-specific representations which gerrymander similar functional data in an appropriate parameter space. The present work develops a fully generalized version of the model first considered in [14]. Whereas this previous work was only developed for one-dimensional functional data, $f : \mathbb{R} \rightarrow \mathbb{R}$, we considerably advance the framework by formulating an extension capable of supporting vector-valued functions, i.e. $f : \mathbb{R} \rightarrow \mathbb{R}^q$. The vector-valued extension allows us to incorporate localized feature curves, allowing for a richer functional representation. For the remainder of the paper, we will refer to this new framework as *Classification by Discriminative Interpolation with Features (CDIF)*.

Figure 1 demonstrates the warping effect of CDIF on the *SwedishLeaf* dataset. Fig. 1 (a) graphs the original curves from the fifteen classes (each color represents a different class), while Fig. 1 (b) plots the curves after the CDIF training. The curves within the same class better resemble each other resulting in better classification performance, as shown in our experiments. The proposed CDIF method makes the following unique contributions to advancing functional data classification:

- Introduces a new vector-valued feature signal representation for functional data;
- Implements a *fully multi-class* FDA classification framework that leverages push-pull k -NN margin-based learning;
- Employs an alternating optimization strategy that first updates neighborhoods and then completes a gradient descent update on the wavelet coefficients.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 briefly covers the requisite background on function spaces and wavelets as well as explains the CDIF work in more detail. This is followed by the development of the training and testing algorithms required for supervised classification. In Section 4, we show that our method is competitive with many functional and standard feature-vector based algorithms, as well as show improvements gained from the original method in [14]. All experimental results are evaluated on UCR datasets [9]. We conclude with Section 5, detailing recommendations and future extensions.

2 Related Work

For several years, the norm for classification or analysis of functional data, such as time series data, has been to represent the data in vector form [2, 4, 12], thus turning a temporal problem into a static one. The authors in [12] proposed Feature-Based Linear (FBL), a method that relied on the multidimensional features of the data, such as basic statistics, correlations, etc. These features are then separated by a greedy forward feature selection algorithm with a linear classifier, which according to the authors is competitive with state-of-the-art classifiers, but is still computationally expensive which might not be practical in many applications.

The authors in [4] suggest a method called Time Series based on Bag-of-Features (TSBF). It first separates a signal into random time intervals, and then computes features such as the start and end interval points, mean, variance, and slope which create a bag-of-features. The likelihood that a set of features, called an instance, belong to a signal (bag) was calculated as a class probability. At the end, the authors used a Support Vector Machine accompanied with a random forest for signal classification. In [16], the authors used the Principal Components Analysis through Conditional Expectation (PACE) method to decompose signals into principal components and then performed classification by applying logistic regression to the scores.

Similar to FBL, the collective of transformation-based ensembles (COTE) framework from [3] creates a large set of data features. Bagnall *et al.* derive three different domains from the data: shapelet, frequency, and autocorrelation. Then, they showed that forming a collective of ensembles of classifiers— k NN, Naive Bayes, C5.4 Decision Tree, Support Vector Machines (SVM), Random Forests, Rotation Forest, and a Bayesian network—on these three data representations improved the classification of time-series data. While these methods have shown promising results, they fail to address the functional aspect of the data.

The authors in [1, 2, 5, 6, 14] acknowledge the importance of incorporating the robustness of the data. These include, but are not limited to, representing the data in a different basis such as splines [1], Fourier [6] and wavelets [5] or utilizing the continuous aspects and differentiability of the functional data [2]. Berlinet *et al.* [5] expand the observations on a wavelet basis as well and then they pick the set of basis functions that carry the most significant information, through a multiple step algorithm, while learning an optimal classifier. These classifiers include k -nearest neighbors, Quadratic Discriminant Analysis (QDA) [11] and classification and regression trees (CART) [8].

3 Classification by Discriminative Interpolation with Features

In this section, we detail the theory and algorithmic steps for the proposed *Classification by Discriminative Interpolation with Features* (CDIF) method. We begin by providing a brief overview of the function spaces considered in this work and the necessary results therein. Next, we detail the localized feature extraction process that maps an input one-dimensional function to a vector-valued function. Since CDIF is a supervised classification method, we proceed to discuss the required training and testing procedures.

3.1 Background

CDIF relies on the fundamental characteristics of functional data. Even though the curves are discretely sampled, they are intrinsically infinite dimensional. The consecutive measurements within a curve are highly correlated and they are assumed to have an underlying smooth function.

It is important to emphasize that the functional data, f , we consider are continuous functions that belong to the space of square-integrable functions $\mathcal{L}^2([a, b] \subset \mathbb{R})$ that have a defined inner product, $\langle f, g \rangle := \int_t fg$ for $f, g \in \mathcal{L}^2$, and a norm, $\|f\| := (\int_{\mathcal{T}} f^2)^{1/2}$. This premise allows the analysis to transition from the functions themselves to the coefficients of their wavelet basis expansion which has many benefits such as easily handling irregularly sampled functions, missing data, and function interpolation [15].

Though many different bases exist for the Hilbert space \mathcal{H} , CDIF uses compactly supported wavelets to get a faithful reconstruction. Functions $f \in \mathcal{L}^2([a, b])$ can be represented as a linear combinations in a wavelet basis [10]:

$$f(t) = \sum_k c_{j,k} \phi_{j,k}(t) \quad (1)$$

where $t \in \mathbb{R}$, $\phi(x)$ is the *scaling* basis function, and $c_{j,k}$ is the scaling coefficients; the j -index represents the resolution level and the k -index the integer translation value. Note how the expansion is done only using the scaling bases functions because using a full multiresolution expansion with both the scaling $\phi(x)$ and the wavelet $\psi(x)$ basis function is primarily useful to obtain a sparse representation of the signal with thresholded coefficients. This is not crucial for an accurate representation. Also, CDIF, like all other computational efforts, adopts a projection onto a finite d -dimensional subspace \mathcal{P} . Given a discretely sampled function $\mathbf{f} = \{f(t_i)\}_{1 \leq i \leq m}$, the optimal coefficients can be found by minimizing the quadratic objective function

$$\min_{\mathbf{c}} \|\mathbf{f} - \phi \mathbf{c}\|_2^2, \quad (2)$$

where ϕ is an $m \times d$ matrix with entries $\phi_{i,l} = \phi_l(t_i)$ and \mathbf{c} is a $d \times 1$ column vector of the coefficients. For an orthonormal basis such as the wavelets basis Φ , defined by $\Phi_{r,s} = \langle \phi_r, \phi_s \rangle$ is equivalent to $\Phi = \mathbf{I}$. This allows results, such as $\|\mathbf{f}^h - \mathbf{f}^j\|_2^2$, to reduce to the basis coefficients

$$\|\mathbf{f}^h - \mathbf{f}^j\|_2^2 = (\mathbf{c}^h - \mathbf{c}^j)^T \Phi (\mathbf{c}^h - \mathbf{c}^j) = \|\mathbf{c}^h - \mathbf{c}^j\|_2^2. \quad (3)$$

This result is used in both the pull and push terms of CDIF.

3.2 Localized Feature Representation

For CDIF, we derive several feature functions from the original signal and we use a selective combination of these features when classifying our data. Each feature results in a function that is of the same length of the original signal. Note that this is a design choice not a limitation. Aggregating multiple feature curves result in a vector-valued function.

The features we utilized in the present work are: identity map, instantaneous frequency [7], derivative (first, second and third), a few localized statistical measures (standard deviation, skewness, kurtosis, median, and autocorrelation), and local binary patterns (LBP) [17]. Some of the feature extraction methods naturally produce functional outputs, e.g. derivatives. For others, we induce a function output by using a n -point sliding window over the original signal. A summary of all the features can be found in Table 1.

This feature extraction process can be formalized by defining a feature mapping operator $\Psi : \mathcal{H} \rightarrow \mathcal{H}$ such that $\Psi f = g$, where $g \in \mathcal{L}^2([a, b] \subset \mathbb{R})$. The inner product and norm defined on f also hold for g . In CDIF, we consider a variety of operators Ψ that act on our original f and produce a set of feature functions g_q . We aggregate those g_q 's into a generalized vector-valued function, $\mathbf{g}(t) = [g_1(t), g_2(t), \dots, g_q(t)]^T$. As an example, consider the case where we use the derivative of a function as a feature, i.e. $\Psi \equiv D$ and $D : f \rightarrow g$. We assume the data to be smooth, meaning that the data f possesses one or more derivatives, indicated by $Df, D^2f, \dots, D^m f$ where m is the order of the derivative.

In another set of mappings, the Ψ 's are not as straightforward as the derivative. They are computed using neighboring data samples, $\Psi f[t_{j-\alpha}, t_{j+\alpha}] = g(t_j)$ where $\alpha \in \mathbb{Z}_+$. This technique is known as the sliding window technique allowing

Table 1. Detailed Feature Extraction Equations. Some features naturally produce functional outputs. For others, we use sliding windows (SW) to induce functional representations.

ID	Feature Name	Feature Function
1.	Identity Map	$g^i(t) = f^i(t)$
2.	Instantaneous Frequency [7]	$g^i(t) = \frac{1}{2\pi} \frac{d\omega(t)}{dt}$ where $\omega(t) = \arg\{f^i(t)\}$
3-5.	Derivative	$g^i(t) = \frac{d(f^i(t))^n}{d^n t}$ where $n = 1, 2, \text{ or } 3$
6.	SW Standard Deviation	$g^i(t) = \sqrt{\frac{1}{n} \sum_{j=1}^n (f^i(j) - \bar{f}^i)^2}$ where $\bar{f}^i = \frac{1}{n} \sum_{j=1}^n f^i(j)$ and n is the window size
7.	SW Skewness	$g^i(t) = \frac{1}{s^3} \sum_{j=1}^n (f^i(j) - \bar{f}^i)^3$ where s is the SW standard deviation
8.	SW Kurtosis	$g^i(t) = \frac{1}{s^4} \sum_{j=1}^n (f^i(j) - \bar{f}^i)^4$ where s is the SW standard deviation
9.	SW Median	$g^i(t) = \begin{cases} f^i(\lceil \frac{n}{2} \rceil) & \text{if } n \text{ is odd} \\ \frac{1}{2} (f^i(\frac{n}{2}) + f^i(\frac{n}{2} + 1)) & \text{if } n \text{ is even} \end{cases}$ where n is the window size
10.	SW Local Binary Pattern [17]	$g^i(t) = \sum_{j=1}^{n-1} I[f^i(j), f^i(t)] \times 2^{(j-1)}$ where $I[f^i(j), f^i(t)] = \begin{cases} 0, & f^i(j) < f^i(t) \\ 1 & f^i(j) > f^i(t) \end{cases}$ and j is the position of the neighbor cell
11.	SW AutoCorrelation	$g^i(t) = \sum_{j=1}^n f^i(j) f^i(j - t)$

us to compute a variety of localized features based on a neighborhood of contiguous samples. For example, the sample standard deviation $\sigma : \mathbb{R}^n \leftarrow \mathbb{R}$ takes in an n -dimensional data and returns a positive scalar. Standard deviation can have valuable information on the curve but to use it in our CDIF framework, we apply the sliding window technique which allows us to use sections of the curve to find a standard deviation value per discrete time t_j . The output is a function due to the nature of the sliding window technique, where $\sigma(t)$ represents the local standard deviation computed from the window of n samples around the time step t . Since we apply Ψ on the original curve with a much finer discretization, enough discretization such that $n \rightarrow \infty$ then g would be assumed to be intrinsically infinite dimensional as well. Also, the sliding window technique with overlapping "windows" creates a close correlation between the data points at t_j and t_{j+1} .

In our CDIF technique, we are interested in using the best features to better discriminate our data but at the same time, we do not want to lose the functional aspect of the data. Given a set of features in \mathbb{R}^d we can create an overall feature map $G : \mathcal{L}^2([a, b]) \rightarrow \mathcal{L}_1^2([a, b]) \times \mathcal{L}_2^2([a, b]) \times \cdots \times \mathcal{L}_q^2([a, b])$ such that

$$\mathbf{g}(t_j) = [g_1(t_j), g_2(t_j), \dots, g_q(t_j)]^T \quad (4)$$

is a vector-valued function at t_j and q is the number of features used. This is referred to by [18] as a q -dimensional parametric curve that still holds the same Hilbert space properties. This representation of the features of the functional data will be used in the CDIF framework.

3.3 CDIF Training Formulation

Now that we have established the vector-valued feature representation for CDIF, we provide a detailed sketch of the supervised training formulation. We begin with a labeled functional data set $\{(\mathbf{f}^i, y^i)\}_{i=1}^N$, where $\mathbf{f}^i \in \mathcal{H}$ are the one-dimensional functions and $y^i = \{1, \dots, A\}$ are the class labels with A being the number of categories. The discretized function data $\mathbf{f}^i = \{f^i(t_j)\}_{1 \leq j \leq m}$ are first used to derive all the feature vector-valued functions defined in Table 1. Denote the vector-valued feature curve i by \mathbf{g}_q^i where q represents the feature number as per the table. Observe that any single feature curve, $\mathbf{g}_q^i \in \mathbb{R}^{m \times 1}$, has the same dimension as \mathbf{f}^i . Now we can define the multivariate localized feature signal (MLFS) \mathbf{h}^i as the concatenation of a subset of the features:

$$\mathbf{h}^i = [\mathbf{g}_1^i, \mathbf{g}_2^i, \dots, \mathbf{g}_q^i]. \quad (5)$$

This denotes $\mathbf{h}^i \in \mathbb{R}^{m \times q}$ where q can vary between different datasets. Now similar to the discretized data, the MLFS can be approximated in a Hilbert basis,

$$\hat{\mathbf{h}}^i = \phi \mathbf{c}^i.$$

where $\phi = [\phi_1, \phi_2, \dots, \phi_d]$ is the $m \times d$ matrix of the orthonormal basis of \mathcal{H} . Let $\tilde{\mathbf{c}}^i = [c_1^i, c_2^i, \dots, c_d^i]^T$ be the $d \times 1$ vector of coefficients associated with the approximation $\hat{\mathbf{g}}^i$ of \mathbf{g}^i . We can now define $\mathbf{c}^i = [\tilde{c}_1^i, \tilde{c}_2^i, \dots, \tilde{c}_q^i]$ as the $d \times q$ matrix of coefficients associated with the approximation $\hat{\mathbf{h}}^i$ of \mathbf{h}^i .

Getting the best approximation to $\hat{\mathbf{h}}^i$ requires minimizing

$$\min_{\mathbf{c}} \|\mathbf{h} - \phi \mathbf{c}\|_F^2 = \min_{\mathbf{c}} \|\mathbf{h} - \hat{\mathbf{h}}\|_F^2, \quad (6)$$

where $\|\cdot\|_F$ is the Frobenius norm. For any matrix $X \in \mathbb{R}^{n \times m}$, $\|X\|_F^2 = \text{trace}\{X^T X\}$ and $\text{trace}\{A\} = a_{11} + a_{22} + \dots + a_{mm}$. But since CDIF is trying to find the best approximation for the data while making sure it looks more like the data in its class and less like the data in the other classes, we also seek to minimize

$$\sum_{j \text{ s.t. } y_i = y_j} M_{ij} \|\hat{\mathbf{h}}^i - \hat{\mathbf{h}}^j\|_F^2 \quad (7)$$

and to maximize

$$\sum_{j \text{ s.t. } y_i \neq y_j} M'_{ij} \|\hat{\mathbf{h}}^i - \hat{\mathbf{h}}^j\|_F^2 \quad (8)$$

Algorithm 1 Functional Classification by Discriminative Interpolation with Features (CDIF)

Training**Input:** $\mathbf{h}_{\text{train}} \in \mathbb{R}^{N \times m}$, $\mathbf{y}_{\text{train}} \in \mathbb{R}^N$, λ , μ , k , η (stepsize)**Output:** \mathbf{c}^i (optimal interpolation)

1. **For** $i \leftarrow 1$ to N
 2. **Repeat**
 3. Find $\mathcal{N}(i)$ and $\mathcal{M}(i)$ using k NN
 4. Compute M_{ij} and $M'_{ij} \forall i, j$ pairs
 5. Compute \mathbf{c}^i (eq. 10)
 6. **Until** convergence
 7. **End For**
-

Testing**Input:** $\mathbf{h}_{\text{test}} \in \mathbb{R}^{L \times m}$, \mathbf{c}^i (from training), λ , μ , k , η' (stepsize)**Output:** \hat{a} (labels for all testing data)

1. **For** $l \leftarrow 1$ to L
 2. **For** $a \leftarrow 1$ to A
 3. **Repeat**
 4. Find $\mathcal{N}(a)$ for $\tilde{\mathbf{c}}^l$ using k NN
 5. Compute $M_i^a \forall i$
 6. Compute $\tilde{\mathbf{c}}^l$ (eq. 12)
 7. **Until** convergence
 8. compute E_a^l (eq. 11)
 9. **End For**
 10. $\hat{a}^l \leftarrow \{a | \min E_a^l \forall a\}$
 11. **End For**
-

where $M_{ij} \in (0, 1)$ and $\sum_j M_{ij} = 1$ is the nearest neighbor constraint for $y_i = y_j$ where $j \neq i$, similarly, $M'_{ij} \in (0, 1)$ and $\sum_j M'_{ij} = 1$ is the nearest neighbor constraint for $y_i \neq y_j$ where $j \neq i$. Similarly, as shown in eq. 3, given an orthonormal basis, $\|\hat{\mathbf{f}}^i - \hat{\mathbf{f}}^j\|^2 = \|\mathbf{c}^i - \mathbf{c}^j\|^2$, this can be generalized to any Frobenius norm $\|\hat{\mathbf{f}}^i - \hat{\mathbf{f}}^j\|_F^2 = \|\mathbf{c}^i - \mathbf{c}^j\|_F^2$. Combining this result and the objectives in eq. 6 and eq. 7 yields the following objective function and optimization problem:

$$\min_{\mathbf{c}, M, M'} E = \min_{\mathbf{c}, M, M'} \sum_{i=1}^N \|\mathbf{h}^i - \phi \mathbf{c}^i\|_F^2 + \lambda \sum_{i,j \text{ s.t. } y_i = y_j} M_{ij} \|\mathbf{c}^i - \mathbf{c}^j\|_F^2 - \mu \sum_{i,j \text{ s.t. } y_i \neq y_j} M'_{ij} \|\mathbf{c}^i - \mathbf{c}^j\|_F^2. \quad (9)$$

An update equation for the objective in eq. 9 can be found by taking the gradient with respect to \mathbf{c}^i and then setting it to zero. This yields a closed form solution

$$\mathbf{c}^i = (\Phi^T \Phi + \lambda(1 + d_i)I - \mu(1 + b_i)I)^{-1} \dots \quad (10)$$

$$\left(\Phi^T \mathbf{h}^i + \lambda \left(\sum_{v \in C_a} M_{iv} \mathbf{c}^v + \sum_{w \in C_a} M_{wi} \mathbf{c}^w \right) - \mu \left(\sum_{s \notin C_a} M'_{is} \mathbf{c}^s + \sum_{r \notin C_a} M'_{ri} \mathbf{c}^r \right) \right)$$

where $d_i = \sum_{w \in C_a} M_{wi}$ s.t. $M_{wi} \in \{0, 1\}$ and $b_i = \sum_{r \notin C_a} M'_{ri}$ s.t. $M'_{ri} \in \{0, 1\}$, \mathbf{c}^v are the neighbors of curve i from the same class, \mathbf{c}^s are the neighbors of curve i from a different class and \mathbf{c}^w are the curves that consider i to be their neighbor from the same class, \mathbf{c}^r are the curves that consider i to be their neighbor from a different class.

3.4 Testing Formulation

Once we have trained our algorithm and found the coefficients that give the best approximation to our training curves, we use the testing data to evaluate the performance of our algorithm. Similar to the CDI training algorithm, we minimize an objective function that approximates the best wavelet reconstruction of the data and simultaneously updates the neighborhood for each class. This allows us to try and fit a curve to the best class. The objective function is formalized as

$$\hat{a} = \arg \min_a E_a = \arg \min_a \left(\min_{\hat{\mathbf{c}}} \|\hat{\mathbf{h}} - \Phi \hat{\mathbf{c}}\|_F^2 + \lambda \sum_i M_i^a \|\hat{\mathbf{c}} - \mathbf{c}^i\|_F^2 \right) \quad (11)$$

where $\hat{\mathbf{h}}$ is the test feature set function and $\hat{\mathbf{c}}$ is its matrix of reconstruction coefficients and \hat{a} is the assigned label to the incoming test pattern. M_i^a is the nearest neighbor in the set of class a patterns. As before, the membership is defined as $M_i^a \in \{0, 1\}$, again this can be solved by taking the derivative with respect to $\hat{\mathbf{c}}$ and yields a closed form solution

$$\hat{\mathbf{c}} = (\Phi^T \Phi + (\lambda d_a) I)^{-1} \left(\Phi^T \hat{\mathbf{h}} + \lambda \sum_{i \in C_a} M_i^a \mathbf{c}^i \right) \quad (12)$$

The testing formulation consists of two stages:

1. Solve $\min_{\hat{\mathbf{c}}} E_{\text{interp}}^a(\hat{\mathbf{c}})$ using block gradient descent and eq. (12).
2. Assign the label \hat{a} to $\hat{\mathbf{f}}$ by finding the class with the smallest value in (1) when using the feature set found in the training phase.

Details of the CDIF training and testing procedures are found in Algorithm 1.

4 Experimental Results

In this section, we discuss the performance of the *Classification by Discriminative Interpolation with Features* (CDIF) algorithm. The datasets used in our experiments are obtained from the “UCR Time Series Classification Archive” [9]. For a uniform comparison to other algorithms, the UCR database already divides the data into training and testing sets. All of these datasets had been sampled using a constant sampling rate. For a well rounded experimental dataset, we chose datasets with varying number of classes, different curve samplings, and uneven training and testing sizes. Information on each of the datasets can be found in Table 2. Through our empirical tests, we show two significant results. First, we show that our formulation CDIF is competitive with state-of-the-art algorithms, with leading competitors searching over 1,000 features compared to

Table 2. UCR Time Series Data Set with varying number of classes, training set size, testing set size and curve length.

Data Sets	Number of Classes	Training Set Size	Testing Set Size	Curve length
Beef	5	30	30	470
CBF2	3	30	900	128
Coffee	2	28	28	286
Face(all)	14	560	1690	131
Gun-Point	2	50	150	150
SwedishLeaf	15	500	625	128
SyntheticControl	6	300	300	60
TwoPatterns	4	1000	4000	128
Wafer	2	1000	6174	152
Yoga	2	300	3000	426

our eleven. Second, the current generalized CDIF methods performs much better than the original CDI algorithm which did not employ vector-valued functions.

The competing techniques we compare against begin with the baseline methods recommended for UCR: the Euclidean distance and Dynamic Time Warping (DTW) algorithm. We also pick four other leading algorithms [3, 4, 12, 13]. The first three—COTE, TSBF, and FBL—are discussed in the related works. Since COTE greedily searches over hundreds of features and uses ensemble classifiers, the training takes multiple hours. On the other hand, our CDIF method takes minutes to train and test. As detailed earlier, the TSBF training algorithm is said to be computationally complex, but once the model is trained, classifying the test set is very fast. FBL extracts thousands of features by using an extensive database of algorithms. After it computes thousands of features for each dataset, it learns the most informative of the classes by using a greedy forward feature selection with a linear classifier. Our classifier is inherently nonlinear and uses far less features. Grabcocka *et al.* [13] introduce a novel approach to shapelet classification called *Learning Shapelets* (LS) for time series data. They claim it has a much more significant classification accuracy than many state of the art algorithm as well as being faster to implement.

The results of our experiments in comparison with these algorithms can be found in Table 3. We also compare our CDIF method against the original CDI method, which does not utilize vector-value feature functions. The free parameters in our method are the push λ and pull μ regularizing parameters, the nearest neighbor parameter k . The λ and μ parameters typically range in values $[0.1, 10]$ and $[0, .05]$, respectively. While k values of $\{1, 2, 3\}$ consistently produce high classification rates.

The results clearly demonstrate the value of the proposed CDIF method. Our classification accuracies are competitive in all datasets, except SwedishLeaf. We perform better than the previous CDI method in all experiments. Methods like COTE and FBL employ ensemble classifiers and potentially thousands of features to achieve their accuracies. CDIF only utilized the eleven features detailed in Section 3.2 and our method is inherently multiclass and nonlinear, without

Table 3. CDIF Experimental Results comparable with state-of-the-art results.

Data Set	ED [9]	DTW [9]	COTE [3]	TSBF [4]	LS [13]	FBL [12]	CDI [14]	CDIF
Beef	66.7	66.7	86.7	71.3	76.0	56.7	70.0	96.7
CBF2	85.2	99.6	99.9	99.1	99.4	71.1	82.1	95.9
Coffee	100	100	100	99.6	100	100	100	100
Face(all)	71.4	80.6	89.5	76.6	78.2	70.8	75.6	86.8
Gun-Point	91.3	91.3	93.0	98.9	100	92.7	90.0	97.3
SwedishLeaf	78.9	84.6	95.4	92.5	91.3	77.3	76.0	83.0
SyntheticControl	99.3	98.3	99.9	99.2	99.3	96.3	98.7	99.0
TwoPatterns	90.7	100	100	94.7	99.7	92.6	88.7	90.4
Wafer	99.5	99.5	100	99.6	99.6	100	92.3	94.5
Yoga	83.0	83.6	88.7	85.1	85.0	77.4	81.5	81.8

the need to aggregate binary classification results. Even without some of these advantages, CDIF performed well against these techniques.

5 Conclusion

The proposed *Classification by Discriminative Interpolation with Features* (CDIF) framework leverages class-specific neighborhood relationships to discriminatively interpolate functions in a manner that morphs curves from the same class to become more similar in their appearance, while simultaneously pushing away neighbors from competing classes. CDIF formulates a vector-valued generalization of the previous work in [14]. The component functions of the vector-valued representation are obtained from localized features extracted from the original time series function. CDIF takes advantage of the interpolation characteristics of vector-valued functions to adaptively warp functions through their basis coefficients and according to their class labels.

Our CDIF framework can inherently handle multi-class problems, avoiding the *ad-hoc* one-vs-all heuristic employed strictly by binary classifiers. Our experimental evaluation demonstrated competitive performance to contemporary techniques. The results on a number of time-series, multi-class benchmark data sets, had CDIF ranked in the top tier among other approaches.

The present work illustrates how vector-valued functions and their interpolation property can be capitalized for functional data classification. We plan to explore several possible extensions using other functional characteristics and application to higher dimensional functional data like images. We are also going to continue exploring improvements in the optimization framework.

References

1. Abraham, C., Cornillon, P.A., Matzner-Lober, E., Molinari, N.: Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics* 30, 581–595 (2003)
2. Alonso, A.M., Casado, D., Romo, J.: Supervised classification for functional data: A weighted distance approach. *Computational Statistics & Data Analysis* 56, 2334–2346 (2012)

3. Bagnall, A., Lines, J., Hills, J., Bostrom, A.: Time-series classification with COTE: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering* 27, 2522–2535 (2015)
4. Baydogan, M.G., Runger, G., Tuv, E.: A bag-of-features framework to classify time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 2796–2802 (2013)
5. Berline, A., Biau, G., Rouviere, L.: Functional supervised classification with wavelets. *Annale de l'Institut de Statistique de l'Université de Paris* 52 (2008)
6. Biau, G., Bunea, F., Wegkamp, M.: Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory* 51, 2163–2172 (2005)
7. Boashash, B.: *Time-Frequency Signal Analysis and Processing: A Comprehensive Review*. Academic Press (2013)
8. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Wadsworth and Brooks (1984)
9. Chen, Y., Keogh, E., Hu, B., Begum, N., Bagnall, A., Mueen, A., Batista, G.: The UCR time series classification archive. www.cs.ucr.edu/~eamonn/time_series_data/ (2015)
10. Daubechies, I.: *Ten Lectures on Wavelets*. Regional Conference Series in Applied Mathematics, Society of Industrial and Applied Mathematics (1992)
11. Devroye, L., Györfi, L., Lugosi, G.: *A Probabilistic Theory of Pattern Recognition*. Springer New York (1996)
12. Fulcher, B.D., Jones, N.S.: Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering* 26, 3026–3037 (2014)
13. Grabocka, J., Schilling, N., Wistuba, M., Schmidt-Thieme, L.: Learning time-series shapelets. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1, 392–401 (2014)
14. Haber, R., Rangarajan, A., Peter, A.M.: Discriminative interpolation for classification of functional data. *Joint European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* 1, 20–36 (2015)
15. Kreyszig, E.: *Introductory Functional Analysis with Applications*. John Wiley and Sons (1978)
16. Leng, X., Müller, H.G.: Classification using functional data analysis for temporal gene expression data. *Bioinformatics* 22, 68–76 (2006)
17. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 971–987 (2002)
18. Pigoli, D., Sangalli, L.M.: *Wavelets Smoothing for Multidimensional Curves*. Physica-Verlag HD (2011)
19. Ramsay, J., Silverman, B.: *Functional Data Analysis*. Springer, 2nd edn. (2005)
20. Tarlow, D., Swersky, K., Charlin, L., Sutskever, I., Zemel, R.S.: Stochastic k -neighborhood selection for supervised and unsupervised learning. *Proceedings of the 30th International Conference on Machine Learning* 28, 199–207 (2013)
21. Weinberger, K.Q., Saul, L.K.: Fast solvers and efficient implementations for distance metric learning. *Proceedings of the 25th International Conference on Machine Learning* 1, 1160–1167 (2008)

A Modified EM Algorithm for Parameter Estimation in Linear Models with Time-Dependent Autoregressive and t-Distributed Errors

Boris Kargoll¹, Mohammad Omidalizarandi¹, Hamza Alkhatib¹, and Wolf-Dieter Schuh²

¹ Leibniz Universität Hannover, Geodätisches Institut
Nienburger Str. 1, 30167 Hannover, Germany
{kargoll,zarandi,alkhatib}@gih.uni-hannover.de

² Rheinische Friedrich-Wilhelms-Universität Bonn,
Institut für Geodäsie und Geoinformation
Nussallee 17, 53115 Bonn, Germany
schuh@geod.uni-bonn.de

Abstract. We derive an expectation conditional maximization either (ECME) algorithm for estimating jointly the parameters of a linear regression model, of a time-variable autoregressive (AR) model with respect to the random deviations, and of a scaled t-distribution with respect to the white noise components. This algorithm is shown to take the form of iteratively reweighted least squares in the estimation of the parameters both of the regression and time-variability model. The fact that the degree of freedom of that distribution is also estimated turns the algorithm into a partially adaptive estimator. As low degrees of freedom correspond to heavy-tailed distributions, the estimator can be expected to be robust against outliers. It is shown that the initial stabilization phase of an accelerometer on a shaker table can be modeled parsimoniously and robustly by a Fourier series with AR errors for which the time-variability model is defined by cubic polynomials.

Keywords: Linear regression model, time-dependent AR process, partially adaptive estimation, robust parameter estimation, EM algorithm, iteratively reweighted least squares, scaled t-distribution

1 Introduction

Linear regression models are used in many fields of application to approximate numerical measurement results by means of parametric functions. In practice, the random deviations of the observables from these deterministic functions are frequently correlated. In the context of a time series measured by a single sensor, autocorrelations can be expected for instance as a consequence of calibration corrections being applied to all of the measurements (cf. [20]). The resulting colored

noise is often modeled parametrically by means of covariance-stationary autoregressive (AR) or, more generally, by autoregressive moving average (ARMA) models. In the context of geodesy, for instance, such models were estimated to describe the colored noise of the Gravity and Gravity Field and Steady-State Ocean Circulation Explorer (GOCE) satellite gravity gradiometer (see [31]), of inertial sensors (see [37], [29], [28]), and within global navigation satellite system (GNSS) data (cf. [23]). The idea of fusing a linear regression model with an AR error model has been known at least since [14].

Stationary colored noise models are frequently found to be insufficient due to time-variable effects acting on the sensor, the environment or the observed phenomenon. To overcome this limitation, AR(MA) processes with time-dependent coefficients were introduced and methods for their estimation investigated by [34] and [19]. If colored noise is to be removed from the measurements, the estimated AR(MA) process must be invertible. Invertibility conditions for time-variable ARMA processes were formulated by [12]. Many different schemes for modeling the time-variability of AR(MA) processes have been proposed. For instance, [16] used a stochastically perturbed difference equation constraint model to ensure smoothness of estimated time-variable AR processes. Another stochastic approach is based on the formulation of a time-variable AR process as a state-space model, leading to a Kalman filter (see [35] and the references therein). Furthermore, the modeling of AR models with coefficients changing throughout different regimes in the sense of a Markov chain was considered by [7]; see also the comparative study [1].

The usual approach to modeling the time-variability of AR(MA) coefficients is to assume a certain set of basis functions and to express a particular AR(MA) coefficient, at every time instance, as a point on the best-fitting linear combination of the basis functions. For instance, polynomials in terms of truncated power series [4], Legendre polynomials [30], wavelets [36], trigonometric [11, 6], sigmoid functions [10], and discrete prolate spheroidal sequences [8] have been used for this purpose. The previous studies made effective use of least squares techniques for the purpose of parameter estimation. When the white noise error component of the AR(MA) process is expected to be outlier-afflicted or heavy-tailed, a robust estimator should be used instead. Probability distributions that take care of both of these issues are found (amongst others) within the family of scaled t-distributions. When their degree of freedom is estimated alongside the regression or AR(MA) model parameters, one speaks then of partially adaptive estimation.

On the one hand, partially adaptive estimation for linear regression models with t-distributed random deviations was suggested by [18]. In their approach, an EM algorithm was used for the purpose of maximum likelihood (ML) estimation, which takes the form of numerically convenient iteratively reweighted least squares (as already indicated by [5]). [27] and [21] suggested expectation conditional maximization (ECM), expectation conditional maximization *either* (ECME) and multicycle ECM as variants of EM to speed up convergence. On the other hand, [3] carried out a Bayesian type of partially adaptive estimation

of pure AR processes with t-distributed errors. [32] considered even a time-dependent AR process with t-distributed innovations, but did not estimate the degree of freedom. It should be mentioned that [25] and [24] introduced partially adaptive estimation, respectively, for linear regression and ARMA models in connection with a generalized t-distribution, which however does not seem to allow for the development of an EM algorithm in the spirit of [18].

Partially adaptive estimation of linear regression models with autoregressive random deviations and t-distributed white noise component appears to have been investigated first in [15]. In the following, this model is further extended to include an additional component that allows for time-variability of the AR coefficients. We choose for this purpose the aforementioned approach based on basis functions. After defining the specifics of the observation model, we derive a corresponding ECME algorithm, showing in particular that the coefficients of the regression model and the coefficients of the AR model can be estimated via two separate iteratively reweighted least squares schemes. This algorithm is applied to a measured time series of accelerometer data in a vibration analysis experiment. It is shown that the initial stabilization phase of the induced vibration can be modeled efficiently and estimated robustly by using a combination of a low-order Fourier series and a low-order, time-variable AR process with rather heavily tailed, t-distributed white noise components.

2 The Observation Model

The basic time series model consists of n linear observation equations

$$Y_t = \mathbf{A}_t \boldsymbol{\xi} + E_t \quad (t = 1, \dots, n), \quad (1)$$

where Y_t represents an observable, $\mathbf{A}_t \boldsymbol{\xi}$ a purely deterministic functional model, and E_t a random deviation. The observables and random deviations are collected in corresponding $(n \times 1)$ -random vectors \mathbf{Y} and \mathbf{E} . We assume that the functional model includes m unknown parameters $\boldsymbol{\xi} = [\xi_1, \dots, \xi_m]^T$ and that the row vectors $\mathbf{A}_1, \dots, \mathbf{A}_n$ give rise to a known $(n \times m)$ -coefficient matrix \mathbf{A} with full rank m . Furthermore, we assume that the random deviations are autocorrelated via a time-dependent p -th order autoregressive (AR) model

$$E_t = \alpha_{1,t} E_{t-1} + \dots + \alpha_{p,t} E_{t-p} + U_t \quad (t = 1, \dots, n), \quad (2)$$

where U_1, \dots, U_n are independently and identically distributed random variables with mean 0 and variance σ_0^2 . The time variability of each of the p AR coefficients will be described by linear models

$$\alpha_{j,t} = \mathbf{X}_t \boldsymbol{\beta}_j \quad (j = 1, \dots, p; t = 1, \dots, n) \quad (3)$$

involving a $(q \times 1)$ -vector of unknown parameters $\boldsymbol{\beta}_j = [\beta_{1,j}, \dots, \beta_{q,j}]^T$ and known coefficient vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$. Thus, we employ the same family of functions for all p AR coefficients. We consider for the distribution of the white noise

components U_1, \dots, U_n the scaled t-distribution $U_t \sim t_\nu(0, \sigma^2)$ with generally unknown degree of freedom ν and unknown scale parameter σ . That family of distributions is defined by the (family of) probability density functions (pdf)

$$f(u_t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\sigma\Gamma\left(\frac{\nu}{2}\right)} \left[1 + \left(\frac{u_t}{\sigma}\right)^2 / \nu\right]^{-\frac{\nu+1}{2}} \quad (t = 1, \dots, n), \quad (4)$$

where Γ is the gamma function. Due to the stochastic independence of the white noise components, their joint pdf factorizes into $f(\mathbf{u}) = \prod_{t=1}^n f(u_t)$. To fix the initial conditions of the AR(p) model (2), we assume that E_0, \dots, E_{1-p} take a constant value of 0. Assuming in addition that this model can be inverted, we can write in view of (1) – (3) for the t -th realization of U_t

$$u_t = e_t - \alpha_{1,t}e_{t-1} - \dots - \alpha_{p,t}e_{t-p} \quad (5)$$

$$= (y_t - \mathbf{A}_t\boldsymbol{\xi}) - \mathbf{X}_t\boldsymbol{\beta}_1(y_{t-1} - \mathbf{A}_{t-1}\boldsymbol{\xi}) - \dots - \mathbf{X}_t\boldsymbol{\beta}_p(y_{t-p} - \mathbf{A}_{t-p}\boldsymbol{\xi}) \quad (6)$$

$$= \left(y_t - \sum_{j=1}^m A_{t,j}\xi_j\right) - \sum_{k=1}^q X_{t,k}\beta_{k,1} \left(y_{t-1} - \sum_{j=1}^m A_{t-1,j}\xi_j\right) \\ - \dots - \sum_{k=1}^q X_{t,k}\beta_{k,p} \left(y_{t-p} - \sum_{j=1}^m A_{t-p,j}\xi_j\right), \quad (7)$$

setting the initial conditions $y_0 = \dots = y_{1-p} = 0$ and $\mathbf{A}_0 = \dots = \mathbf{A}_{1-p} = \mathbf{0}_{[1 \times m]}$. Using the notation $L^j \mathbf{Z}_t := \mathbf{Z}_{t-j}$ in connection with $\boldsymbol{\alpha}_t(L) := 1 - \alpha_{1,t}L - \dots - \alpha_{p,t}L^p$ and $\bar{\mathbf{Z}}_t = \boldsymbol{\alpha}_t(L)\mathbf{Z}_t$ for an arbitrary sequence of matrices $(\mathbf{Z}_t)_{t \in T}$ with $T \subseteq \mathbb{Z}$, we also have for every $t = 1, \dots, n$

$$u_t = \bar{e}_t = \boldsymbol{\alpha}_t(L)e_t = \boldsymbol{\alpha}_t(L)(y_t - \mathbf{A}_t\boldsymbol{\xi}) = \bar{y}_t - \bar{\mathbf{A}}_t\boldsymbol{\xi}. \quad (8)$$

This enables an interpretation of the quantities \bar{e}_t , \bar{y}_t and $\bar{\mathbf{A}}_t$ as the outputs of the digital filter $\boldsymbol{\alpha}_t(L)$, applied respectively to a segment of the random deviations \mathbf{e} , of the observations \mathbf{y} and of the coefficient matrix \mathbf{A} . Thus, $\boldsymbol{\alpha}_1(L), \dots, \boldsymbol{\alpha}_n(L)$ may be viewed as turning the colored noise sequence \mathbf{e} progressively into white noise \mathbf{u} , acting thus jointly as a *decorrelation filter*. Equations (4) and (7) define the basic probabilistic and parametric observation model. As $f(\mathbf{u})$ is actually a function of the observations \mathbf{y} , depending also on the values of all model parameters, we could use this joint pdf to define the likelihood function $\mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2, \nu; \mathbf{y})$ for the purpose of parameter estimation. Note that this function is conditional on the previously fixed values for E_0, \dots, E_{1-p} ; the use of such a conditional likelihood function (cf. [13]) is justified if the number of observations is sufficiently large in order for the 'warm-up effect' of the initial conditions on the subsequent autoregressive values to fade out.

Since maximum likelihood (ML) estimation based on the preceding likelihood function (or on its natural logarithm) cannot be based on closed-form expressions due to the intricacy of the t-distribution, we apply a well-known latent-variables approach (see [18]), which will enable ML estimation by means of a relatively

simple form of expectation maximization (EM) algorithm also for our specific time series model. The general idea is to firstly introduce independently and identically gamma-distributed latent variables $W_t \sim G(\nu/2, \nu/2)$ ($t = 1, \dots, n$), where ν is the degree of freedom of the desired t-distribution $t_\nu(0, \sigma^2)$. This distribution is defined by the pdf

$$f(w_t) = \begin{cases} \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \cdot w_t^{\frac{\nu}{2}-1} \cdot e^{-\frac{\nu}{2}w_t} & \text{if } w_t > 0, \\ 0 & \text{if } w_t \leq 0, \end{cases} \quad (9)$$

and the stochastic independence allows us to factorize $f(\mathbf{w}) = \prod_{t=1}^n f(w_t)$. Secondly, instead of assuming the white noise components to follow a scaled t-distribution at the outset, it is assumed that each random variable U_t follows a normal distribution conditional on the occurrence of the value w_t of the latent variable W_t . More specifically, we choose for the conditional pdf

$$f(u_t|w_t) = \frac{1}{\sqrt{2\pi(\sigma/\sqrt{w_t})^2}} \exp\left\{-\frac{u_t^2}{2(\sigma/\sqrt{w_t})^2}\right\}, \quad (10)$$

where each U_t is assumed to be conditionally independent from $U_1, W_1, \dots, U_{t-1}, W_{t-1}, U_{t+1}, W_{t+1}, \dots, U_n$ and W_n . In other words, the values of the latter random variables shall not affect the density of u_t , in the sense that

$$f(u_t|u_1, w_1, \dots, u_{t-1}, w_{t-1}, u_{t+1}, w_{t+1}, \dots, u_n, w_n, w_t) = f(u_t|w_t). \quad (11)$$

This form of conditional independence can be interpreted as a hidden Markov property for which the hidden variables are real-valued (see Section 4 in [2]). In light of (10), we now see that the variance of every white noise component U_t is rescaled by an unknown (latent) weight w_t , independently of the white noise components and weights associated with time instances other than t . We obtain then from (9) and (10) the joint pdf $f(u_t, w_t) = f(w_t) f(u_t|w_t)$, which in turn gives the desired pdf (4) of the scaled t-distribution as a marginal distribution (see Sect. 2.6 in [26]). This joint pdf also yields $f(w_t, u_t) = f(u_t) f(w_t|u_t)$, where the conditional pdf $f(w_t|u_t)$ can be shown to define the gamma distribution $G(a, b)$ with parameters $a = (\nu + 1)/2$ and $b = (\nu + u_t^2/\sigma^2)/2$, given the value u_t (cf. [17], equations (27)). In connection with the initially made two assumptions of conditional independence, this allows us to establish the joint pdf of the white noise and the latent weights in the form of the factorization $f(\mathbf{u}, \mathbf{w}) = \prod_{t=1}^n f(w_t) f(u_t|w_t)$, which we define to be the likelihood function $\mathcal{L}(\xi, \beta, \sigma^2, \nu; \mathbf{y}, \mathbf{w})$ of the extended observation model.

3 The Modified EM Algorithm

Combining the preceding pdf with (8) – (10), we can write the log-likelihood function in the form

$$\begin{aligned} \log \mathcal{L}(\xi, \beta, \sigma^2, \nu; \mathbf{y}, \mathbf{w}) = & -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + \frac{n\nu}{2} \log\left(\frac{\nu}{2}\right) - n \log \Gamma\left(\frac{\nu}{2}\right) \\ & - \frac{1}{2} \sum_{t=1}^n \log w_t - \frac{1}{2\sigma^2} \sum_{t=1}^n w_t [\alpha_t(L)(y_t - \mathbf{A}_t \xi)]^2 + \frac{\nu}{2} \sum_{t=1}^n (\log w_t - w_t), \end{aligned} \quad (12)$$

where each $\alpha_t(L)$ -filter is a function of the parameters β . Collecting for brevity of expressions all unknown parameters ξ , β , σ^2 and ν within the vector θ and following the idea of expectation maximization (EM) in [5], we iteratively aim for a solution $\theta^{(i+1)}$ that maximizes $E_{\mathbf{Y}, \mathbf{W}|\mathbf{y}; \theta^{(i)}} \{\log \mathcal{L}(\theta; \mathbf{y}, \mathbf{W})\}$. Here, $i \in \{0, 1, 2, \dots\}$ denotes the iteration step within the EM algorithm, so that the conditional expectation is evaluated by using both the given measurements \mathbf{y} and the parameter estimates $\theta^{(i)}$ from the preceding iteration step.

3.1 The E-Step

Following the general approach by [5], we restate the conditional expectation as the Q -function (see also [9])

$$Q(\theta|\theta^{(i)}) = E_{\mathbf{W}|\mathbf{y}; \theta^{(i)}} \{\log \mathcal{L}(\theta; \mathbf{y}, \mathbf{W})\}. \quad (13)$$

As the likelihood function was previously defined in terms of white noise \mathbf{U} rather than the observables \mathbf{Y} , we will condition directly on the realizations \mathbf{u} . We then find with (12), in analogy both to the pure regression case without AR models in [18] and to the regression model with time-constant AR models in ([15]), the Q -function to be

$$\begin{aligned} Q(\theta|\theta^{(i)}) = & -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + \frac{n\nu}{2} \log\left(\frac{\nu}{2}\right) - n \log \Gamma\left(\frac{\nu}{2}\right) \\ & - \sum_{t=1}^n \frac{1}{2} \left[\nu + \left(\frac{u_t}{\sigma}\right)^2 \right] E_{\mathbf{W}|\mathbf{u}; \theta^{(i)}} \{W_t\} + \sum_{t=1}^n \frac{1}{2} (\nu - 1) E_{\mathbf{W}|\mathbf{u}; \theta^{(i)}} \{\log W_t\} \end{aligned} \quad (14)$$

with

$$w_t^{(i)} := E_{\mathbf{W}|\mathbf{u}; \theta^{(i)}} \{W_t\} = \frac{\nu^{(i)} + 1}{\nu^{(i)} + \left(\frac{\alpha_t^{(i)}(L)(y_t - \mathbf{A}_t \xi^{(i)})}{\sigma^{(i)}} \right)^2} \quad (15)$$

and (employing the digamma function ψ)

$$E_{\mathbf{W}|\mathbf{u}; \theta^{(i)}} \{\log W_t\} = \log w_t^{(i)} + \psi\left(\frac{\nu^{(i)} + 1}{2}\right) - \log\left(\frac{\nu^{(i)} + 1}{2}\right). \quad (16)$$

Substitution of the findings (15) and (16) into (14) gives us now

$$\begin{aligned} Q(\theta|\theta^{(i)}) = & \text{const.} - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^n w_t^{(i)} [\alpha_t(L)(y_t - \mathbf{A}_t \xi)]^2 + \frac{n\nu}{2} \log \nu \\ & - n \log \Gamma\left(\frac{\nu}{2}\right) + \frac{n\nu}{2} \left[\psi\left(\frac{\nu^{(i)} + 1}{2}\right) - \log(\nu^{(i)} + 1) + \frac{1}{n} \sum_{t=1}^n (\log w_t^{(i)} - w_t^{(i)}) \right]. \end{aligned} \quad (17)$$

3.2 The M-Step

To carry out the M-Step, we determine the first partial derivatives of the Q -function (17) with respect to the individual parameters ξ , β , σ^2 and ν in θ , and set these equal to zero. It is not difficult to show that the first-order condition with respect to the j -th parameter in ξ becomes

$$0 = \frac{\partial}{\partial \xi_j} Q(\theta|\theta^{(i)}) = \frac{1}{\sigma^2} \sum_{t=1}^n w_t^{(i)} \bar{A}_{t,j} (\bar{y}_t - \bar{\mathbf{A}}_t \xi).$$

Writing these m equations in matrix notation, for which purpose we denote by $\mathbf{W}^{(i)}$ the diagonal matrix of the weights $w_1^{(i)}, \dots, w_n^{(i)}$, we obtain

$$\mathbf{0} = \begin{bmatrix} \bar{A}_{1,1} & \cdots & \bar{A}_{n,1} \\ \vdots & & \vdots \\ \bar{A}_{1,m} & \cdots & \bar{A}_{n,m} \end{bmatrix} \mathbf{W}^{(i)} \begin{bmatrix} \bar{y}_1 - \bar{\mathbf{A}}_1 \xi \\ \vdots \\ \bar{y}_n - \bar{\mathbf{A}}_n \xi \end{bmatrix} = \bar{\mathbf{A}} \mathbf{W}^{(i)} (\bar{\mathbf{y}} - \bar{\mathbf{A}} \xi).$$

As these normal equations for the parameter group ξ involve also the unknown parameters β through the filter operations, we fix values for the latter by setting $\beta = \beta^{(i)}$. In doing this, we perform a so-called conditional maximization (CM) step in the sense of [27]. Then, $\beta^{(i)}$ allows us to compute the time variable AR coefficients $\alpha_{1,1}^{(i)}, \dots, \alpha_{p,n}^{(i)}$ by means of the equations (3); these coefficients define decorrelation filters, which we can subsequently employ to calculate the filtered quantities (for every $t = 1, \dots, n$)

$$\bar{y}_t^{(i)} := \alpha_t^{(i)}(L)y_t, \quad \bar{A}_{t,j}^{(i)} := \alpha_t^{(i)}(L)A_{t,j}, \quad \bar{\mathbf{A}}_t^{(i)} := \alpha_t^{(i)}(L)\mathbf{A}_t. \quad (18)$$

The new solution $\xi^{(i+1)}$ for ξ can then be computed from

$$\xi^{(i+1)} = \left((\bar{\mathbf{A}}^{(i)})^T \mathbf{W}^{(i)} \bar{\mathbf{A}}^{(i)} \right)^{-1} (\bar{\mathbf{A}}^{(i)})^T \mathbf{W}^{(i)} \bar{\mathbf{y}}^{(i)}, \quad (19)$$

which estimates give rise to the colored noise residuals $e_t^{(i+1)} := y_t - \mathbf{A}_t \xi^{(i+1)}$. Next, we consider the first-order conditions with respect to the previously fixed parameter vectors β_1, \dots, β_p , for which we obtain

$$\mathbf{0} = \frac{\partial}{\partial \beta_h} Q(\theta|\theta^{(i)}) = \frac{1}{\sigma^2} \sum_{t=1}^n w_t^{(i)} e_{t-h} \mathbf{X}_t^T (e_t - \mathbf{X}_t \beta_1 e_{t-1} - \dots - \mathbf{X}_t \beta_p e_{t-p}).$$

Having already determined estimates $\xi^{(i+1)}$, the joint solution of the equation systems arising for all $h = 1, \dots, p$ can be determined as a second CM step via

$$\begin{aligned} \mathbf{0} &= \begin{bmatrix} e_0^{(i+1)} \mathbf{X}_1^T & \cdots & e_{n-1}^{(i+1)} \mathbf{X}_n^T \\ \vdots & & \vdots \\ e_{1-p}^{(i+1)} \mathbf{X}_1^T & \cdots & e_{n-p}^{(i+1)} \mathbf{X}_n^T \end{bmatrix} \mathbf{W}^{(i)} \begin{bmatrix} e_1^{(i+1)} - e_0^{(i+1)} \mathbf{X}_1 \beta_1 - \dots - e_{1-p}^{(i+1)} \mathbf{X}_1 \beta_p \\ \vdots \\ e_n^{(i+1)} - e_{n-1}^{(i+1)} \mathbf{X}_n \beta_1 - \dots - e_{n-p}^{(i+1)} \mathbf{X}_n \beta_p \end{bmatrix} \\ &=: (\mathbf{E}^{(i+1)})^T \mathbf{W}^{(i)} (\mathbf{e}^{(i+1)} - \mathbf{E}^{(i+1)} \beta), \end{aligned}$$

using the initial conditions $e_0^{(i+1)} = \dots = e_{1-p}^{(i+1)} = 0$ and the stacked vector $\boldsymbol{\beta}^T = [\boldsymbol{\beta}_1^T \dots \boldsymbol{\beta}_p^T]$. The reweighted least squares solution for $\boldsymbol{\beta}$ then reads

$$\boldsymbol{\beta}^{(i+1)} = \left((\mathbf{E}^{(i+1)})^T \mathbf{W}^{(i)} \mathbf{E}^{(i+1)} \right)^{-1} (\mathbf{E}^{(i+1)})^T \mathbf{W}^{(i)} \mathbf{e}^{(i+1)}. \quad (20)$$

For every time instance t , the resulting AR coefficients will be denoted by $\alpha_{j,t}^{(i+1)}$, and we write for the corresponding decorrelation filter $\boldsymbol{\alpha}_t^{(i+1)}(L)$, which allows us to estimate the white noise residuals through $u_t^{(i+1)} = \boldsymbol{\alpha}_t^{(i+1)}(L) e_t^{(i+1)}$. The third CM-Step applies to the scale factor of the underlying t-distribution and requires the solution of

$$0 = \frac{\partial}{\partial \sigma^2} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(i)}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=1}^n w_t^{(i)} u_t^2.$$

Substituting the current estimates $\boldsymbol{\xi}^{(i+1)}$ and $\boldsymbol{\beta}^{(i+1)}$, we obtain for this solution the average sum of squared residuals

$$(\sigma^2)^{(i+1)} = \frac{1}{n} \sum_{t=1}^n w_t^{(i)} \left(u_t^{(i+1)} \right)^2 = \frac{(\mathbf{u}^{(i+1)})^T \mathbf{W}^{(i)} \mathbf{u}^{(i+1)}}{n}. \quad (21)$$

The fourth CM step would then follow from solving the equation $0 = \frac{\partial}{\partial \nu} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(i)})$, completing the current step of the EM algorithm (in the form of ECM). In view of the findings of [21] and [22] (see also [26]) in the context of estimating the degree of freedom of the scaled t-distribution, the number of iteration steps can generally be reduced greatly by replacing the Q -function with the original log-likelihood function within the preceding first-order condition for ν . This modification of the ECM algorithm is called ECM *either* (ECME) and increases the likelihood in each iteration step as well. Applying this idea to our specific model (4) – (8), we thus seek the zero of the equation

$$0 = \frac{\partial}{\partial \nu} \log \mathcal{L}(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2, \nu; \mathbf{y}) = \frac{n}{2} \psi \left(\frac{\nu+1}{2} \right) - \frac{n}{2} \psi \left(\frac{\nu}{2} \right) + \frac{n}{2} (\log \nu + 1) - \frac{1}{2} \sum_{t=1}^n \log \left[\nu + \left(\frac{u_t}{\sigma} \right)^2 \right] - \frac{1}{2} (\nu + 1) \sum_{t=1}^n \left[\nu + \left(\frac{u_t}{\sigma} \right)^2 \right]^{-1}$$

Replacing \mathbf{u} by the currently available estimated residuals $\mathbf{u}^{(i+1)}$, denoting the solution for ν by $\nu^{(i+1)}$, and defining $w_t^{(i+1)}$ according to (15), we finally obtain

$$0 = \log \nu^{(i+1)} + 1 - \psi \left(\frac{\nu^{(i+1)}}{2} \right) + \psi \left(\frac{\nu^{(i+1)} + 1}{2} \right) - \log \left(\nu^{(i+1)} + 1 \right) + \frac{1}{n} \sum_{t=1}^n \left(\log w_t^{(i+1)} - w_t^{(i+1)} \right). \quad (22)$$

We conclude this section with a few comments on our implementation of the preceding ECME algorithm.

If initial values for ξ are unknown, they are computed via unweighted least squares through $\xi^{(0)} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$. Based on the resulting initial residuals $\mathbf{e}^{(0)}$, initial values for β are computed next via (20), applying again the neutral weight matrix $\mathbf{W}^{(0)} = \mathbf{I}$. This solution gives rise to the initial decorrelation filters $\alpha_t^{(0)}(L)$, which allow for the computation of the residuals $\mathbf{u}^{(0)}$. Subsequently, we determine the initial value for σ^2 through $(\sigma^2)^{(0)} = (\mathbf{u}^{(0)})^T \mathbf{u}^{(0)} / n$. Furthermore, we choose for the initial degree of freedom $\nu^{(0)} = 30$. With these values, the weight determination (15) within the E step and the CM(E) steps (19) – (22) are iterated until the maximum number of iterations is reached or until the following stop criterion is satisfied. We check whether the greatest absolute value of the differences between the estimates of two subsequent iteration steps is less than 10^{-8} for the parameters ξ , β and σ^2 , and less than 10^{-4} for ν . Since the normal distribution represents the limiting case $\nu \rightarrow \infty$ of the t-distribution, it is possible that the zero of (22) is infinite. To circumvent numerical problems created by this case, we check if a sign change of the function on the right-hand side of (22) occurs between 10^{-8} and 10^8 ; if not, we set $\nu^{(i+1)}$ to a very large value in correspondence to a normal distribution. It should be mentioned that we did not find it necessary in our real-data applications to enforce stability on the time-variable AR-processes, which issue is beyond the scope of the current paper.

4 An Application to Vibration Analysis

We applied the ECME algorithm to estimate the non-stationary behavior of a highly accurate single-axis PCB Piezotronics accelerometer within a vibration analysis experiment, which was carried out at the Institute of Concrete Construction at the Leibniz Universität Hannover. The sensor was mounted on a shaker table, which consists of a plexiglass plate fixed between two wooden supports and two imbalance motors in the center. This shaker induced an oscillation frequency of 16 Hz throughout the measurement period of approximately 45 minutes. The sampling frequency of the accelerometer was approximately 195 Hz, so that a maximum frequency of about 95 Hz can be detected. Usually, the first few seconds of the data set are discarded as transient oscillation. The data set without this initial stabilization phase was modeled in [15]. In the following, we analyze only that initial phase, which we defined to consist of the first 1500 accelerometer values (i.e., of the initial approximately 7.7 seconds). Apart from the main frequency, multiples of 8 Hz with small amplitudes can be expected to occur as a consequence of the sampling of the originally continuous-time phenomenon and due to the physical properties of the shaker table. We modeled this signal content by means of the truncated Fourier series

$$y_t = \frac{a_0}{2} + \sum_{j=1}^{12} a_j \cos(2\pi f_j x_t) + b_j \sin(2\pi f_j x_t) + e_t, \quad (t = 1, \dots, n) \quad (23)$$

with fixed frequencies $f_j = j \cdot 8$ Hz; the unknown Fourier coefficients a_0, a_1, \dots, a_{12} and b_1, \dots, b_{12} , are collected within the parameter vector ξ . Concerning the

colored noise, we specified on the one hand a time-variable $\text{AR}(p)$ -process using the global polynomials x^0, x^1, \dots, x^q , and on the other hand a time-constant $\text{AR}(p)$ -process (which constitutes the special case $q = 0$ of the preceding model). We tried different autoregressive and polynomial model orders, beginning with $p = q = 1$, and identified the least orders for which the estimated, decorrelation-filtered residuals $\hat{\mathbf{u}}$ (obtained as the values $\mathbf{u}^{(i+1)}$ after convergence of the ECME algorithm) pass a periodogram-based white noise test. We used for this purpose the MATLAB routine `periodogram` to compute the onesided periodogram I_1, \dots, I_M , which values give the normalized cumulated periodogram

$$S_0 = 0, \quad S_i = \frac{\sum_{k=1}^i I_k}{\sum_{k=1}^M I_k} \quad (i = 1, \dots, M),$$

where M is the lower integer of $n/2$. The test compares the maximum cumulated periodogram excess $T = \max_i |S_i - i/M|$ over a cumulated, theoretical white noise periodogram with $1 - \alpha$ significance bounds (cf. Sect. 7.3.3 in [33]). We thus obtained as the most parsimonious colored noise description a time-variable $\text{AR}(6)$ model with cubic polynomials ($q = 3$) and a time-constant $\text{AR}(21)$ model (see Fig. 1 for the depiction of the two periodogram tests and Fig. 2 for the estimated AR coefficients). The adjustment involving the time-variable AR

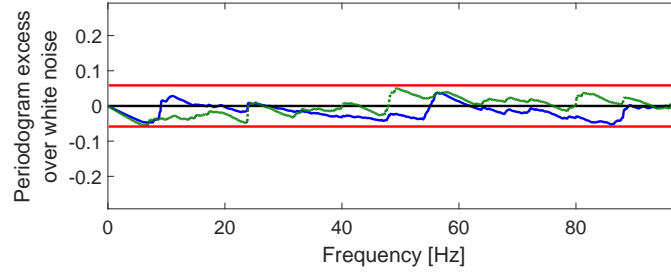


Fig. 1. Excess of the estimated periodogram of the decorrelated residuals for the time-variable $\text{AR}(6)$ model (blue) and for the stationary $\text{AR}(21)$ model (green) with respect to the theoretical white noise periodogram (black) and 99% significance bounds (red).

model yields for the estimated degree of freedom $\hat{\nu} = 4.8$ (indicating a rather heavy-tailed t-distribution), whereas we obtained the Gaussian limit $\hat{\nu} \rightarrow \infty$ for the time-constant model. The difference between these two models in terms of adjusted observations $\mathbf{A}\hat{\boldsymbol{\xi}}$ is also clearly discernible (see Fig. 3). Whereas the time-variable model reproduces the eventual oscillation amplitude quite accurately, much of the oscillation signal is absorbed into the colored noise residuals of the time-constant model. We therefore conclude that it is more reasonable to interpret and model the initial measurement phase, where the oscillation amplitude changes greatly before reaching a stable value, as a combination of outliers (leading to heavy tails) and non-stationary autocorrelation patterns interacting

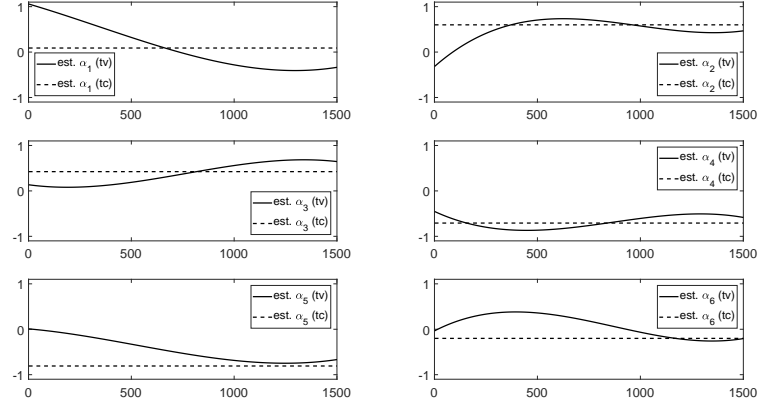


Fig. 2. Estimated coefficients of the time-variable (tv) AR(6) model and of the first six coefficients for the time-constant (tc) AR(21) model over the 1500 time instances.

with the Fourier series model. These patterns can be displayed as a time-variable power spectral density (see Fig. 4), defined by (cf. [35])

$$PSD(f, t) = \frac{\hat{\sigma}_u^2}{\left| 1 - \sum_{j=1}^p \hat{\alpha}_{j,t} e^{-i2\pi j f} \right|^2}, \quad (24)$$

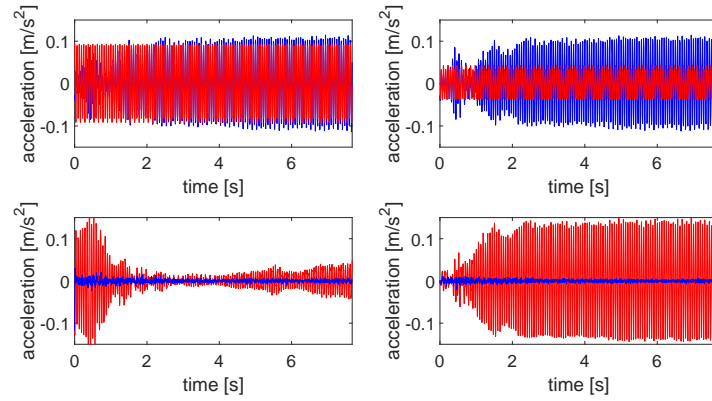


Fig. 3. Top row: plot of the complete dataset \mathbf{y} (blue), of the adjusted observations involving the time-variable (tv) AR(6) model (in red on the left subplot), and of the adjusted observations $\mathbf{A}\hat{\xi}$ involving the time-constant (tc) AR(21) model (in red on the right subplot). Bottom row: plots of the corresponding estimated residuals (decorrelation-filtered residuals $\hat{\mathbf{u}}$ in blue, colored noise residuals $\hat{\mathbf{e}}$ in red).

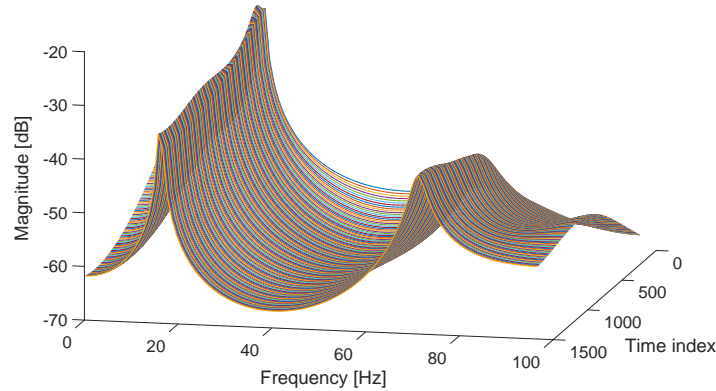


Fig. 4. Power spectral density based on the time-variable AR(6) processes.

where the standard deviation of the t-distributed white noise components is related to the estimated scale factor and degree of freedom via $\hat{\sigma}_u^2 = \frac{\hat{\nu}}{\hat{\nu}-2}\hat{\sigma}^2$. The evident fact that the PSDs have peaks around 16 Hz demonstrates that the oscillation signal is still partially captured by the colored noise model. As both the Fourier and the AR model have relationships with the frequency domain, this kind of interaction appears to be unavoidable.

Acknowledgments. The presented application of the PCB Piezotronics accelerometer within the vibration analysis experiment was performed as a part of the collaborative project "Spatio-temporal monitoring of bridge structures using low cost sensors" with ALLSAT GmbH, which is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) and the Central Innovation Programme for SMEs (ZIM Kooperationsprojekt, ZF4081803DB6). In addition, the authors would like to acknowledge the Institute of Concrete Construction (Leibniz Universität Hannover) for providing the shaker table and the reference accelerometer used within this experiment.

References

1. Azrak, R., Mélard, G.: AR Models with Time-Dependent Coefficients - a Comparison Between Several Approaches. Technical Report 0642, IAP Statistics Network, Interuniversity Attraction Pole (2006)
2. Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov models. International Computer Science Institute, Berkeley/California, http://lasa.epfl.ch/teaching/lectures/ML_PhD/Notes/GP-GMM.pdf (1989)
3. Christmas, J., Everson, R.: Robust Autoregression: Student-t Innovations Using Variational Bayes. IEEE Transactions on Signal Processing 59, 48-57 (2011)

4. Dahlhaus, R.: Fitting Time Series Models to Nonstationary Processes. *The Annals of Statistics* 25, 1-37 (1997)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society (Series B)* 39, 1-38 (1977)
6. Eom, K.B.: Analysis of Acoustic Signatures from Moving Vehicles Using Time-Varying Autoregressive Models. *Multidimensional Systems and Signal Processing* 10, 357-378 (1999)
7. Francq, C., Gautier, A.: Large Sample Properties of Parameter Least Squares Estimates for Time-Varying ARMA Models. *Journal of Time Series Analysis* 25, 765-783 (2004)
8. Grenier, Y.: Time-Dependent ARMA Modeling of Nonstationary Signals. *IEEE Transactions in Acoustics, Speech, and Signal Processing ASSP-31*, 899-911 (1983)
9. Gupta, M.R., Chen, Y.: Theory and Use of the EM Algorithm. *Foundations and Trends in Signal Processing* 4, 223-296 (2011)
10. Härmä, A., Juntunen, M., Kaipio, J.P.: Time-Varying Autoregressive Modeling of Audio and Speech Signals. In: *Proceedings of the 10th European Signal Processing Conference*, Tampere, Finland, 2037-2040 (2000)
11. Hall, M.G., Oppenheim, A.V., Willsky, A.S.: Time-Varying Parametric Modeling of Speech. *Signal Processing* 5, 267-285 (1983)
12. Hallin, M.: Mixed Autoregressive-Moving Average Multivariate Processes with Time-Dependent Coefficients. *Journal of Multivariate Analysis* 8, 567-572 (1978)
13. Hamilton, J.D.: *Time series analysis*. Princeton University Press (1994)
14. Hildreth, C.: Asymptotic Distribution of Maximum Likelihood Estimators in a Linear Model with Autoregressive Disturbances. *The Annals of Mathematical Statistics* 40, 583-594 (1969)
15. Kargoll, B., Omidalizarandi, M., Loth, I., Paffenholz, J.A., Alkhatib, H.: An Iteratively Reweighted Least-Squares Approach to Adaptive Robust Adjustment of Parameters in Linear Regression Models with Autoregressive and t-Distributed Deviations. *Journal of Geodesy*, <https://doi.org/10.1007/s00190-017-1062-6> (2017)
16. Kitagawa, G., Gersch, W.: A Smoothness Priors Time-Varying AR Coefficient Modeling of Nonstationary Covariance Time Series. *IEEE Transactions on Automatic Control* 30, 48-56 (1985)
17. Koch, K.R., Kargoll, B.: Expectation Maximization Algorithm for the Variance-Inflation Model by Applying the t-Distribution. *Journal of Applied Geodesy* 7, 217-225 (2013)
18. Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust Statistical Modeling Using the t-Distribution. *Journal of the American Statistical Association* 84, 881-896 (1989)
19. Liporace, L.A.: Linear Estimation of Nonstationary Signals. *The Journal of the Acoustical Society of America* 58, 1288-1295 (1975)
20. Lira, I., Wöger, W.: Comparison Between the Conventional and Bayesian Approaches to Evaluate Measurement Data. *Metrologia* 43, S249-S259 (2006)
21. Liu, C.H., Rubin, D.B.: ML Estimation of the t Distribution using EM and its Extensions, ECM and ECME. *Statistica Sinica* 5, 19-39 (1995)
22. Liu, C.H., Rubin, D.B.: The ECME Algorithm: a Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika* 81:633-648 (1994)
23. Luo, X., Mayer, M., Heck, B.: Analysing Time Series of GNSS Residuals by Means of AR(I)MA Processes. In: *VII Hotine-Marussi Symposium on Mathematical Geodesy*, International Association of Geodesy Symposia 137, pp. 129-134, Springer, Berlin, Heidelberg (2012)

24. McDonald, J.B.: Partially Adaptive Estimation of ARMA Time Series Models. *International Journal of Forecasting* 5, 217-230 (1989)
25. McDonald, J.B., Newey, W.K.: Partially Adaptive Estimation of Regression Models via the Generalized t Distribution. *Econometric Theory* 4, 428-457 (1988)
26. McLachlan, G.J., Krishnan, T.: *The EM Algorithm and Extensions*. John Wiley & Sons, Hoboken, New Jersey (2008)
27. Meng, X., Rubin, D.B.: Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* 80, 267-278 (1993).
28. Nassar, S., El-Sheimy, N.: Accuracy Improvement of Stochastic Modeling of Inertial Sensor Errors. *Zeitschrift für Geodäsie, Geoinformation und Landmanagement* 130, 146-155 (2005)
29. Park, M., Gao, Y.: Error and Performance Analysis of MEMS-based Inertial Sensors with a Low-cost GPS Receiver. *Sensors* 8, 2240-2261 (2008)
30. Rudoy, D., Quatieri, T.F., Wolfe, P.J.: Time-Varying Autoregressions in Speech: Detection Theory and Applications. *IEEE Transactions in Acoustics, Speech, and Signal Processing* 19, 977-989 (2011)
31. Schuh, W.D.: The Processing of Band-limited Measurements; Filtering Techniques in the Least Squares Context and in the Presence of Data Gaps. *Space Science Reviews* 108, 67-78 (2003)
32. Sanubari, J., Tokuda, K.: Non-Stationary Spectral Estimation Based on Robust Time Varying AR Model Excited by a t -Distribution Process. In: *Proceedings of the IEEE Conference on Speech and Image Technologies for Computing and Telecommunications (TENCON'97)*, 51-54 (1997)
33. Schlittgen, R., Streitberg, B.H.J.: *Zeitreihenanalyse*. 9th edition. R. Oldenbourg Verlag, Munich (2001)
34. Subba Rao, T.: The Fitting of Non-Stationary Time-Series Models with Time-Dependent Parameters. *Journal of the Royal Statistical Society (Series B)* 32, 312-322 (1970)
35. Tary, J.B., Herrera, R.H., van der Baan, M.: Time-Varying Autoregressive Model for Spectral Analysis of Microseismic Experiments and Long-Period Volcanic Events. *Geophys. J. Int.* 196, 600-611 (2014)
36. Tsatsanis, M.K., Giannakis, G.B.: Time-Varying System Identification and Model Validation Using Wavelets. *IEEE Transactions on Signal Processing* 41, 3512-3523 (1993)
37. Wang, K., Xiong, S., Li, Y.: Modeling with Noises for Inertial Sensors. In: *Position Location and Navigation Symposium (PLANS) 2012, IEEE/ION*, 625-632 (2012)

Copulas for Modeling the Relationship between the Inflation and the Exchange Rates

Laila Ait Hassou¹, Fadoua Badaoui*², Okou Guei Cyrille³, Amine Amar⁴,
Abdelhak Zoglat¹, and Elhadj Ezzahid⁵

¹ Laboratoire de Mathématiques, Statistique et Applications, Département de Mathématiques, Faculté des sciences, Université Mohammed V de Rabat.

² Département Statistique, Démographie et Actuariat, Institut National de Statistique et d'Economie Appliquée, Rabat.

³ Unité de Formation et Recherche (UFR) Environnement, Université Jean Lorougnon Guédé de Daloa, Côte d'Ivoire.

⁴ Moroccan Agency for Sustainable Energy (Rabat, Morocco)

⁵ Dpartement d'conomie, Universit Mohammed V. Rabat, Morocco.

Abstract. *Copulas are useful tools for formalizing the dependence structure between variables, especially for the economics field, where the dependence plays a key role. In this paper, we analyze the dependence between the inflation and the US/Euro exchange rates in the Euro area, from different periods, a crisis and a non-crisis periods. First, we analyze the dependence profiles using a non-parametric approach. In the second, we select an appropriate parametric Copulas, depending on the nature of the periods. Results confirm the sensibility of Copulas to the macroeconomic fluctuations, which occur during the analyzed periods.*

Keywords: Copulas, Forecasting, GoF tests, Inflation, Exchange rate, Non-parametric approaches.

1 Introduction

The economic literature presents the inflation as one of the key macroeconomic indicators. This concept is defined as persistent increases in the general level of the prices, which referred to the devaluation of the money worth (ILO, IMF, OECD, UNECE, Eurostat, and The World Bank, 2004). Concerning susceptible causes, the economic literature provides two main reasons of the inflation. The most common is the demand-pull inflation while the second is the cost-push inflation (Jongwanich and Park, 2008). The first one occurs when a demand for a good or a service increases much more than it outstrips supply. This situation can occur only in some circumstances, namely a growing economy which conduct people to be more confident, a discretionary fiscal policy where the government's ability to spend more or tax less increases the demand, in the situation of the marketing and a new technology deployment or, in the situation of over expansion of the money supply. The cost-push inflation is a result of a supply shortage combined with an enough demand. This can occur through a wage inflation, a monopoly, a natural disasters and a depletion of a natural resources,

a government regulation and a taxation or, through the currency exchange rates.

Economists highlight the dependence between the inflation and several key macroeconomic variables. Chollete and Ning (2009) mention a negative dependence between outputs and prices, Munyeka (2014) studies a positive dependence between the inflation and the real GDP, and Fitzgerald and Nicolini (2014) analyze a negative linear dependence between the unemployment and the inflation .

The exchange rate movements are one of the factors that influence the inflation. This indicator is of a great importance from the perspective of the monetary policy. The exchange rate is defined as the price of one country's' currency in relation to another. It may be expressed as the average rate for a period of a time or as the rate at the end of a period.

A variety of empirical studies and time series models show that there is a relationship between the exchange rate fluctuations and the inflation (Kano (2016), Burstein and Gopinath (2014) and Engel (2014)).

In fact, the exchange rate can influence the inflation, directly through the price of the imported final consumer goods, and indirectly via the price of the imported intermediate goods used in the Euro area domestic production. However, the effect on the inflation depends on what the causes of the exchange rate movements are. Moreover, the size and the speed of the exchange rate effects differ across the product categories and depends on the macroeconomic environment (ECB, 2014).

Most of the results about these macroeconomic dependencies are formulated with some variant of a covariance. However, covariances and correlations are not enough to identify the forms of a dependence. For this reason, the Copulas are introduced as useful extensions and generalizations of the approaches for modeling a joint distributions and a dependence (Sklar, 1959). There are many econometric studies which use a Copulas in an explicit manner. Granger et al. (2006) use the Copulas to examine common factors in a conditional distributions for the income and the consumption. Miller and Liu (2002) mention the Copula approach to recover a joint distributions from a limited information. Prieger (2002) and van Ophem (1999, 2000) use Copulas to model the bivariate latent variable distributions. Zimmer and Trivedi (2006) use a trivariate Copula framework to analyze a model with counted outcomes. Xiongtoua and Sriboonchitta (2014) use a Copulas-based GARCH to analyze the volatility and the dependence between the exchange and the inflation rates.

The interest in the Copulas arises from a several perspectives. First, the Copulas are the best useful method for deriving the joint distributions given the marginal ones, especially in the situation of the econometricians who often possess more information about the marginal distributions. Second, the Copulas

allow a convenient choice for modeling a potentially nonlinear dependence.

In this paper, we are interested in the establishment of a relationship between the *US/Euro* exchange and the inflation rates in Euro area for different periods. Afterwards, we present our proposed methodology with a detailed theoretical background. In the third section, we present the used data. The last section concerns the results and a brief conclusion.

2 Methodology and a Theoretical Background

Given two random variables X and Y with the continuous marginals $F(x) = u$ and $G(y) = v$, the Sklar theorem (Sklar, 1959) states that the joint distribution function $H(x, y)$ of (X, Y) can be written in terms of a unique function $C(u, v)$, where:

$$H(x, y) = C(u, v) \quad (1)$$

$C(u, v)$ is known as the Copula of (X, Y) . It describes how $H(x, y)$ is coupled with the marginal functions F and G .

Many families of Copulas $C(u, v)$ have been proposed in the literature (Table 1). We can mention for example the Archimedean Copulas (e.g.: a Gumbel Copula) and the Elliptical Copulas (e.g.: a Normal Copula). To identify the best

Table 1. Some examples of Copulas

Copulas	θ	$C(u, v)$
Gumbel	$[1, \infty)$	$\exp[-((- \log(u))^\theta + (- \log(v))^\theta)^{\frac{1}{\theta}}]$
Clayton	$\theta \in [-1, \infty) \setminus \{0\}$	$\max([u^{-\theta} + v^{-\theta} - 1]^{-\frac{1}{\theta}}, 0)$
Frank	$\theta \in \mathbb{R} \setminus \{0\}$	$-\frac{1}{\theta} \ln(1 + \frac{(\exp(-\theta u) - 1)(\exp(-\theta v) - 1)}{\exp(-\theta) - 1})$
Gaussienne	$[-1, 1]$	$\Phi_\Sigma(\Phi^{-1}(u), \Phi^{-1}(v))$
Student	$[-1, 1]$	$t_{\Sigma, \nu}(t_\nu^{-1}(u), t_\nu^{-1}(v))$
Plackett	$(\theta > 0, \theta \neq 1)$	$\frac{(1 + (\theta - 1)(u + v)) - \sqrt{(1 + (\theta - 1)(u + v))^2 - 4uv\theta(\theta - 1)}}{2(\theta - 1)}$
Galambos	$[0, \infty)$	$uv \exp(((- \log(u))^{-\theta} + (- \log(v))^{-\theta})^{-\frac{1}{\theta}})$

Copula which describes the data, we use first a non-parametric approach, based on graphical tools which allow to identify susceptible families of Copulas.

2.1 The Chi-plot

The Chi-plot allows to identify a nature of a dependence between X and Y . Let $(X_i, Y_i)_{1 \leq i \leq n}$ be a random sample from a bivariate cumulative distribution H .

Let F and G be the marginal distributions of X and Y , respectively. Fisher and Switzer (1985 ,2001) propose a plot of the pairs (λ_i, χ_i) , defined by:

$$\lambda_i = 4 \operatorname{sign}(\widetilde{F}_i \widetilde{G}_i) \max(\widetilde{F}_i^2, \widetilde{G}_i^2) \quad \text{and} \quad \chi_i = \frac{H_i - F_i * G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}, \quad (2)$$

Where:

$$F_i = \frac{1}{(n-1)} \operatorname{rank}(X_i), G_i = \frac{1}{(n-1)} \operatorname{rank}(Y_i), H_i = \frac{1}{(n-1)} \operatorname{rank}(X_i, Y_i),$$

$$\operatorname{rank}(X_i) = \sum_{j \neq i} \mathbf{1}_{X_j \leq X_i}, \operatorname{rank}(X_i, Y_i) = \sum_{j \neq i} \mathbf{1}_{X_j \leq X_i, Y_j \leq Y_i}, \widetilde{F}_i = F_i - 0.5, \text{ and}$$

$$\widetilde{G}_i = G_i - 0.5.$$

To predict the χ_i 's values, Fisher and Switzer (1985 ,2001) built confidence intervals of the form $\pm c_p \sqrt{n}$. They also gave approximate values of the c_p 's for different values of $p \in [0, 1]$. In the case of independence, we expect that $p \times 100\%$ of the pairs (λ_i, χ_i) will be inside the interval $[-c_p \sqrt{n}, c_p \sqrt{n}]$. In the case of a positive dependence, the pairs points go scattered above the band, and conversely for the case of negative dependence.

2.2 The Kendall (K)-Plot

The Kendall plot or the K-plot proposed by Genest and Boies (2003) is a rank based procedure for the detection of dependence. The procedure consists to represent the pair $(W_{i,n}, H_i)$ for $i \in [1, n]$, where $W_{i,n}$ is the expectation of the i^{th} order statistic of a n random sample size. K_0 is a conditional distribution, issued from H , under the independence between X and Y . The form of the bivariate distribution K_0 is given as follows:

$$K_0(w) = P(UV \leq w) = w - w \log(w) \quad (3)$$

Where U and V are independent uniform random variables on the interval $[0, 1]$, and $W_{i,n}$ is given by:

$$W_{i,n} = n \binom{n-1}{i-1} \int_0^1 w \{K_0(w)\}^{(i-1)} (1 - K_0(w))^{(n-i)} dK_0 w \quad (4)$$

The closer the K-plot from the 45-degree line is, less is the association between the random variables.

2.3 Deheuvels Empirical Copula and Mean Squared Error (MSE)

Once we have an idea on the susceptible family of Copulas, we explore another non-parametric approach, to identify a specific sets of Copulas. This approach is based on the comparison regarding the Deheuvels empirical Copula (Deheuvels, 1979 and 1981), using the Mean Squared Error (MSE). The implementation of this approach is done by:

1. The construction of empirical univariate marginal.
2. The construction of the empirical Copula.
3. The construction of the parametric Copula.
4. The minimization of the Mean Squared Error (MSE).

Once the best Copula is identified using the non-parametric approach, we proceed to the identification of the best Copula, using a parametric approach.

2.4 Adjustement of a Parametric Copula

In this paper, the observed data (x_1, \dots, x_n) and (y_1, \dots, y_n) present an auto-correlation. This can be taken into account, by modeling X_t and Y_t as a time series, before adjusting an adequate Copulas (Patton, 2012).

Models such as *ARIMA* and *GARCH* are commonly used for the time series (Box and Jenkins, 1976). If a model is well adjusted to X_t and Y_t , then their residuals ε_{X_t} and ε_{Y_t} are not autocorrelated. In this case, to model the dependence between X_t and Y_t , we adjust a Copula $C(F(\varepsilon_{X_t}), G(\varepsilon_{Y_t}))$, where F and G are the distribution of ε_{X_t} and ε_{Y_t} , respectively.

This Copula can be used for the forecasting, based on the corresponding conditional Copula C_1 , where:

$$P[\varepsilon_{Y_t} \leq \varepsilon_y | \varepsilon_{X_t} = \varepsilon_x] = P[V \leq v | U = u] = \frac{\partial C(u, v)}{\partial u} \equiv C_1(F(\varepsilon_{X_t}), G(\varepsilon_{Y_t})) \quad (5)$$

Let $\tau = C_1(F(\varepsilon_{X_t}), G(\varepsilon_{Y_t}))$, with $\tau \in]0, 1[$. To obtain $Q_{\varepsilon_{X_t}}(\tau | \varepsilon_x)$, which is the τ -th conditional quantile function, given $\varepsilon_{X_t} = \varepsilon_x$, one can solve the equation:

$$Q_{\varepsilon_{X_t}}(\tau | \varepsilon_x) = G^{-1}(C_1^{-1}(\tau, F(\varepsilon_x))) = G^{-1}(C_1^{-1}(\tau, u) | u = F(\varepsilon_x)) = H(\varepsilon_x, \tau) \quad (6)$$

Where G^{-1} and C^{-1} are the inverse function of G and C , respectively.

In this paper, we set τ to 0.5 which corresponds to the case of the median quantile Copula.

Thus, to forecast at the $(t + 1)$ period, the value of the quantile $Q_{\varepsilon_{Y_{t+1}}}$, we can use the adjusted Copula and the value of $\varepsilon_{X_{t+1}}$.

3 Data

The proposed methodology is applied to the monthly inflation and the *US/Euro* exchange rates in the Euro area, from a different periods. The first is from April 2000 to December 2006, devoted to the forecasting of the inflation during 2007. The second adds observations of 2007 to those of the first period, to forecast inflation during 2008. The third period is from September 2009 to March 2016,

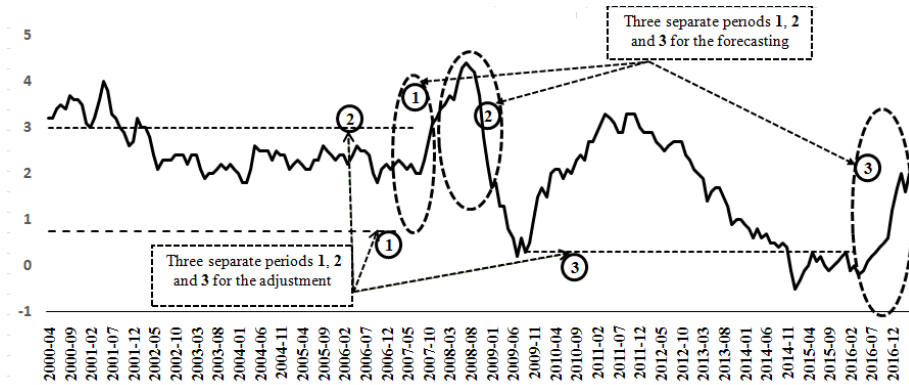


Fig. 1. The evolution of the Inflation in the Euro area from April 2000 to December 2016

devoted to the forecasting of the second quarter of 2016. The reason behind choosing different periods, is to show the sensibility of Copulas according to occurred macroeconomic events.

The representation of the fluctuations in the Inflation in EU and the *US/Euro* exchange rates shows an obvious relationships. Thus, while the exchange rate manifests an uptrend, the inflation behaves erratically.

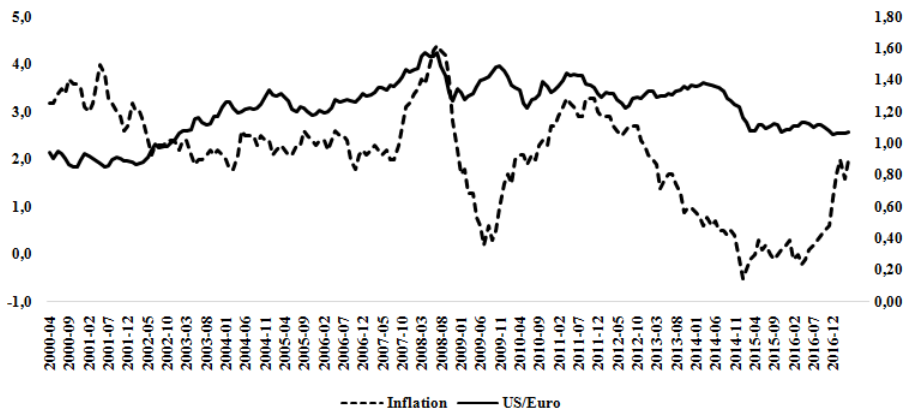


Fig. 2. The exchange rate and the Inflation evolution during sixteen years.

4 Results and Concluding Remarks

To identify an adequate model for the inflation and the *US/Euro* exchange rates, both for the three specified periods, we use the Box-Jenkins modeling strategy. By matching the patterns of the observed sample autocorrelations with the theoretical autocorrelations of a specific times series models, and by using a some goodness of fit criteria and tests (AIC, BIC and a significance of estimated parameters), we identify the following models:

For more details on Exponential smoothing models, see (Holt, 1957; Charles,1957).

Table 2. Residuals deduced from adjusting different time series models

Series	Adjusted models	Residuals
The inflation (First period)	SARIMA(0,1,0)(0,0,1)	$\varepsilon_{Inf,period1}$
The <i>US/Euro</i> exchange rate (First period)	SARIMA(0,1,1)(0,0,0)	$\varepsilon_{Exch,period1}$
The inflation (Second period)	SARIMA(1,1,0)(0,0,1)	$\varepsilon_{Inf,period2}$
The <i>US/Euro</i> exchange rate (Second period)	SARIMA(0,1,1)(0,0,0)	$\varepsilon_{Exch,period2}$
The inflation (Third period)	Exponential smoothing (Holt)	$\varepsilon_{Inf,period3}$
The <i>US/Euro</i> exchange rate (Third period)	SARIMA(1,1,0)(0,0,0)	$\varepsilon_{Exch,period3}$

A non-parametric approach and a goodness of fit test are used to specify the distributions of the residuals, deduced from each adjusted model. The results confirms that the residuals are distributed as Normal variables.

The non-parametric approach is used also to choose a preliminary copulas families, to be tested. The representation of residuals of the inflation and the *US/Euro* exchange rates, doesn't present any particular structure.

However, the examination of the K and Chi-plots (figure 4) gives a specific information. For the first period, this examination indicates an upper tail dependence between the studied variables, since the dots show a clear departure from the diagonal line of 45 degrees, especially from the upper tails of the distribution.

The chi-plot confirms the aforementioned constatation from the K-plot and reveals an upper tail dependence. This result suggests that the underlying Copula belongs to the Archimedean and Elliptical copulas families.

The table 3 above presents the results of the empirical investigation. Based on the values of the Mean Squared Error, some preliminary Copulas are retained. The corresponding values of the MSE are small and close to each others.

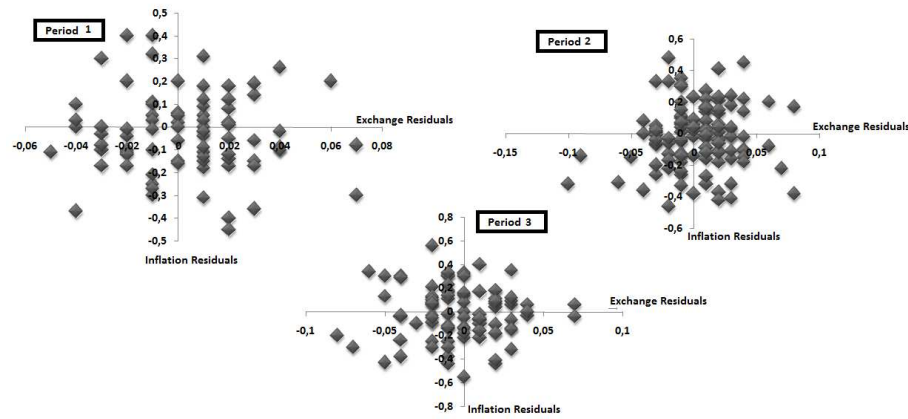


Fig. 3. Dependence patterns of residuals and empirical marginal values of the inflation and exchange rates

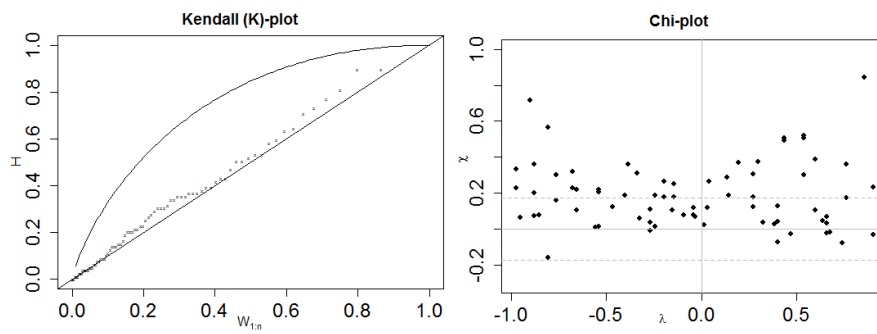


Fig. 4. Chi-plot and K-plot for the first period

Table 3. The retained Copulas from the empirical investigation

Period	Copulas	Parameters	SME
2000-2006	t-Copula	0,02	0,000194476
	Normal	0,02	0,00019655
	Frank	0,1	0,000198502
	Plackett	1,04	0,00019857
	Clayton	0,01	0,000199487
	Gumbel	1	0,000201036
	Galambos	1	0,003321712

The preliminary Copulas are tested using a GoF test. The final results are produced in the following table 4.

Table 4. The final retained Copulas, based on a GoF tests

Periods	Copulas families	Adjusted Copulas	P-values	Parameters
First period	Archimedean	Clayton	0.6848	-0.066658
	Plackett	Plackett	0.5759	0.85718
	Elliptical Copulas	t-Copula	0.9615	-0.054132
		Normal	0.9805	-0.054132

For the first period, the best adequate description of the dependence between the elements of the pairs is given by the Elliptical Copulas (a t-Copula with a p-value of 0.96 and a Normal Copula with a p-value of 0.98), other families of Copulas describe well the studied dependence structure, namely the Archimedean and the Plackett Copulas, but they have a lower p-value, comparatively to the two first one. The Elliptical Copulas include the Gaussian Copula and the Student t-Copula. The first one, constructed from a bivariate normal distribution using the Sklar theorem, allows equal degrees of a positive and a negative dependence. The t-Copula specifies an additional dependence parameter which captures fatness in a distribution's tails.

The Elliptical Copulas, adjusted for the first period, do not have closed form expressions and are restricted to have radial symmetry. Thus, some asymmetries cannot be modeled with elliptical Copulas, such as in many finance and insurance applications, where it seems reasonable that there is a stronger dependence between big losses than between big gains. One practical problem with elliptical distributions in multivariate risk modeling is that all marginal are of the same types. This presents a big problem in the concern of estimation.

Based on the same methodology, the Elliptical Copulas are also adjusted for the second period (a t-Copula with a p-value of 0.93 and a Normal Copula with a p-value of 0.96). For the third period, the best adequate description is given by the Archimedean Copulas (a Clayton Copula with a p-value of 0.95) and the Plackett Copula with a p-value of 0.94. The Elliptical Copulas give also a high p-value, but it cannot reach the two first one. The Archimedean Copulas have the great advantage to capture wide ranges of dependence. Examples of the Archimedean Copulas include the product Copula which corresponds to independence of the examined variables, the Clayton Copula which is used in the case of strong left tail dependence, the Gumbel Copula which is employed in the case of highly correlated variables at high values but less correlated at low values, and the Frank Copula which is applied when a tail dependence is weak. Unlike the Elliptical Copulas, the Archimedean Copulas are not derived from

multivariate distribution functions using Sklar's theorem.

We propose to retain the Elliptical and Archimedean Copulas for the forecasting. This step allows identifying the best model that captures better, the evolution of the dependence between the inflation and the *US/Euro* exchange rate.

Results shows that for the first period, characterized by a stability, Copulas perform well in terms of the forecasting.

Table 5. Predicted values based on the retained Copulas

Observed values	t-Copula	Normal	Clayton	Plackett
-0.19	-0.085	-0.092	-0.096	-0.096
-0.09	-0.033	-0.033	-0.037	-0.037
-0.07	-0.01	-0.008	-0.011	-0.011
-0.05	-0.046	-0.049	-0.052	-0.052
0.02	0.001	0.005	0.002	0.002
0	-0.047	-0.05	-0.053	-0.053

In the opposite, for a crisis and instable periods, the forecasting seems to be biased. This is due to the volatility of the major macroeconomic aggregates. To overcome this problem, we suggest to model the updated series, using a dynamic Copulas.

References

1. International Labour Office (ILO), International Monetary Fund (IMF), Organization for Economic Co-operation and Development (OECD), Statistical Office of the European Communities (Eurostat), United Nations (UN) and the World Bank: Consumer price index manual. Theory and practice. International Labour Office. ISBN, 92-2-113699-X, Geneva (2004).
2. Jongwanich, J., and D. Park: Inflation in developing Asia: Pass-through from global food and oil price shocks. Working Paper, Asian Development Bank (2008).
3. Chollete, L. and Ning, C.: The Dependence structure of macroeconomic variables in the US. Working Papers. Journal of Economics and Finance, 31, University of Stavanger (2009).
4. Munyeka, W.: An In-Depth Look at Economic Growth and Employment in Post-Apartheid South Africa: Analysis and Policy Implications. Journal of Educational and Social Research, 4(3) MCSER, Rome-Italy (2014).
5. Fitzgerald, T.J. and Nicolini, J.P.: Is There a Stable Relationship between Unemployment and Future Inflation? Evidence from U.S. Cities. Working Paper 713, Federal Reserve Bank of Minneapolis Research Department (2014).
6. Kano, T.: Exchange rates and fundamentals: a general equilibrium exploration, Hitotsubashi Institute for Advanced Study, Discussion Paper HIAS-E-19 (2016).

7. Burstein A. and Gopinath, G.: International prices and exchange rates. Handbook of International Economics, vol 4, 391-451, North Holland/Elsevier, London (2014).
8. Engel, C.: Exchange rates and interest parity. Handbook of International Economics, vol 4, 453-522, North Holland/Elsevier, London (2014).
9. European Central Bank (ECB): Annual Report. Statistical section, ISSN 1725-2865 (2016).
10. Sklar, A.: Fonctions de rpartition n dimensions et leurs marges. Publications de l'Institut Statistique de l'universit de Paris 8, 229-231. Paris (1959).
11. Granger, C., Terasvirta, T., Patton, A.: Common factors in conditional distributions for bivariate time series. Journal of Econometrics 132, 43-57 (2006).
12. Miller, D. J. and Liu, W.: On the recovery of joint distributions from limited information. Journal of Econometrics 107, 259-274 (2002).
13. Prieger, J.: A flexible parametric selection model for non-Normal data with application to health care usage. Journal of Applied Econometrics 17(4), 367- 392 (2002).
14. Van Ophem, H.: a general method to estimate correlated discrete random variables. Econometric Theory 15, 228-237 (1999).
15. Van Ophem, H.: modeling selectivity in count data models. Journal of Business and Economic Statistics 18, 503-511 (2000).
16. Zimmer, D. M. and P. K. Trivedi: Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. Journal of Business and Economic Statistics 24, 63-76 (2006).
17. Xiongtoua, T. and Sriboonchitta, S.: Analysis of volatility and dependence between exchange rate and inflation rate in Lao people's Democratic Republic using Copula-Based GARCH Approach. Modeling Dependence in Econometrics 251, 201-214 (2014).
18. Fisher, N.I. and Switzer, P. :Chi-plots of assessing of dependence. Biometrika 72, 253-265, (1985).
19. Fisher, N.I. and Switzer, P. :Graphical assessment of dependence: is a picture worth 100 tests. The American Statistician 55 (3), 233-239, (2001).
20. Genest, C. and Boies, J.C.:Detecting dependence with kendall plots. The American Statistician 57 (4), 275-284 (2003).
21. Deheuvels, P.:La fonction de dpndance empirique et ses proprits. Un test non paramtrique d'indpendance. Acad. Roy. Belg. Bull. Cl. Sci.(5), 65(6), 274- 292 (1979).
22. Holt, C.C.: Forecasting Trends and Seasonal by Exponentially Weighted Averages. Office of Naval Research Memorandum. 52 (1957).
23. Deheuvels, P.: Multivariate tests of independence. In Analytical methods in probability theory. Lecture Notes in Math. Springer, Berlin , 861, 42-50 (1981).
24. Patton, A. J.: A review of copula models for economic time series.J. Multivariate Anal. 110,418 (2012).
25. Box, G.E.P. and Jenkins, G.M.: Time Series Analysis Forecasting and Control. San Francisco: Holden-Day (1976).

Fractal analysis applied to light curves of pulsating stars

Sebastiano de Franciscis^{1*}, Javier Pascual Granado¹, Rafael Garrido Haba¹, and Juan Carlos Suárez^{1 2}

July 21, 2017

¹ CSIC-IAA, Instituto de Astrofísica de Andalucía, Stellar Variability Group
Glorieta de Astronomía s/n, 18008, Granada (Spain)
Email: sebas@iaa.es

² UGR, Universidad de Granada, Facultas de Ciencias, Departamento de Física Teórica y del Cosmos
Avenida de la Fuente Nueva S/N C.P. 18071, Granada (Spain) [3mm]

Abstract

Fractal behaviours, i.e. scale invariance in spatio-temporal dynamic, have been found to describe and model many systems in nature, in particular fluid mechanics and geophysical related geometrical objects, as the convective boundary layer of cumulus cloud fields, topographic landscapes, solar granulation patterns, and observational astrophysical time series, like light curves of pulsating stars.

The main interest in the study of fractal properties in such physical phenomena lies in the close relationships they have with chaotic and turbulent dynamic.

In this poster we introduce some statistical tools for fractal analysis of light curves: Rescaled Range Analysis, Multifractal Spectra Analysis, and Coarse Graining Spectral Analysis (CGSA), an FFT based algorithm, which can discriminate in a time series the stochastic fractal power spectra from the harmonic one.

An interesting application of fractal analysis in asteroseismology concerns the joint use of all these tools in order to develop classification criteria and algorithms for δ -Scuti, γ -Doradus and Solar-like pulsating stars. In fact from the fractal and multi-fractal fingerprints in light curves we could infer the mechanism of modes excitation and/or on the magnetic activity in the outer convective region.

1 Introduction

- **Fractals** are mathematical sets that are defined through [7] **Self-Similarity**, i.e. geometrical objects invariant under **homogeneous scaling**, emerging from infinite iteration rules, and **no integer dimension**, measured by the so called box counting method, generalizing the Euclidean dimension. Fractal dimension is the exponent of the **power law dependence** between the minimal number N_r of boxes that embeds the object and their linear dimension r .
- Fractals patterns in nature emerge by means of an interaction between stochasticity and the accomplishment of a few simple deterministic dynamic rules: in this last case we are dealing with **stochastic fractals**.
- Fractal behaviours have been found in several fluid mechanics dynamical systems: the convective boundary layer of cumulus cloud fields, topographic landscapes, rivers branching, geological formation, thin film growth by molecules deposition. **The main interest in the study of fractal properties in such physical phenomena lies in the close relationships they have with chaotic and turbulent dynamic.**

*Corresponding Author

- **In stellar physics fractal fingerprints** in statistical observables (power law distributed) have been found in perimeter/area correlations [9] and size and lifetime distributions of solar granules [5], as well as in sunspot number variability [3] and in light curves from pulsating stars [1, 2, 8].
- Fractal **time series** $y(t)$ are called **self-affine**, and their scaling relation is [6]:

$$y(\lambda t) = \lambda^\alpha y(t) \quad (1)$$

where α is the so-called **Hurst exponent, characterizing long term correlations and the type of self-affinity in time series**. It is possible to generate series with different values of α by time integration of Gaussian white noise, obtaining Fractional Brownian Motions [6].

The oscillation modes of δ -Scuti stars are excited by the κ -mechanism, driven by the radiative flux excitation and the opacity modulation in He partial ionization regions, which coincide with convective zone, dominated by turbulent and convective dynamic, and thus eventually affected at least by chaos.

In Solar-like pulsating stars, the mechanism is explicitly called stochastic excitation mechanism, and is modeled by a stochastically driven oscillator, whose driving force is a noise that models the dynamic of their typical convective thick outer layer pushing the inner radiative zone.

Our hypothesis is that an appropriate fractal analysis of light curves from pulsating stars could give strong indications on the role played by the stochastic and/or chaotic dynamic of excitation mechanism on their background spectra, and finally characterize it to solar-like or δ -Scuti stars type.

2 Main Objectives

Our work hypothesis establishes that an appropriate fractal analysis of light curves from pulsating stars could give strong indications on the role played by stochastic and/or chaotic dynamic of mode excitation mechanism as well as by magnetic phenomena.

1. Find in fractal characterization of light curves a **robust physical observable associated with chaos and turbulence phenomena typical of the convective envelope**.
2. Correlate the different **fractal fingerprints** in light curves emerging from **mode excitation mechanisms** and, in the solar-like cases, from **magnetic activity**.

3 Coarse Graining Spectral Analysis

The Coarse Graining Spectral Analysis splits in a time series the self-affine component and the harmonic one, giving as output the percentage of (stochastic) fractal power in time series [10]. While the majority of Fourier based analysis consider only the amplitudes of the harmonic components, disregarding the half of the information resulting from Fourier transform, i.e. the phases associated at each harmonic, CGSA method focused also on phases distribution.

CGSA is based on the consideration that in **a self-similar time series FFT phases Θ_k follows a uniform distribution $\Theta_k \in [0, 2\pi]$** .

We consider the original time series $y(i)$ and the series obtained by scaling $y(i)$ by a factor 2 and 1/2:

$$y_2 = \{y(2), y(4), y(6), \dots\} \quad y_{\frac{1}{2}} = \{y(1), y(1), y(2), y(2), \dots\},$$

and we cut them into N_s partially overlapping windows. For each window l we compute the power spectra $S_{yy,l}(k)$ and the cross spectra between the original series and the rescaled ones, i.e. $S_{yy_2,l}(k)$ and $S_{yy_{1/2},l}(k)$. If $y(i)$ is constituted by a sum of few harmonics with fixed phase relationship it is possible to exploit the phase difference between windows $l - 2$ and $l - 1$ to orthogonalize $S_{yy,l}$ with the rotating factor

$$S_{yy,l}^o(k) = S_{yy,l}(k) e^{\pi/2 - (\Theta_{l-1,k} - \Theta_{l-2,k})}.$$

The residuals of such orthogonalization process are non zero in self-affine series, because any rescaled harmonic will find its counterpart in the original series and the phase relationships are always randomly distributed. Taking advantage of Schwartz's inequality we can calculate the fractal module cross correlations

$$\langle || S_{yy,l-1}^{frac}(k) || \rangle_l \equiv \frac{\langle || S_{yy,l-1}(k) \cdot S_{yy,l}^o(k) || \rangle_l}{\langle S_{yy,l-1}(k) \rangle_l} \leq \langle || S_{yy,l}^o(k) || \rangle_l.$$

Finally considering the possible distortions that could emerge by the finite size of the original series and the coarse graining of y_2 and $y_{1/2}$, we define the fractal power and the percentage of fractal power as

$$|| S^{frac}(k) || \equiv \sqrt{|| S_{yy_2}^{frac}(k) || \cdot || S_{yy_{1/2}}^{frac}(k) ||} \quad \% \text{Frac} \equiv \frac{\sum_k || S^{frac}(k) ||}{\sum_k || S(k) ||}.$$

4 Rescaled Range Analysis

An often used approach to the quantification of correlations in self-affine time series, determined by α exponent in 1, is rescaled-range (RR) analysis [4, 6]. Let us consider the running sum of the N long time series relative to its mean value, i.e. $ys(n) = \sum_{i=1}^n (y(i) - \langle y \rangle_N)$. The Hurst exponent, α , is obtained from

$$\lim_{N \rightarrow +\infty} \left(\frac{R_N}{S_N} \right) = \left(\frac{N}{2} \right)^\alpha, \quad (2)$$

here the range is defined by $R_N = \text{Max}(ys(n)) - \text{min}(ys(n))$, and $S_N = \sigma_N$ is the total series standard deviation.

5 Multifractal Singularity Spectrum

Some time series do not exhibit a simple monofractal scaling behavior, which can be accounted for by a single α scaling exponent, and such different scaling behavior can be observed for many interwoven fractal subsets of the time series. Thus a multitude of scaling exponents, associated with different behaviors of small and large fluctuations, is required for a full description of the scaling behavior and a multifractal analysis must be applied [4].

Two general types of multifractality in time series can be distinguished: (i) Multifractality due to a broad probability distribution (density function) for the values of the time series, e. g. a Levy distribution. In this case the multifractality cannot be removed by shuffling the series. (ii) Multifractality due to different long-term correlations of the small and large fluctuations. In this case the probability density function of the values can be a regular distribution with finite moments, e. g., a Gaussian distribution.

The corresponding shuffled series will exhibit non-multifractal scaling, since all long-range correlations are destroyed by the shuffling procedure. Randomly shuffling the order of the values in the time series is the easiest way of generating surrogate data.

A multifractal analysis of time series will also reveal higher order correlations. Multifractal scaling can be observed if, e. g., three or four-point correlations scale differently from the standard two-point correlations studied by classical autocorrelation analysis. In addition, multifractal scaling is observed if the scaling behavior of small and large fluctuations is different. For example, extreme events might be more or less correlated than typical events.

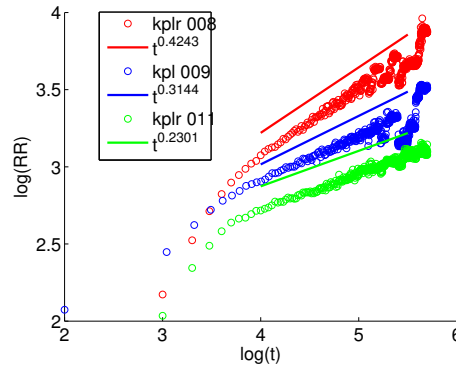
Multifractal singularity spectrum (MSS) gives us the whole broad range of $\bar{\alpha}$ scaling exponents with their relative weights. In this framework we work with the cumulative sum of an $y(n)$ series, i.e. $Y(n) = \sum_{i=1}^n y(i)$. Once we cut our N long $Y(n)$ time series in N_s non-overlapping segments with size s , such that $s = \text{int}(N/N_s)$, we compute the q - th momentum of the fluctuations,

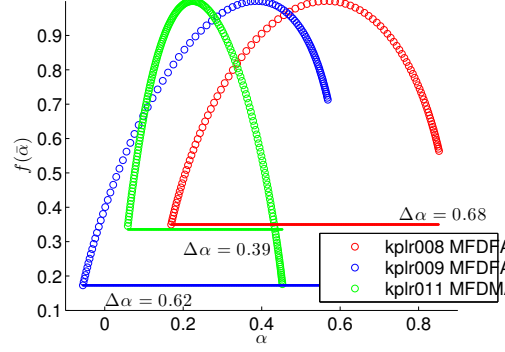
$$F_q(s) = \left\{ \frac{1}{N_s} \sum_{\nu=1}^{N_s} \frac{1}{s} \sum_{i=1}^s \epsilon_{\nu,s}^q(i) \right\}^{\frac{1}{q}} = s^{\frac{1+\tau(q)}{q}} = s^{h(q)}, \quad (3)$$

which define a the generalized multifractal Hurst exponent $h(q)$. Residuals series $\epsilon_{\nu,s}(i)$ could be computed by the Multifractal Detrending Moving Average Analysis (MFDMA), which consists in considering for each segment ν of size s the detrended series using the moving average function $\tilde{Y}(i) = \frac{1}{s} \sum_{k=0}^{s-1} Y(i-k)$ resulting in $\epsilon_{\nu,s}(i) = Y(i) - \tilde{Y}(i)$. Another choice is the Multifractal Detrended Fluctuation Analysis (MFDFA), which estimates an m grade polynomial trend $P^m(\nu)$ by least-square fitting and subtracting this trend from the original profile $\epsilon_{\nu,s}(i) = Y(i) - P^m(i)$. Multifractal singularity spectrum $f(\bar{\alpha})$ is related to $\tau(q)$ via a Legendre transform,

$$\bar{\alpha} = \frac{d}{dq} \tau(q), \quad f(\bar{\alpha}) = q\bar{\alpha} - \tau(q).$$

Here $\bar{\alpha}$ is the singularity strength or Hölder exponent, while $f(\bar{\alpha})$ denotes the dimension of the subset of the series that is characterized by $\bar{\alpha}$. Such multifractal approach can be considered as a generalized version of the fluctuation analysis method, that make use of the second order fluctuation to find the standard (mono) fractal self-affine exponent in eq.1, i.e. $\alpha = \frac{1+\tau(2)}{2}$.





6 Results and Conclusions

ID	CGSA %Frac	RR α	MSS $\Delta\bar{\alpha}$
kplr008 (Solar-like)	1.05	0.42	0.68
kplr009 "	0.85	0.31	0.62
kplr011 "	0.48	0.23	0.39
Active Sun (reg. I and II)	0.68	0.99/0.13	0.51/0.21
Quiet Sun "	0.55	1.00/0.29	0.54/0.4
HD50844 (δ -Scuti)	0.054	0.25	0.22
HD174936 "	0.12	0.38	0.08

We perform fractal analysis of 3 light curves of Kepler solar-like stars, 2 CoRoT δ -Scuti and SoHO/GOLF data from the sun.

- **CGSA %Frac values of all the Kepler stars are compatible with the ones obtained for SoHO/GOLF data, giving us a good indication on the solar-like nature of them.** High %Frac values are a fingerprint of granulation due to the outer convective layer that excites stochastically the oscillating modes, resulting in a background spectra with fractal features.
- **RR Analysis confirms the fractal nature of Kepler, SoHO/GOLF and CoRoT δ -Scuti light curves, since for all of them there is a power law dependence.** While Kepler curve have only 1 fractal regime, SoHO/GOLF displays 2 regimes in the range $t \in (2, 31)$ hours and for $t > 7$ days, and for CoRoT δ -Scuti light curves fractal regime, depending from the star, emerges between $t \in (1, 8)$ hours and breaks down at $t \in (1, 7)$ days (data not shown).
- **MSS width, typical fingerprint of turbulent dynamic, is broader in solar-like than in δ -Scuti stars, indicating the predominant role played of convective layer in the excitation of solar-like modes.**

References

- [1] D. B. de Freitas and et al. New suns in the cosmos? *ApJL*, 773:L18, 2013.
- [2] D. B. de Freitas and et al. New suns in the cosmos. iii. multifractal signature analysis. *ApJ*, 831:87, 2016.

- [3] S. Drozd and P. Oswiecimka. Detecting and interpreting distortions in hierarchical organization of complex time series. *Physical Review E*, 91(3):030902, 2015.
- [4] J. W. Kantelhardt. *Fractal and Multifractal Time Series*, 2008.
- [5] B. Lemmerer, A. Hanslmeier, H. Muthsam, and I. Pianschitsch. Dynamics of small-scale convective motions. *A&A*, 598:A126, 2017.
- [6] B. D. Malamud and D. L. Turcotte. Self-affine time series: measures of weak and strong persistence. *Journal of Statistical Planning and Inference*, 80(12):173 – 196, 1999.
- [7] B. Mandelbrot. *The fractal geometry of nature*. Freeman, 1977.
- [8] J. Pascual-Granado. Fractal analysis of noise in pulsating stars. In M. R. Zapatero Osorio and et al., editors, *Highlights of Spanish Astrophysics VI*, pages 744–748, 2011.
- [9] T. Roudier and R. Muller. Structure of the solar granulation. *solphys*, 107:11–26, 1986.
- [10] Y. Yamamoto and R. L. Hughson. Extracting fractal components from time series. *Physica D Nonlinear Phenomena*, 68:250–264, 1993.

Method for modeling and analysis of natural time series

O.V. Mandrikova, N.V. Fetisova and Yu.A. Polozov

Institute of Cosmophysical Research and Radio Wave Propagation,
Mirnaya Str., 7, Kamchatka Region, Elizovskiy District, Paratunka 684034, Russia;
e-mails: oksanam1@mail.ru, nv.glushkova@ya.ru, up_agent@mail.ru

Abstract. The paper considers a multicomponent model for natural time series (MCM) which was developed by the authors. Model identification is based on the combination of a wavelet transform with the class of autoregressive integrated moving average model (ARIMA). The MCM is capable of studying the characteristic changes of natural time series and of detecting anomalies determined by its structure change. To make a detailed analysis of anomalous changes in the data, computing solutions were developed. They are based on the continuous wavelet transform. They allow the authors to detect different-scale anomalies and to estimate the moments of their occurrences, duration and intensity. On the example of ionospheric critical frequency f_oF2 data, the efficiency of the suggested method is illustrated (*Paratunka site data (Kamchatka, Russia, 53.0 N, 158.7 E, IKIR FEB RAS) and Gakona site data (USA, 62.40 N, 145.0 W.) were under analysis*). Typical changes of f_oF2 variations were investigated in the conditions of calm ionosphere and during disturbed periods (increased solar activity and magnetic storms). During the increased solar activity, ionospheric anomalies (anomalous increases and decreases of electron density in the ionosphere) were detected. They had large spatial-time scales. The detected anomalous increases in electron densities occurring before magnetic storms (pre-storm increases) were of the greatest interest. Analysis showed that the pre-storm increases are typical for the strongest magnetic storms with sudden commencements.

Keywords: wavelet transform, autoregressive-integrated moving average model, ionosphere critical frequency, ionospheric disturbances

1 Introduction

The work is focused on the development of methods and algorithms for analysis of natural time series of complex structure and the construction of automatic systems for their realization. The present paper is concerned with the problem associated with the analysis of ionospheric parameters and detection of anomalous effects occurring during ionospheric disturbances. Ionospheric disturbances cause serious failures in the

operation of ground and space technical equipment [1, 2] that determines applied significance of the investigation.

The Earth's ionosphere is a part of the atmosphere, stretching from 80 to 1000 km and affecting radio wave propagation [1, 3, 4]. Its structure is changeable and heterogeneous, and its investigation is based on the variation analysis of environmental parameters. The ionospheric parameters clearly change with the height, depend on the solar activity cycle, geomagnetic conditions, and geographic coordinates, and have characteristic diurnal and seasonal variations [1-5]. Among the main parameters of the ionosphere are the variations of ionospheric F2-layer critical frequency (foF2). Data of foF2 are registered by vertical radio sounding using an impulse radar (ionosonde). These data are represented as time series. The registered foF2 time series describe electronic concentration of the ionosphere, abrupt fluctuations of which (increase or decrease) leads to their anomalous behavior [1, 2, 4, 5]. During the anomalies in foF2 time series, local features, having different form and duration, occur [6-10]. In most cases ionospheric anomalies are observed during solar flares events and magnetic storms [1-5].

The problems associated with the analysis of ionospheric conditions and detection of anomalies have been addressed by many authors [1, 2, 4, 5, 9, 11-17, 19]. The main approaches include the traditional moving median method [15], ionosphere empirical models [1, 11-13], the application of adaptive algorithms based on neural networks [1, 4, 14, 16], and the wavelet transform [3, 6-10, 17]. At present, the International Reference Ionosphere (IRI) model is the best ionospheric empirical model, based on a wide range of ground and space data [10-12]. Its accuracy significantly depends on the presence of recorded data as well as on the level of solar activity and decrease rapidly with the growth of the latter [1, 11-13]. New developments of ionospheric data empirical models based on the methods of pattern recognition and neuron networks [1, 4, 14, 16] are the most effective in comparison with the IRI model. They are easily realized in automatic mode and are quite flexible. However, at the stage of identification, to describe the pattern space, these models require long training samplings, prone to re-training and can show unpredictable results in the case of too noisy data. To apply the models in real-time mode (or close to it) we need operative data on a complex of geophysical parameters that is not always realizable [1, 4].

Identification of the MCM suggested in the paper is based on the application of ARIMA methods [18] which allow us to obtain quite accurate estimates of model parameters when the samplings are limited, besides the methods are easily realized in automatic mode. However, their main advantage is the possibility to obtain the results with a given confidence level. As long as ARIMA methods are linear, in the case of data complicated structure, their direct application is not effective. Extending the application of ARIMA methods, we developed a new MCM [6-8], based on the combination of multiscale wavelet decomposition with ARIMA models. As the recent investigations [3, 6-10, 19] show, nonlinear adaptive wavelet decompositions are natural and one of the most effective ways for representation of complicated structure data. Adaptive wavelet decomposition is being intensively developed at present [3, 6-10, 19]. Given the large variety of orthogonal basis wavelets and the presence of numerically stable fast algorithms for data transformation, wavelet decomposition pro-

vides many possibilities for the analysis of data with a complex structure [20], including geophysical data [3, 6-10, 19]. Based on the application of ionospheric critical frequency foF2 time series, the papers [6-8] describe a technique for MCM identification. The comparison of MCM with IRI carried out in the article [8] showed that MCM allows us to describe the data time series more accurately, especially during solar activity maxima that proves the efficiency of the suggested approach. This paper is a sequel of those articles. It proves the MCM efficiency for the analysis of typical changes of natural time series and the detection of anomalies. Computing solutions are suggested for a more detailed analysis of foF2 time series structure. They are based on a continuous wavelet transform and threshold functions and allow us to obtain quantitative estimates of times of occurrence, duration and intensity of ionospheric anomalies with high accuracy.

2 Methods

2.1 MCM identification

Considering a random time series $f(t)$ containing stationary components and noise, based on the multiscale wavelet decomposition (MRA) up to the m -th level, the $f(t)$ time series was presented as a linear combination of multiscale components [6-8, 20]:

$$f_0(t) = \sum_{j=-l}^{-m} g[2^j t] + f[2^{-m} t] + e(t), \quad (1)$$

where $f[2^{-m} t] = \sum_k c_{-m,k} \phi_{-m,k}(t)$ is a smoothed component; $c_{-m,k} = \langle f, \phi_{-m,k} \rangle$ are decomposition coefficients describing time series trend; $\phi_{-m,k}(t) = 2^{-m/2} \phi(2^{-m} t - k)$ is a scaling function; $g[2^j t] = \sum_k d_{j,k} \Psi_{j,k}(t)$ are detailing components; $d_{j,k} = \langle f, \Psi_{j,k} \rangle$ are decomposition coefficients describing local changes in a time series; $\Psi_{j,k}(t) = 2^{j/2} \Psi(2^j t - k)$ is a wavelet basis; $e(t)$ is noise; the lower index 0 corresponds to the data initial resolution $j=0$.

By changing the decomposition level m , we could obtain various representations of a time series. Our task was to determine the representation scheme that allowed the extraction of the stationary components from the noise:

1. We consequently make MRA of the time series to the decomposition levels $m = \overline{1, M}$ (the maximum acceptable decomposition level M is determined by the

length N of the time series: $M \leq \log_2 N$ [20]). We obtain smoothed and detailing components $f[2^{-m}t] = \sum_k c_{-m,k} \phi_{-m,k}(t)$ and $g[2^j t] = \sum_k d_{j,k} \Psi_{j,k}(t)$, $j = \overline{-I, -m}$.

2. We test the components $f[2^{-m}t]$ and $g[2^j t]$ for stationarity by estimating their character [18].

3. When the conditions of strict stationarity are fulfilled for the components $f[2^{-m^*}t]$ and $g[2^{j^*}t]$ corresponding to the decomposition level $m = m^*$, we consider that these components describe data regular changes.

To detect regular components $f[2^{-m^*}t]$ and $g[2^{j^*}t]$, we identify the models from ARIMA class based on the traditional approaches [18]. Joining the obtained models into a general multi-component construction, we obtain data representation in the form:

$$f_{0\text{regular}}(t) = \sum_{\mu=1, \overline{T}} \sum_{k=1, N_j^\mu} s_{j^*,k}^\mu b_{j^*,k}^\mu(t), \quad (2)$$

where $s_{j^*,k}^\mu = \sum_{l=1}^{p_j^\mu} \gamma_{j^*,l}^\mu \omega_{j^*,k-l}^\mu - \sum_{n=1}^{h_j^\mu} \theta_{j^*,n}^\mu a_{j^*,k-n}^\mu$ is an estimated μ th component,

p_j^μ and $\gamma_{j^*,l}^\mu$ are the order and parameters of the μ th component autoregression, h_j^μ and $\theta_{j^*,n}^\mu$ are the order and parameters of a moving average of the μ th component, $\omega_{j^*,k}^\mu = \nabla^{\nu^\mu} \beta_{j^*,k}^\mu$, ν^μ is the difference order of the μ th component, $\beta_{j^*,k}^1 = c_{j^*,k}$, $\beta_{j^*,k}^\mu = d_{j^*,k}$, $\mu = \overline{2, \overline{T}}$, \overline{T} is the number of modeled components; $a_{j^*,k}^\mu$ are the residual errors of the μ th component model; N_j^μ is the length of the μ th component; $b_{j^*,k}^1 = \phi_{j^*,k}$ is a scaling function, and $b_{j^*,k}^\mu = \Psi_{j^*,k}$, $\mu = \overline{2, \overline{T}}$ is a wavelet basis of the μ th component.

During the anomalous behavior of data, their structure changes and, as a consequence, the model errors increase. Following the papers [6-8], detection of anomalies in a time series can be based on the analysis of MCM residual errors

$$\mathcal{E}_\mu = \sum_{q=1}^{Q_\mu} |a_{j^*,k+q}^\mu| > T_\mu, \quad (3)$$

where Q_μ is the time of data prediction based on the μ th component model, T_μ is the threshold value of the μ th component, determining the presence of an anomaly,

and $a_{j^*,k+q}^\mu = s_{j^*,k+q}^{\mu, fact} - s_{j^*,k+q}^{\mu, model}$ are the residual errors of the μ th component model at a point $k + q$.

2.2 Modeling of ionospheric critical frequency data for the Kamchatka region

For the model construction, we used hourly data of the ionospheric critical frequency f0F2 (Paratunka station, 52° 58'N, 158° 15'E, Kamchatka, Russia, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS (IKIR FEB RAS)) from 1968 to 2013. To determine the degree of geomagnetic disturbance, we used the K-index (IKIR FEB RAS). To model the foF2 data for a quiet period, the data for calm near earth space (NES), without strong seismic events that occurred in Kamchatka (without earthquakes of $K_s \geq 12$, within a 300 km radius from the station), were used as estimates. Considering the seasonality of ionospheric processes, the different seasons were modeled separately. The level of solar activity (SA) was also considered. The SA was estimated according to the average monthly radio radiation at a wavelength of f10.7. For $f10.7 < 100$, the activity was considered low, while for $f10.7 > 100$, it was considered high. The model identification was performed using the method described in "MCM identification" section. The multiresolution wavelet decomposition of the foF2 data (Eq. 1) was performed using Daubechies wavelet of third order, which was determined by minimization of the approximation error [8]. Based on the operations 1-3 ("MCM identification" section), representation of foF2 time series was determined as

$$f_0(t) = f[2^{-3}t] + g[2^{-3}t] + e(t), \quad (4)$$

where $f[2^{-3}t] = \sum_k c_{-3,k} \phi_{-3,k}(t)$ is the regular smoothed component containing periods of more than 8 h, $g[2^{-3}t] = \sum_k d_{-3,k} \Psi_{-3,k}(t)$ is the regular detailing component containing periods of 8-16 h and $g[2^j t] = \sum_k d_{j,k} \Psi_{j,k}(t)$ $j = \overline{-1, -2}$ are the detailing components taken as noise ones $e(t)$.

The identification results showed that the model parameters depend on season and SA level [8]. MCM general parameters were obtained for winter season for high and low SA according to Eq. 2:

$$s_{3,k}^1 = -0.62 \cdot \omega_{3,k-1}^1 - 0.63 \cdot \omega_{3,k-2}^1 + 0.36 \cdot \omega_{3,k-3}^1 + a_{3,k}^1(t)$$

for the estimated component $f[2^{-3}t]$ and $s_{3,k}^2 = -0.97 \cdot \omega_{3,k-1}^2 - 0.93 \cdot \omega_{3,k-2}^2 + a_{3,k}^2(t)$ for the estimated component $g[2^{-3}t]$. We obtained the following models for summer season according to Eq. 2. For a high SA, we obtained:

$s_{3,k}^1 = -0.50 \cdot \omega_{3,k-1}^1 - 0.58 \cdot \omega_{3,k-2}^1 + a_{3,k}^1(t)$ for the estimated component $f[2^{-3}t]$ and $s_{3,k}^2 = -0.88 \cdot \omega_{3,k-1}^2 - 0.80 \cdot \omega_{3,k-2}^2 + a_{3,k}^2(t)$ for the estimated component $g[2^{-3}t]$. For a low SA, we obtained: $s_{3,k}^1 = -0.83 \cdot \omega_{3,k-1}^1 - 0.73 \cdot \omega_{3,k-2}^1 + a_{3,k}^1(t)$ for the estimated component $f[2^{-3}t]$ and $s_{3,k}^2 = -0.95 \cdot \omega_{3,k-1}^2 - 0.86 \cdot \omega_{3,k-2}^2 + a_{3,k}^2(t)$ for the estimated component $g[2^{-3}t]$.

The investigation on the estimation of the threshold value of the μ th component (T_μ , see Eq. (3)), determining the anomaly in foF2 time series, also showed its dependence on SA and season. Values of T_μ were determined on the basis of foF2 data prediction error dispersion [8, 18]. For Paratunka (Kamchatka) site it is:

- for winter time: $T_1 = 1.37/1.22$ (high/low SA), $T_2 = 0.97/0.73$ (high/low SA);
- for summer time: $T_1 = 1.60/1.30$ (high/low SA), $T_2 = 0.88/0.80$ (high/low SA).

Statistic modeling was performed to evaluate the efficiency of MCM. The model time series included the following commutative components: foF2 time series median values, white noise and “triangle impulse” anomalies. Duration of the anomalies was from 3 to 17 hours, the anomalies amplitude changed from 1.5 to 6 MHz, noise amplitude was 1.5 MHz. Figure 1 illustrates the evaluation results of the probability of anomalies detection depending on their amplitude and duration. The analysis of Figure 1 shows that on the basis of the model, long-period anomalies (from 7 h and more) can be detected with a probability of 85% if their amplitude exceeds the noise amplitude by a factor of 1.5.

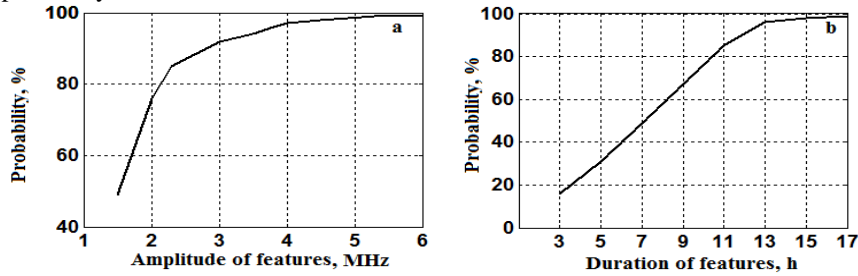


Fig. 1. The graph of the dependence of the probability of anomalies detection on their amplitude and duration (winter, low solar activity, noise amplitude is 1.5 MHz): **a** the duration of the features is 7 hours, and **b** the amplitude of the features is 1.5 MHz

Figure 2 illustrates the modeling results of the foF2 data (Paratunka station) for December 11-22, 2015. The analyzed period is characterized by increased geomagnetic activity. The results show that during disturbed periods (K-index exceeds 3), the obtained model errors increase and go beyond the limits of a typical deviation that indicates anomalous changes in the foF2 data. During a strong magnetic storm which occurred on December 20-22, 2015, the most significant increase of errors is ob-

served, that is more than 5 standard deviations for component $f[2^{-3}t]$ and more than 2.5 standard deviations for $g[2^{-3}t]$ component.

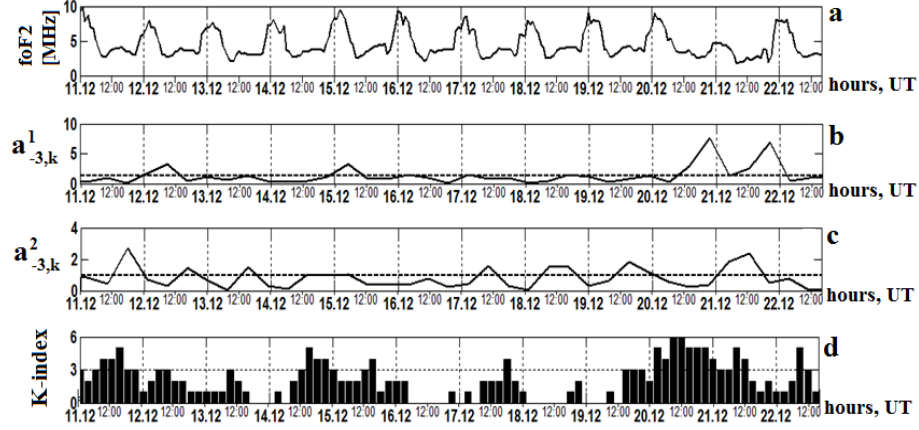


Fig. 2. Analysis results of the foF2 data (Paratunka station, Kamchatka) for December 11-22, 2015: **a** the foF2 data, **b** – errors of the $f_{-3}(t)$ smoothed component model, **c** errors of the $g_{-3}(t)$ detailing component model, and **d** K-index of geomagnetic activity. Graphs **b** and **c** show standard deviations of component model errors (dashed lines)

2.3 Ionospheric anomaly detection and estimation of their parameters based on the continuous wavelet transform and threshold functions

Regarding each basic wavelet Ψ , the continuous wavelet transform was given by the following formula [20]:

$$W_{\Psi}f_{b,a} := |a|^{-1/2} \int_{-\infty}^{\infty} f(t)\Psi\left(\frac{t-b}{a}\right)dt, \quad f \in L^2(R), a, b \in R, a \neq 0.$$

A decrease in the $|W_{\Psi}f_{b,a}|$ coefficient amplitudes depending on scale a is associated with the Lipschitz's uniform and dot smoothness of the Lipschitz's function f [20]. According to the Zhaffar's theorem [20], when scale a decreases, the amplitudes of the $|W_{\Psi}f_{b,a}|$ coefficients rapidly decrease to zero where the function f is smooth and has no local features. Based on this property of the wavelet transform, we used the following threshold function to detect local features in the time series of the foF2 critical frequency and identify ionospheric anomalies:

$$P_{T_a}(W_{\Psi}f_{b,a}) = \begin{cases} W_{\Psi}^+ f_{b,a}, & \text{if } (W_{\Psi}f_{b,a} - W_{\Psi}f_{b,a}^{med}) \geq T_a \\ 0, & \text{if } |W_{\Psi}f_{b,a} - W_{\Psi}f_{b,a}^{med}| < T_a \\ W_{\Psi}^- f_{b,a}, & \text{if } (W_{\Psi}f_{b,a} - W_{\Psi}f_{b,a}^{med}) \leq -T_a \end{cases} \quad (5)$$

where $T_a = U * St_a$ is the threshold detects the presence of an anomaly for an a scale near point ξ included in the carrier $\Psi_{b,a}$ (see below), U is a threshold coefficient and $St_a = \sqrt{\frac{1}{\Phi-1} \sum_{k=1}^{\Phi} (W_{\Psi}f_{b,a} - \overline{W_{\Psi}f_{b,a}})^2}$, $\overline{W_{\Psi}f_{b,a}}$ and $W_{\Psi}f_{b,a}^{med}$ are the average and median for a moving time window of length Φ . Taking into account the diurnal variation of the ionospheric data, the average $\overline{W_{\Psi}f_{b,a}}$ and median $W_{\Psi}f_{b,a}^{med}$ were calculated separately for each hour.

Since the $\Psi_{b,a}$ carrier for an a scale is $[b - \Omega a, b + \Omega a]$, the cone of influence of ξ on a was defined by the following inequality [20]

$$|b - \xi| \leq \Omega a.$$

The anomaly duration for a was then defined by the influence cone of ξ and equal to:

$$H_a = 2\Omega a$$

The anomaly intensity for $t = b$ was defined as:

$$Y_b = \sum_a \frac{|P_{T_a}(W_{\Psi}f_{b,a})|}{\|W_{\Psi}f_{b,a}\|_2}, \quad (6)$$

where $\|W_{\Psi}f_{b,a}\|_2 = \sqrt{\sum_{N_a} (P_{T_a}(W_{\Psi}f_{b,a}))^2}$ is the norm and N_a is the series length for scale a .

Figure 3 shows the application results of Eq. 5 and Eq. 6, during the magnetic storm of August 25–26, 1987. Analyze of the results shows that a negative anomaly (it is shown in Figure 3b in blue), lasting for more than 1 day and characterizing by a decrease in the ionospheric electron density, occurred in the foF2 data during a magnetic storm. Its intensity increased from the beginning of the storm and was maximum during the main phase of the storm. After the magnetic storm, the electron density increased, as indicated by the positive anomalies (it is shown in Figure 3b in red) from 28 August 1987. During the storm, small-scale anomalies, associated with local variations of the ionospheric electron density also occurred.

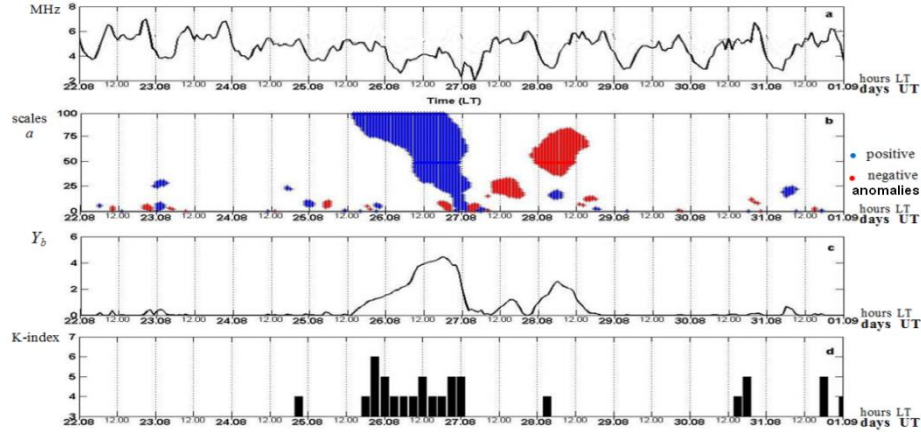


Fig. 3. Results of the foF2 data processing (Paratunka station, Kamchatka) for August 22-31, 1987: **a** the foF2 data, **b** detected anomalies for a threshold coefficient $U = 2.3$ and a moving time window length Φ of 336 h, **c** estimation of the anomaly intensity and **d** K-indices above three (Paratunka station, Kamchatka).

3 Application of the method to analyze ionospheric critical frequency data during magnetic storms

Fig. 4 illustrates the results of application of the developed method for the foF2 data (Paratunka site (Kamchatksiy kray, Russia), 53.0 N, 158.70 E) and Gakona site (USA), 62.40 N, 145.0 W). To analyze the near earth space (NES), Fig. 4 shows geomagnetic disturbance intensity determined on the basis of the value $E_b = \sum_a |W_\psi f_{b,a}|$

[10], Dst-index and solar wind speed. Within the period under analysis from May 30 to June 3, 2013, based on space weather data [http://ipg.geospace.ru/], the gradual increase of solar wind speed which began at about 14:00 UT on May 31 was accompanied by a strong magnetic storm with gradual commencement at 00:00 on June 1, 2013. Analysis of the results of foF2 data processing demonstrates a common character of the behavior for the ionosphere. Modeling results (Fig. 4 d, k) show anomalous processes in foF2 data before and during the event. It is a significant increase of MCM errors at Paratunka site ($f[2^{-3}t]$ is more than 2.5 of standard deviations (SD); $g[2^{-3}t]$ is more than 4.3 of SD) and at Gakona site ($f[2^{-3}t]$ is more than 3.4 of SD). Before the storm, gradual increase of electron concentration was observed in the ionosphere. Positive anomaly maximum was observed on May 31, at 20:00 UT at Paratunka site and at 00:00 at Gakona site (Fig. 4 c, j). Comparison of the occurrence time, duration and intensity dynamics of the anomaly with NES data and taking into account that integral SA was low from May 30 to June 1 [http://ipg.geospace.ru/], allow us to assume the relation of the detected anomaly with the oncoming magnetic storm. During the magnetic storm, electron concentration decrease (10:00-11:00 UT) and

formation of the negative phase of ionospheric storm was observed (Fig.4 b, c, i, j). Similar character of ionosphere behavior is observed during magnetic storms on April 5 and August 3, 2010 (event on April 5 was discussed in the paper [9], the event on August 3 was considered in the paper [10]), on March 7, 2012 [10], on March 17 and October 2, 2013 (events are discussed in the papers [8, 10]), on February 19 and November 4, 2014 [8]. The electron concentration increased before the events. During the magnetic storms and, in some cases, at the recovery stage, significant and long decreases of electron concentration were observed.

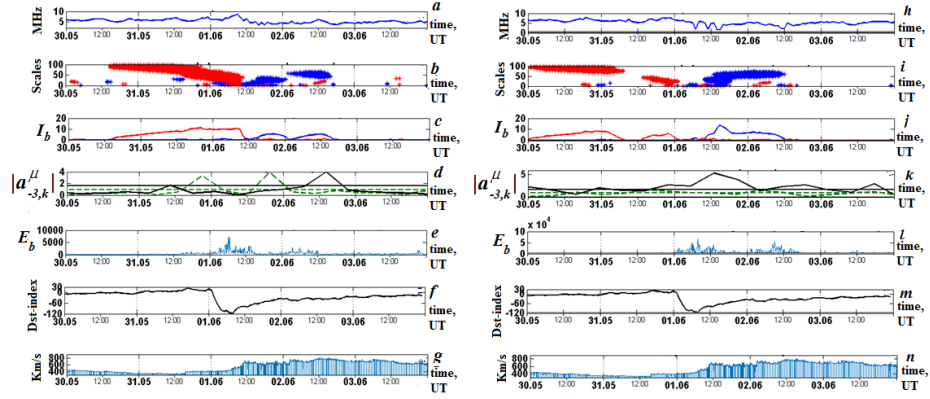


Fig. 4. Results of data processing for Paratunka site (left) and Gakona site (right) for the period of May 30 – June 3, 2013: **a, h** the foF2 data, **b, i** blue color shows negative anomalies, red color shows positive ones, **c, j** estimation of the anomaly intensity, **d, k** MCM errors of smoothed (black) and detailing (green) components and their standard deviations, **e, l** estimation of the geomagnetic disturbance intensities, **f, m** Dst-index, and **g, n** solar wind speed

Table 1 presents the results of analysis of ionospheric data during magnetic storms which occurred within the period of 2004-2014. Analysis of Table 1 shows that the pre-storm increase effect is typical for strong storms with sudden commencement. The detected effects were observed independently on local time at the background of calm and slightly disturbed geomagnetic field. The duration was from several hours to a day and a half. Significant decrease of electron concentration in the ionosphere was observed during the strongest events. Similar behavior of the ionosphere during ground measurements and in TEC data was mentioned in the review [5]. The results obtained in this paper agree well with the effects described in the article [5]. We hold to the opinion of the authors [5] and think that such ionospheric effects are associated with some channel of energy penetration from interplanetary space and the magnetosphere. In this case, the pre-storm ionospheric anomalies may be a signal of the oncoming geomagnetic storm that has high applied significance.

Table 1. Results of ionospheric data processing of Paratunka site

Event date / starting time (UT) / anomaly character: G – with gradual commencement, Sd – with sudden commencement / storm strength: M-moderate, S-strong	SA	Presence (absence) of anomaly before a storm: P - positive anomaly, N – negative anomaly / its duration/ how many hours before a storm it occurred	Presence (absence) of anomaly during a storm: P - positive, N – negative / its duration	Threshold U
5.04.2010 / 8.26 / Sd/ S	low	P/ 17 h / 16 h	N/ 29 h	2
3.08.2010/18.40-18.50/ G / S	low	P / 38 h/ 19 h	N/10 h	2
7.03.2012 / 05.28-05.45 / G/S	high	P/ 37 h/ 12 h	N/ 17 h	2.5
17.03.2013 / 06.05/ Sd/ S	high	P/ 31 h / 27 h	N/22 h	2.5
1.06.2013 / 00.40/ G / S	high	P/ 42 h / 4 h	N/ 25 h	2.5
2.10.2013 / 01.52/ Sd/ M	high	P/ 38 h /27 h	N/ 20 h	2.5
8.12.2013/1.00-2.20/G / S	high	N/ 28 h / 14 h	P/14 h	2.5
14.12.2013/15.00-16.00/G/M	high	---	P /21 h	2.5
19.02.2014 / 03.50/ G / S	high	P / 23 h / 10 h	N /14 h	2.5
11.09.2014/23.30/G/ S	high	---	N/ 31 h	2.5
04.11.2014/10.10/Sd/ M	high	P/ 25 h / 14 h	N/ 22 h	2.5

4 Conclusions

The method, based on the application of wavelet transform and traditional methods, developed by the authors to analyze natural time series of complicated structure showed the efficiency in the tasks of analysis of ionospheric parameters and in detection of anomalies during ionospheric disturbances. Analysis of foF2 data during increased solar activity and magnetic storms has confirmed the fact of possible occurrence of electron concentration increase pre-storm effect in the regions under analysis and showed the possibility of the application of the developed method in detection of similar effects.

Acknowledgements.

The paper was supported by RSF Grant No. 14-11-00194. The authors are grateful to the organizations carrying out the registration of ionospheric and magnetic data which were applied in the paper. The authors also appreciate the organizations supporting the site <http://ipg.geospace.ru/>.

References.

1. Nakamura, M., Maruyama, T., Shidama, Y.: Using a neural network to make operational forecasts of ionospheric variations and storms at Kokubunji, Japan. *Journal of the National Institute of Information and Communications Technology*. 56, 391-406 (2009).
2. Afraimovich, E., Kosogorov, E., Leonovich, L.: The use of the international GPS network as the global detector (GLOBDET) simultaneously observing sudden ionospheric disturbances. *Earth, Planets, and Space*. 52(11), 1077-1082 (2000).

3. Kato, H., Takiguchi, Y., Fukayama, D., Shimizu, Y., Maruyama, T., Ishii, M.: Development of automatic scaling software of ionospheric parameters. *Journal of the National Institute of Information and Communications Technology*. 56, 465-474 (2009).
4. Watthanasangmechai, K., Supnithi, P., Lerkvaranyu, S., Tsugawa, T., Nagatsuma, T., Maruyama, T.: TEC prediction with neural network for equatorial latitude station in Thailand. *Earth, Planets and Space*. 64(6), 473-483 (2012).
5. Danilov, A.: Ionospheric F-region response to geomagnetic disturbances. *Advances in Space Research*. 52(3), 343-366 (2013).
6. Mandrikova, O., Glushkova, N., Zhivet'ev, I.: Modeling and analysis of ionospheric parameters by a combination of wavelet transform and autoregressive models. *Geomagnetism and Aeronomy*. 54(5), 593-600 (2014).
7. Mandrikova, O., Fetisova (Glushkova), N., Al-Kasasbeh, R., Klionskiy, D., Geppener, V., Ilyash, M.: Ionospheric parameter modelling and anomaly discovery by combining the wavelet transform with autoregressive models. *Annals of Geophysics*. 58, (2015).
8. Mandrikova, O., Fetisova, N., Polozov, Y., Solovev, I., Kupriyanov, M.: Method for modeling of the components of ionospheric parameter time variations and detection of anomalies in the ionosphere coupling of the high and mid latitude ionosphere and its relation to geospace dynamics. *Earth, Planets and Space*. 67(1), 131-146 (2015).
9. Baishev, D., Moiseyev, A., Boroyev, R., Kobayakova, S., Stepanov, A., Mandrikova, O., Solovev, I., Khomutov, S., Polozov, Yu., Yoshikawa, A., Yumoto, K.: Magnetic and ionospheric observations in the Far Eastern region of Russia during the magnetic storm of 5 April 2010. *Sun and Geosphere*. 10(2), 133-140 (2015).
10. Mandrikova, O., Polozov, Yu., Solovev, I., Fetisova (Glushkova), N., Zalyaev, T., Kupriyanov, M., Dmitriev, A.: Methods of analysis of geophysical data during increased solar activity. *Pattern recognition and image analysis*. 26(2), 406-418 (2016).
11. Klimenko, M., Klimenko, V., Zakharenkova, I., Karpov, I.: Numerical modeling of the global ionospheric effects of storm sequence on September 9-14, 2005-comparison with IRI model. *Earth, Planets and Space*. 64(6), 433-440 (2012).
12. Bilitza, D., Reinisch, B.: International Reference Ionosphere 2007: Improvements and new parameters. *Advances in space research*. 42, 599-609 (2007).
13. Oyekola, O., Fagundes, P.: Equatorial F_2 -layer variations: Comparison between F_2 peak parameters at Ouagadougou with the IRI-2007 model. *Earth, Planets and Space*. 64(6), 553-566 (2012).
14. Zhao, X., Ning, B., Liu, L., Song, G.: A prediction model of short-term ionospheric foF2 based on AdaBoost. *Advances in Space Research*. 53(3), 387-394 (2014).
15. Mikhailov, A., Morena, B., Miro, G., Marin, D.: A method for foF2 monitoring over Spain using the El Arenosillodigisonde current observations. *Annals of Geophysics*. 42(4), (1999).
16. Wang, R., Zhou, C., Deng, Z., Ni, B., Zhao, Z.: Predicting foF2 in the China region using the neural networks improved by the genetic algorithm. *Journal of Atmospheric and Solar-Terrestrial Physics*. 92, 7-17 (2013).
17. Hamoudi, M., Zaourar, N., Mebarki, R., Briquieu, L., Parrot, M.: Wavelet analysis of ionospheric disturbances. In: *EGU General Assembly 2009, Geophysical Research Abstracts*, pp. 8523. Copernicus, Vienna (2009).
18. Box, G., Jenkins, G.: *Time series analysis: Forecasting and control*. Holden-Day, San Francisco (1970).
19. He, L., Wu, L., Liu, S., Ma, B.: Seismo-ionospheric anomalies detection based on integrated wavelet. In: *Geoscience and Remote Sensing Symposium*, pp. 1834-1837. IEEE, Vancouver (2011).
20. Mallat, S.: *A wavelet tour of signal processing*. Academic Press, London (1999).

A New Estimation Technique for AR(1) Model with Long-tailed Symmetric Innovations

Ayşen Dener Akkaya^{1*} (0000-0002-0886-3295) and

Özlem Türker Bayrak² (0000-0003-0821-150X)

¹Middle East Technical University, Ankara, Turkey
akkay@metu.edu.tr

²Cankaya University, Ankara, Turkey
ozlemt@cankaya.edu.tr

Abstract. In recent years, non-normal innovations in time series models are observed in many applications and the estimation problem is considered newly through different distributions by the use of modified maximum likelihood (MML) estimation technique which assumes the shape parameter to be known. This becomes a drawback in machine data processing where the underlying distribution cannot be determined but assumed to be a member of a broad class of distributions. Therefore, in this study, the shape parameter is also assumed to be unknown and the MML technique is combined with Huber technique to estimate the model parameters of autoregressive (AR) models of order 1; named as adaptive modified maximum likelihood (AMML) estimation. After derivation of the AMML estimators, their efficiency and robustness properties are discussed through simulation study and compared with both MML and the least-squares (LS) estimators.

Keywords: Adaptive modified maximum likelihood. Autoregressive models. Least squares estimators. Modified maximum likelihood. Estimation. Efficiency. Robustness.

1 Introduction

The classical AR models assume that the innovations are normally distributed which might not be the case in applications. Therefore, in recent studies, this assumption is relaxed and MML method developed by Tiku [9] is utilized to estimate unknown parameters in such situations [1 – 4, 13 – 14]. The LS estimators are neither efficient nor robust and maximum likelihood (ML) estimators are elusive due to the implicit nature of likelihood functions under non-normality. Use of iterative approach is weary due to convergence problems and produce bias in estimators especially for small samples. The MML estimators capture all the good statistical properties of ML estimators and they are explicit functions of sample observations. Besides they are, i) considerably more efficient (unbiased and smaller variance) than the LS estimators for all sample sizes, particularly for large n , ii) asymptotically fully efficient under

very general regularity conditions and almost fully efficient for small samples and iii) robust to plausible deviations from the assumed distribution and mild data anomalies (outliers, inliers etc). For the detailed information about MML and its applications, one can refer to [10, 12]. On the other hand, MML method is based on the assumption of a particular distribution; i.e. the shape parameter is known. Many different ways are suggested in literature for determining the shape parameter including the use of Q-Q plots [6]. Due to the intrinsic robustness of MML estimators [10], the values obtained by any of these methods will yield essentially the same estimates and standard errors for plausible alternatives. However, when the data is huge and machine learning methods will be applied, it is important to estimate this parameter also since one has no opportunity to investigate the nature of the underlying distribution in this case. It can only be assumed that it is a member of a broad class of distributions. Inserting a likelihood equation related to the shape parameter into the likelihood equation system makes it unsolvable analytically even MML estimation method is used. Thus, there is a need to extend the MML method so that the assumption on the shape parameter is relaxed.

In studies [7 – 8], M-estimators which are efficient and robust under a broad class of long-tailed symmetric (LTS) distributions are developed. In this study, following [5, 11, 15], we use an adapted form of MML estimators which combine the logic of MML with M-estimators named as AMML estimation in memory of Moti Lal Tiku who actually initiated the idea and thought this name. The parameters are estimated under the assumption that the innovations in AR(1) model belong to LTS family. The efficiency and robustness properties of them are discussed via simulation as well as their comparison with LS and MML estimators.

2 Estimation of the Model Parameters

Consider the time series model

$$y_t = \mu + \phi y_{t-1} + \varepsilon_t, \quad (1 \leq t \leq n), \quad (-1 < \phi < 1) \quad (1)$$

where the innovations ε_t are independent and identically distributed (iid), and have one of the distributions in LTS family

$$f(\varepsilon) = \frac{\Gamma(p)}{\sigma \sqrt{k} \Gamma(1/2) \Gamma(p-1/2)} \left(1 + \frac{\varepsilon^2}{k\sigma^2}\right)^{-p}, \quad -\infty < \varepsilon < \infty; \quad (2)$$

where $k = 2p-3$ and $p \geq 2$. $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. For $p = \infty$, Equation (2) reduces to normal $N(0,1)$. Note that the distribution of $t = \sqrt{v/k} (\varepsilon/\sigma)$ is Student's t with $v = 2p - 1$ degrees of freedom. The likelihood function is

$$L \propto \sigma^{-n} \prod_{i=1}^n \left(1 + \frac{\varepsilon_i^2}{k\sigma^2}\right)^{-p}. \quad (3)$$

In fact, Equation (3) is the likelihood function conditional on $y_0 = \varepsilon_0/\sqrt{1-\phi^2}$ where ε_0 is an independent innovation that has the same distribution as ε_i ($1 \leq i \leq n$). Actually, this is Model 2 of [16] which is more general than their Model 1.

2.1 Modified Maximum Likelihood Estimators

The likelihood equations for known p are obtained in terms of $z_i = (y_i - \phi y_{i-1} - \mu)/\sigma = \varepsilon_i/\sigma$, ($1 \leq i \leq n$) as follows:

$$\begin{aligned}\frac{\partial \ln L}{\partial \mu} &= \frac{2p}{k\sigma} \sum_{i=1}^n g(z_i) = 0, \\ \frac{\partial \ln L}{\partial \phi} &= \frac{2p}{k\sigma} \sum_{i=1}^n g(z_i) y_{i-1} = 0, \\ \frac{\partial \ln L}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^n z_i g(z_i) = 0,\end{aligned}\tag{4}$$

where $g(z_i) = \frac{z_i}{\left(1 + \frac{z_i^2}{k}\right)}$.

Since Equations (4) include nonlinear function, $g(z_i)$, they have no explicit solutions and iterative solutions are problematic. MML method is used to find estimators which are known to be asymptotically equivalent to ML estimators [10]. Estimation procedure is carried out in three steps: (i) the maximum likelihood equations are expressed in terms of the order statistics of $z_{(i)} = (y_{[i]} - \phi y_{[i]-1} - \mu)/\sigma$ where $(y_{[i]}, y_{[i]-1})$ are the concomitants of $z_{(i)}$, i.e., the pair (y_j, y_{j-1}) ($j = [i]$) associated with the i th ordered value, $z_{(i)}$ so that the ordering of the time series data is not lost; (ii) the nonlinear function $g(z_{(i)})$ are replaced by linear approximations $g(z_{(i)}) \cong \alpha_i + \beta_i z_{(i)}$, $1 \leq i \leq n$ where the constant coefficients α_i and β_i are obtained from the first two terms of a Taylor series expansion of $g(z_{(i)})$ around the i th population quantile, $t_{(i)} = E(z_{(i)})$. Here we use approximate values of $t_{(i)}$ calculated from

$$\frac{\Gamma(p)}{\sqrt{k}\Gamma(1/2)\Gamma(p-1/2)} \int_{-\infty}^{t_{(i)}} \left(1 + \frac{z^2}{k}\right)^{-p} dz = \frac{i}{n+1} \quad (1 \leq i \leq n).\tag{5}$$

The resulting α_i and β_i are

$$\alpha_i = (2/k)t_{(i)}^3 / \{1 + (1/k)t_{(i)}^2\}^2 \text{ and } \beta_i = [1 - (1/k)t_{(i)}^2] / \{1 + (1/k)t_{(i)}^2\}^2.\tag{6}$$

(iii) incorporating Equations (6) in Equations (4) and by solving the modified (linearized) likelihood equations $\partial \ln L^* / \partial \mu = 0$, $\partial \ln L^* / \partial \phi = 0$ and $\partial \ln L^* / \partial \sigma = 0$, the MML estimators are obtained as:

$$\begin{aligned}\hat{\mu} &= \sum_{i=1}^n \beta_i (y_{[i]} - \hat{\phi} y_{[i-1]}) / m, \\ \hat{\phi} &= K + D \hat{\sigma}, \quad \hat{\sigma} = (B + \sqrt{B^2 + 4nC}) / 2n\end{aligned}\quad (7)$$

where

$$\begin{aligned}m &= \sum_{i=1}^n \beta_i, \quad K = \frac{\sum_{i=1}^n \beta_i y_{[i]} y_{[i-1]} - \frac{1}{m} \sum_{i=1}^n \beta_i y_{[i]} \sum_{i=1}^n \beta_i y_{[i-1]}}{\sum_{i=1}^n \beta_i y_{[i-1]}^2 - \frac{1}{m} (\sum_{i=1}^n \beta_i y_{[i-1]})^2}, \\ D &= \frac{\sum_{i=1}^n \alpha_i y_{[i-1]}}{\sum_{i=1}^n \beta_i y_{[i-1]}^2 - \frac{1}{m} (\sum_{i=1}^n \beta_i y_{[i-1]})^2}, \\ B &= \frac{2p}{k} \sum_{i=1}^n \alpha_i (y_{[i]} - \bar{y}_{[.]} - K(y_{[i-1]} - \bar{y}_{[.]-1})), \\ C &= \frac{2p}{k} \sum_{i=1}^n \beta_i (y_{[i]} - \bar{y}_{[.]} - K(y_{[i-1]} - \bar{y}_{[.]-1}))^2, \\ \bar{y}_{[.]} &= \sum_{i=1}^n \beta_i y_{[i]} / m, \quad \bar{y}_{[.]-1} = \sum_{i=1}^n \beta_i y_{[i-1]} / m.\end{aligned}$$

Comment: The coefficients β_i ($1 \leq i \leq n$) increase until the middle value and then decrease in a symmetric fashion. Therefore, if β_1 is positive then all the β_i coefficients are positive and $\hat{\sigma}$ is real and positive. For small p and large n , however, β_1 (and a few other β_i coefficients) can be negative and needed to be rectified. Thus, if β_1 turns out to be negative, we replace α_i by $\alpha_i^* = (1/k)t_{(i)}^3 / \{1 + (1/k)t_{(i)}^2\}^2$ and β_i by $\beta_i^* = 1 / \{1 + (1/k)t_{(i)}^2\}^2$ ($1 \leq i \leq n$).

Computations: The estimates of the parameters require figuring out the concomitants found out by sorting the innovations. Therefore, there is a need to obtain the initial estimates for the parameters in the model. This is done by the use of LS estimators $\tilde{\mu}$ and $\tilde{\phi}$, given in Equations (8) since LS estimators do not need any distributional assumptions. Then the initial estimates of innovations $\tilde{\varepsilon}_i = y_i - \tilde{\phi} y_{i-1} - \tilde{\mu}$ ($1 \leq i \leq n$) are ordered to obtain the concomitants $(y_{[i]}, y_{[i-1]})$ corresponding to the i th ordered value of the estimated residual. By the use of these concomitants, the MML estimators are calculated from Equations (7). To eliminate the effects of the initial estimates, the LS estimators $\tilde{\mu}$ and $\tilde{\phi}$ are then replaced by $\hat{\mu}$ and $\hat{\phi}$, respectively and the corresponding innovations $\hat{\varepsilon}_i = y_i - \hat{\phi} y_{i-1} - \hat{\mu}$ are ordered to obtain the new concomitants. The revised MML estimators are computed from these new concomitants. The process is repeated one more time for the estimates to stabilize sufficiently.

2.2 Least Squares Estimators

Regardless of the underlying distribution, the LS estimators are

$$\begin{aligned}\tilde{\mu} &= \frac{\sum_{i=1}^n y_i}{n} - \tilde{\phi} \frac{\sum_{i=1}^n y_{i-1}}{n}, \quad \tilde{\phi} = \frac{\sum_{i=1}^n y_i y_{i-1} - \sum_{i=1}^n y_i \sum_{i=1}^n y_{i-1} / n}{\sum_{i=1}^n y_{i-1}^2 - (\sum_{i=1}^n y_{i-1})^2 / n}, \\ \text{and} \quad \tilde{\sigma} &= \frac{\sqrt{\sum_{i=1}^n (y_i - \tilde{\phi} y_{i-1} - \tilde{\mu})^2}}{n-2}.\end{aligned}\quad (8)$$

2.3 Adaptive Modified Maximum Likelihood Estimators

Since the shape parameter p is unknown, the coefficients α_i and β_i have to be estimated from the sample data. The idea of Huber [8] is implemented for this purpose. Let T_0 and S_0 be the initial estimators of μ and σ , respectively given as:

$$T_0 = \text{med}\{y_i - \hat{\phi}_0 y_{i-1}\} \text{ and } S_0 = 1.483 \text{ med}\{|y_i - \hat{\phi}_0 y_{i-1} - T_0|\}, (1 \leq i \leq n) \quad (9)$$

where $\hat{\phi}_0 = \text{med}\left\{\frac{y_2 - y_1}{y_1 - y_0}, \frac{y_3 - y_2}{y_2 - y_1}, \dots, \frac{y_n - y_{n-1}}{y_{n-1} - y_{n-2}}\right\}, (i = 1, 2, \dots, n-1)$.

Then t_i values in Equations (6) can be estimated by $\hat{t}_i = \frac{y_i - \hat{\phi}_0 y_{i-1} - T_0}{S_0}$ and the revised estimated values of coefficients α_i and β_i are obtained as follows:

$$\hat{\alpha}_i = (2/k)\hat{t}_i / \{1 + (1/k)\hat{t}_i^2\}^2 \text{ and } \hat{\beta}_i = 1 / \{1 + (1/k)\hat{t}_i^2\}^2. \quad (10)$$

Realize that the MML estimators do not have bounded influence functions so coefficients α_i and β_i are revised to make them bounded. Besides, they completely depend on the observations not the presumed values of the parameter ϕ and p now.

Replacing Equations (9) and (10) in the modified likelihood equations, the adaptive modified maximum likelihood estimators are obtained as exactly the same as MML estimators given in Equation (7) except the coefficients α_i and β_i are replaced by coefficients $\hat{\alpha}_i$ and $\hat{\beta}_i$ and the concomitants $(y_{[i]}, y_{[i]-1})$ are replaced by the original observations (y_i, y_{i-1}) since t_i values are not obtained from the quantiles of the distribution but estimated directly from the sample and complete sums are invariant to ordering.

Computations: i) First by using sample observations calculate the initial values of $\hat{\phi}_0$, T_0 and S_0 ii) use these initial values to calculate $\hat{\mu}$, $\hat{\phi}$ and $\hat{\sigma}$ from Equations (7) by the use of coefficients $\hat{\alpha}_i$ and $\hat{\beta}_i$ given in Equations (10); iii) replace $\hat{\phi}_0$, T_0 and S_0 by $\hat{\phi}$, $\hat{\mu}$ and $\hat{\sigma}$, respectively; iv) repeat the process one more time and calculate $\hat{\mu}$, $\hat{\phi}$ and $\hat{\sigma}$ which are the desired AMML estimators.

3 Efficiency and Robustness Comparisons of the Estimators

To evaluate the efficiency and robustness of the AMML, MML and LS estimators, [100,000/n] (integer value) Monte Carlo runs (simulations) are used. The distribution of ϵ , known as population model, is taken as LTS with $p = 16.5$. As alternative sample models first

- (1) Normal with mean 0 and variance σ^2 , and the LTS family with
- (2) $p = 5.0$;
- (3) $p = 3.5$;
- (4) $p = 2.5$;
- (5) $p = 2.0$

Table 1. Simulated Values of the Mean and Variance of LS, MML and AMML Estimators under Alternative Models (1) – (5)

n		μ	ϕ	σ	μ	ϕ	σ	μ	ϕ	σ
			LS			MML			AMML	
Model (1)										
30	Mean	0.00	0.42	0.97	0.01	0.42	1.04	0.00	0.42	0.92
	Var	1.44	0.83	0.50	1.47	0.87	0.63	1.48	0.84	0.52
50	Mean	0.00	0.45	0.98	0.00	0.45	1.06	0.00	0.45	0.94
	Var	1.36	0.79	0.51	1.30	0.86	0.63	1.40	0.80	0.53
100	Mean	0.00	0.48	0.99	0.00	0.48	1.06	0.00	0.48	0.96
	Var	1.10	0.71	0.49	1.23	0.78	0.58	1.12	0.72	0.52
Model (2)										
30	Mean	0.00	0.42	0.97	0.00	0.42	1.01	0.00	0.42	0.89
	Var	1.46	0.84	0.77	1.31	0.80	0.75	1.38	0.82	0.61
50	Mean	0.00	0.45	0.98	0.00	0.46	1.03	0.00	0.45	0.91
	Var	1.27	0.76	0.81	1.24	0.79	0.77	1.17	0.73	0.62
100	Mean	0.00	0.48	0.99	0.00	0.48	1.03	0.00	0.48	0.93
	Var	1.17	0.73	0.83	1.05	0.69	0.77	1.10	0.69	0.62
Model (3)										
30	Mean	0.00	0.41	0.97	0.00	0.42	0.97	0.00	0.42	0.86
	Var	1.39	0.79	1.00	1.20	0.75	0.75	1.24	0.75	0.64
50	Mean	0.00	0.45	0.98	0.00	0.46	1.02	0.00	0.45	0.88
	Var	1.28	0.78	1.17	1.09	0.72	0.95	1.09	0.72	0.62
100	Mean	-0.01	0.48	0.98	0.00	0.47	1.02	-0.01	0.48	0.89
	Var	1.06	0.78	1.01	0.97	0.66	0.81	0.90	0.71	0.78
Model (4)										
30	Mean	0.00	0.42	0.95	0.00	0.43	0.94	0.00	0.43	0.80
	Var	1.43	0.82	1.91	1.09	0.73	1.03	1.09	0.73	0.65
50	Mean	0.00	0.45	0.96	0.00	0.46	0.98	0.00	0.45	0.82
	Var	1.29	0.79	1.91	0.99	0.68	1.29	0.98	0.67	0.65
100	Mean	0.00	0.47	0.97	0.00	0.48	0.97	0.00	0.48	0.83
	Var	1.07	0.78	2.16	0.88	0.59	1.14	0.82	0.64	0.61
Model (5)										
30	Mean	0.00	0.42	0.92	0.00	0.43	0.86	0.00	0.43	0.70
	Var	1.41	0.77	4.02	0.81	0.67	1.95	0.80	0.66	0.56
50	Mean	0.00	0.45	0.94	0.00	0.46	0.90	0.00	0.46	0.71
	Var	1.26	0.72	5.80	0.77	0.62	1.90	0.71	0.56	0.57
100	Mean	0.00	0.47	0.96	0.01	0.48	0.91	0.00	0.48	0.73
	Var	1.05	0.68	6.29	0.68	0.51	2.82	0.57	0.50	0.55

are used. The results for $n = 30, 50$ and 100 where $\phi = 0.5$, $\mu = 0.0$ and $\sigma = 1.0$ are given in Table 1.

It can be seen from Table 1 that all estimators are unbiased except AMML estimate of σ which is slightly less and its bias decreases as sample size increases as expected. However in all cases the variances of AMML estimates are smaller than that of the LS estimates and similar with MML estimates except Model (1) where they are close. Thus, although AMML slightly under-estimates σ , its mean squared error is less than the others. Realize that all methods underestimate σ in Model (5). Therefore, we can conclude that AMML is efficient and robust to misspecification errors. Similar results are obtained for other presumed values of ϕ , μ and σ so they are not reported here for conciseness.

Then, the outlier models where $(n - r)X_i$ come from $N(0, \sigma^2)$ and r (we do not know which) come from

$$(6) N(0, 4\sigma^2);$$

$$(7) N(0, 16\sigma^2); r = [0.5 + 0.1n] \text{ (integer value),}$$

and the mixture models

$$(8) 0.90N(0, \sigma^2) + 0.10N(0, 4\sigma^2);$$

$$(9) 0.90N(0, \sigma^2) + 0.10N(0, 16\sigma^2)$$

are taken as alternative sample models. The innovations are scale corrected to make their variances equal to σ^2 . The results for $n = 30, 50$ and 100 where $\phi = 0.5$, $\mu = 0.0$ and $\sigma = 1.0$ are given in Table 2.

It can be seen from Table 2 that the results are similar to misspecification ones given in Table 1. Finally, the extreme alternative sample models

$$(10) \text{ Student's } t \text{ distribution with 2 degrees of freedom;}$$

$$(11) \text{ Cauchy distribution; and}$$

$$(12) \text{ Slash (Normal/Uniform) distribution}$$

are taken as alternative sample models. It must be noted that model (10) has finite mean but non-existent variance, and models (11)-(12) have non-existent mean and variance. Since the differences between the AMML and the others become very striking due to exploding variances, only the results for AMML estimators are given in Table 3. However, it must be noted that Model (10) is still comparable due to negligible bias in μ and ϕ with unacceptable variances. Besides, σ is overestimated by the other methods in this case again with unacceptable variances. Therefore, under such extreme alternatives, only AMML method is valid and robust.

Table 2. Simulated Values of the Mean and Variance of LS, MML and AMML Estimators under Alternative Models (6) – (9)

n		μ	ϕ	σ	μ	ϕ	σ	μ	ϕ	σ	
			LS			MML			AMML		
Model (6)											
30	Mean	0.00	0.41	1.10	-0.01	0.42	1.15	0.00	0.41	1.01	
	Var	1.83	0.95	0.93	1.80	0.89	0.83	1.72	0.93	0.70	
50	Mean	0.00	0.45	1.12	0.00	0.45	1.18	0.00	0.45	1.03	
	Var	1.60	0.95	0.90	1.51	0.91	0.83	1.49	0.90	0.72	
100	Mean	0.00	0.47	1.13	0.00	0.48	1.18	0.00	0.47	1.05	
	Var	1.39	1.06	0.95	1.39	0.95	0.86	1.28	0.98	0.72	
Model (7)											
30	Mean	0.00	0.40	1.05	0.00	0.40	0.98	0.00	0.40	0.79	
	Var	1.63	1.55	2.08	1.05	1.17	0.95	1.05	1.18	0.54	
50	Mean	0.01	0.43	1.05	0.00	0.45	1.03	0.00	0.43	0.81	
	Var	1.40	1.86	2.17	0.95	1.36	1.32	0.94	1.29	0.54	
100	Mean	0.00	0.46	1.09	-0.01	0.47	1.02	0.00	0.46	0.83	
	Var	1.26	2.17	2.48	0.79	1.53	1.18	0.74	1.44	0.59	
Model (8)											
30	Mean	0.00	0.42	0.97	0.00	0.42	1.01	0.00	0.42	0.88	
	Var	1.48	0.81	0.76	1.29	0.80	0.75	1.39	0.80	0.58	
50	Mean	0.00	0.45	0.98	0.00	0.46	1.03	0.00	0.45	0.91	
	Var	1.38	0.80	0.80	1.13	0.78	0.74	1.26	0.77	0.57	
100	Mean	0.00	0.47	0.99	0.00	0.48	1.04	0.00	0.47	0.92	
	Var	1.17	0.78	0.89	1.04	0.75	0.72	1.10	0.74	0.63	
Model (9)											
30	Mean	0.00	0.43	0.95	0.00	0.43	0.88	0.00	0.44	0.71	
	Var	1.35	0.73	2.63	0.82	0.63	1.23	0.77	0.59	0.66	
50	Mean	0.00	0.45	0.96	0.00	0.46	0.94	0.00	0.46	0.73	
	Var	1.22	0.73	2.63	0.85	0.58	1.72	0.67	0.56	0.60	
100	Mean	0.00	0.48	0.97	0.00	0.48	0.92	0.00	0.49	0.74	
	Var	1.03	0.80	2.78	0.67	0.49	1.50	0.59	0.55	0.63	

Table 3. Simulated Values of the Mean and Variance of AMML Estimators under Alternative Models (10) – (12)

		Model (10)			Model (11)			Model (12)		
N		μ	ϕ	σ	μ	ϕ	σ	μ	ϕ	σ
30	Mean	0.00	0.45	1.37	-0.01	0.47	1.94	0.01	0.47	2.68
	Var	2.95	0.54	2.96	6.32	0.28	11.73	11.08	0.27	19.08
50	Mean	0.00	0.47	1.40	0.01	0.48	1.93	-0.01	0.49	2.71
	Var	2.59	0.46	2.49	4.55	0.19	10.43	8.95	0.18	16.97
100	Mean	0.00	0.48	1.42	-0.01	0.49	1.92	0.01	0.49	2.72
	Var	2.32	0.30	2.80	4.29	0.09	9.54	7.56	0.10	16.96

4 Conclusion

In this study, for AR(1) models, MML technique is adapted to machine data processing where the distribution family is known rather than the exact distribution. For this purpose, the idea of Huber M-estimation is inserted to MML technique. Then, the efficiency and robustness properties of the most widely used LS estimators, MML and AMML estimators are examined through simulations and observed that MML and AMML estimators are more efficient than LS estimators as expected. However, AMML underestimates σ in all cases but have much smaller variances than the others yielding less mean squared errors. Therefore, if there is an opportunity to examine the distribution, one should prefer the use of MML rather than AMML. Otherwise, like in machine data processing, one can safely use AMML estimators having in mind the bias in σ which cannot be corrected since the exact distribution is not known.

References

1. Akkaya AD, Tiku ML (2001) Estimating parameters in autoregressive models in non-normal situations: asymmetric innovations. *Commun Stat-Theor M* 30:517-536. doi:10.1081/STA-100002095
2. Akkaya AD, Tiku ML (2005) Time series AR(1) model for short-tailed distributions. *Statistics* 39:117-132. doi: 10.1080/02331880512331344036
3. Akkaya AD, Tiku ML (2008) Autoregressive models in short-tailed symmetric distributions. *Statistics* 42(3):207-221. doi:10.1080/02331880701736663
4. Bayrak OT, Akkaya AD (2010) Estimating parameters of a multiple autoregressive model by the modified maximum likelihood method. *J Comput Appl Math* 233:1762-1772. doi:10.1016/j.cam.2009.09.013
5. Dönmez A (2010) Adaptive estimation and hypothesis testing methods. PhD thesis, Middle East Technical University.
6. Hamilton LC (1992) Regression with graphics. Brooks/Cole Publishing Company, California.
7. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. John Wiley, New York.
8. Huber PJ (1981) Robust statistics. John Wiley, New York.
9. Tiku ML (1964) Estimating the mean and standard deviation from a censored normal sample. *Biometrika* 54:155-165. doi:10.2307/2333859
10. Tiku ML, Akkaya AD (2004) Robust estimation and hypothesis testing. New Age International (P) Ltd., New Delhi, India.
11. Tiku ML, Sürücü B (2009) MMLs are as good as M-estimators or better. *Stat Probabil Lett* 79(7):984-989. doi:10.1016/j.spl.2008.12.001
12. Tiku ML, Tan WY, Balakrishnan N (1986) Robust inference. Marcel Dekker, New York.

13. Tiku ML, Wong, WK, Bian, G (1999) Estimating parameters in autoregressive models in non-normal situations: symmetric innovations. *Commun Stat-Theor M* 28:315-341. doi:10.1080/03610929908832300
14. Tiku ML, Wong WK, Vaughan DC, Bian G. (2000) Time series models in non-normal situations. *J Time Ser Anal* 21:571-596. doi:10.1111/1467-9892.00199
15. Ülgen BE (2010) Robust estimation and hypothesis testing in microarray analysis, PhD thesis, Middle East Technical University.
16. Vinod HD, Shenton LR (1996) Exact moments for autoregressive and random walk models for a zero or stationary initial value. *Economet Theor* 12:481-499. doi:10.1017/S0266466600006824

Modeling and analysis of the cosmic rays variations during periods of heliospheric disturbances on the basis of wavelet transform and neural networks

O.V. Mandrikova¹, T.L. Zalyaev¹

¹ Institute of Cosmophysical Researches and Radio Wave Propagation (IKIR FEB RAS),
Mirnaya Str., 7, Kamchatka Region, Elizovskiy District, Paratunka 684034, Russia

² Kamchatka State Technical University, Klyuchevskaya Street 35, 683003 Petropavlovsk-
Kamchatsky, Russia

In this work, using the cosmic ray (CR) data as an example, we proposed a method for analysis of time series with a complex structure, method is based on combination of orthogonal multiple-scale analysis (MRA) and multilayer neural networks (NN). The method allows to perform an analysis of the typical variations of the data, to allocate anomalous changes and to obtain quantitative estimates of their parameters. Proposed method includes following operations: (1) on the basis of MRA the data is represented as: where $f_{a,(-m)}(t) = \sum_n c_{-m,n} \phi_{-m,n}(t)$, $f_{a,j}(t) = \sum_n d_{j,n} \Psi_{j,n}$, $\Psi_j = \{\Psi_{j,n}\}_{n \in \mathbb{Z}}$ – wavelet basis, $\phi_j = \{\phi_{j,n}\}_{n \in \mathbb{Z}}$ – scaling function basis; (2) for isolated smoothed components $f_{a,(-m)}$ on the basis of NN the mapping is performed: $\gamma: f_{a,(-m)} \rightarrow \overline{f_{a,(-m)}}$, where $f_{a,(-m)}$ – input values of NN, $\overline{f_{a,(-m)}}$ – output values of the NN; (3) analysis of the error vector of the NN: $e(n) = \overline{c_{-5,n}} - c_{-5,n}$, where $c_{-5,n}$ – real value of the coefficient in the time moment $t = n$, $\overline{c_{-5,n}}$ – extrapolated value; anomalous changes are allocated on the basis of the condition: $|e(t)| \geq T_s$, where T_s is threshold function for the corresponding NN. In work on the example of data processing of ground stations of neutron monitors, a model of the time course of cosmic rays is constructed, having the form $\overline{c_{-m,n+1}} = \phi_k^3(\sum_i \omega_{ki} \phi_i^2(\sum_l \omega_{il} \phi_l^1(\sum_{z=0}^{\gamma} \omega_{ln} c_{-m,n-z})))$, where $\omega_{ln}, \omega_{il}, \omega_{ki}$ – weight coefficients of the neurons of NN; $\phi_l^1 = \phi_l^2 = \frac{2}{1+e^{-2z}} - 1$; $\phi_k^3 = a * z + b$; γ – length of the input vector of NN (we used $\gamma = 6$).

For the analysis of subtle features of the time series used computing solutions based on continuous wavelet transform (CWT) and threshold functions: (1) continuous wavelet transform of data is formed: $f_{HM}: (W_{\Psi} f_{b,s}) := |s|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} f(t) \Psi\left(\frac{t-b}{s}\right) dt$, where $f \in L^2(R)$, $s, b \in R$, $s \neq 0$, Ψ – wavelet basis, parameter s characterizes the scale, b – time; (2) a threshold function is applied to the obtained array of wavelet coefficients:

$$W_{\Psi} f_{b,s}: P_{T_s}(W_{\Psi} f_{b,s}) = \begin{cases} W_{\Psi} f_{b,s}, & \text{if } (W_{\Psi} f_{b,s} - W_{\Psi} f_{b,s}^{med,l}) \geq T_s^l \\ 0, & \text{if } |W_{\Psi} f_{b,s} - W_{\Psi} f_{b,s}^{med,l}| < T_s^l \\ -W_{\Psi} f_{b,s}, & \text{if } (W_{\Psi} f_{b,s} - W_{\Psi} f_{b,s}^{med,l}) < -T_s^l \end{cases} \quad (1)$$

where $W_{\Psi}f_{b,s}^{med,l}$ – is the median value calculated in a sliding time window of length l .
 $T_s^l = U * \sigma_s^l$ – is threshold function, where $\sigma_s^l = \sqrt{\left(\frac{1}{l} - 1 \sum_{k=1}^l (W_{\Psi}f_{b,s} - \overline{W_{\Psi}f_{b,s}})^2\right)}$ – is standard deviation, calculated in a sliding time window of length l , $\overline{W_{\Psi}f_{b,s}}$ – is average value, U – threshold coefficient (we used $U = 2.5$); the application of function $P_{T_s}(W_{\Psi}f_{b,s})$ allows us to allocate periods of abnormal decreases and abnormal increases in the time series. To evaluate the intensity of anomaly at the time $t = b$ we used the value: $Y_b = \sum_s P_{T_s}(W_{\Psi}f_{b,s})$, which in the case of abnormal increases will be positive, and in the case of abnormal decreases — negative. The result of the application of the method is presented on figures 1 and 2.

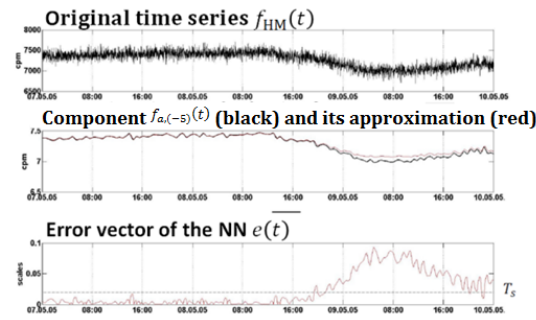


Fig. 1. Analysis of cosmic rays data on the basis of neural network

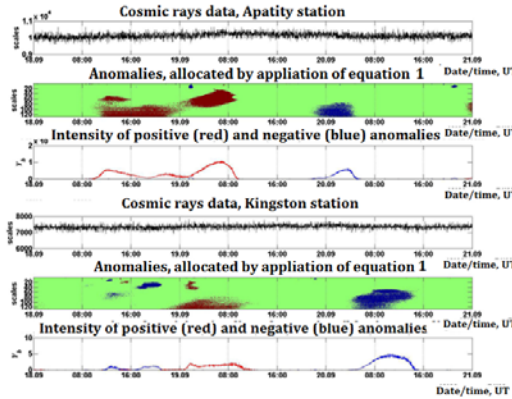


Fig. 2. Detailed analysis of cosmic rays data

On the basis of the data processing during periods of increased solar and geomagnetic activity we confirmed the possibility of an anomalous increase in CR intensity (CR pre-increase) a few hours before the beginning of geomagnetic storms.

These effects can be used as precursors of strong geomagnetic storms (act as additional factor) in the problem of space weather forecast, the solution of which is not satisfactory at the moment

This work was supported by a grant from RSF 14-11-00194.

MULTIDIMENSIONAL TIME-FREQUENCY ANALYSIS OF THE CAPM

Roman MESTRE *, Michel TERRAZA **

LAMETA University of Montpellier

Abstract. The CAPM theory provides a measure of the sensitivity of an asset to the market called the systematic risk. The Beta of an equity is estimated by its market line. According to the OLS hypothesis, it is stable over time but this is not empirically verified. Many studies are in favor with this fact (Unstable Beta), and more particularly the Beta dispersion according to the frequencies which is related to the heterogeneous behavior of agents. Using the wavelets method, we can calculate the coherence and the phase between the stock's returns and those of the market over time, it is also possible to visualize it. In order to confirm the correctness of the methodology, we use three french equities with different Betas (AXA, LVMH and Orange) for the period from 2005 – 2015 including the crisis. We show that the wavelets coherence, associated with the phase, improve our understanding and the classification of equities according to there characteristics. Our study reveal the contagion and interdependence phenomenons between the stocks and the market. The contagion effects (from the market to the stock) is principally located on the High-Frequencies whereas the interdependence effect is located on the Low-Frequencies (Long-run investment). The link between beta and coherence-phase can help the investors to choose more efficiently the time they should invest.

Keywords: Wavelets; Coherence; CAPM; Co-Movement

* Corresponding author, email: roman.mestre@lameta.univ-montp1.fr
LAMETA, University of Montpellier, UFR Economie, Site de Richter, Avenue
Raymond Dugrand, CS 79606, 34960 Montpellier, Cedex 2, France
** LAMETA, University of Montpellier

For more efficiency, the portfolio managers must have the higher level of returns with the lesser risk. They can use the CAPM of Sharpe (1962) providing a measure of risk (The Beta) corresponding to the sensitivity of an asset to the market. In the CAPM theory, the risk premium of an asset (return above the risk free rate) are related with the market premium. The main equation of the CAPM is the Securities Market Line (SML):

$$r_{i,t} = \alpha + \beta \cdot r_{m,t} + \epsilon_t \quad (1)$$

With $r_{i,t}$ the risk premium of asset i and $r_{m,t}$ market premium, ϵ_t is an $i.i.d(0, \sigma_\epsilon)$ process.

The systematic risk of an asset is given by the β of the SML. The higher Beta, the more sensitive is the equity to the market movements (the system). The β is a regression parameter, so it constant by hypothesis, it is similar for the correlation between $r_{i,t}$ and $r_{m,t}$. However, due to the erratic market fluctuations, the equity-market link does not have the same intensity over time. On the other hand, an action may be weakly correlated with the market in long-run but have a more tight link in short run. When agents are informed about the variability of the Action-Market relationship, they can adjust their portfolios more optimally. The volatility of Beta is thus related to the hypothesis of heterogeneous behaviors of agents¹.

To analyze the Beta instability phenomenon, we use the technique of wavelets in order to calculate the coherence between two time series. This instrument was developed by Haar in 1909, popularized by Morlet and Grossmann in 1984 (which give the name "wavelet"), Meyer in 1986 – 1987² and Mallat [1989-2009], over-takes the limits of the spectral analysis, in particular its timelessness, and reduces time-frequency arbitration. In finance, wavelets play an important role because they become the preferred tool to take into account the behavioral hypothesis of agents. Univariate wavelet approaches have been used in the analysis of exchange rate volatility and in the construction of a VaR that takes into account the heterogeneity of agents, called WVaR (Wavelets Value-at-Risk).

¹ Cf. Bibliography[14]

² Abel Prize laureate in Mathematics 2017

The wavelets were used by Gencay et al (2005) to estimate the systematic risk of US equities, they show the possibility of estimate the Beta by frequency. In addition, Auth (2013) used wavelet coherence to appreciate the links between hedge fund portfolios and other financial assets. It is thus possible to show the factors influencing the diversification of these portfolios.

The multivariate extensions with the wavelet coherence give the possibility to dynamically appreciate the frequency causality . The financial series have,indeed rarely a constant variance over time and hence the co-spectral analysis ³ is not a reliable method. In this multivariate time-frequency case, the wavelets coherence can be compared to the correlation by frequency over time between two stationary signals. Coherence associated with the phase calculation, test the significance of the relationships (links) between variables and give the sens of the causality.

The main goal of this article is to demonstrate that the wavelet coherence-phase association is a powerful tool for measuring risk for investors according to their investment horizons. For this purpose, we apply it to the stocks prices of AXA, LVMH and Orange equities listed on the French market (the CAC40 index are used as a reference) for the daily period from 2005 to 2015. The equities are selected according to their betas values (estimated by OLS) : 1.51 for AXA, 1 for LVMH and 0.72 for Orange. Supposing that the Beta of the market is equal to 1, AXA is an highly sensitive stock overreacting the market movement , Orange is relatively less sensitive to the market fluctuation (under reaction), and LVMH "follows" the market.

We present in a first part a synthesis of the methodology used before applying it to the selected data, we conclude on the benefits of using the wavelet coherence-phase method to manage the portfolio.

³ Cf. Bibliography [12; 13]

1 Time-Frequency analysis

The Wavelets

The wavelets improve the interpretation of the Fourier coherence for which the temporal information is lost for more information in the frequency domain. The first works on this, are those of Gabor (1946) concerning the Short-Time Fourier Transform (STFT) using a constant size rolling window. With this method we choose the resolution degree about the temporal or frequency accuracy, but we don't solve totally the arbitration between Time information and Frequency information. The main parameter of this method is the window size, we have a good frequency localization at the detriment of the temporal information if the window is large. The largest window, the better frequency localization at the detriment of the temporal information. To the contrary, a tight window improve the temporal resolution but deteriorate the frequency aspects. This fact is a Physics Dilemma explain by Heisenberg (eponymous Uncertainty Principle): It is impossible to accurately and simultaneously measure the position and momentum of a particle. In our case, we can calculate, concomitantly, the temporal and frequency position of a variable. The Wavelets are constrained by this principle but they "reduce" the arbitration effect by providing simultaneous time-frequency analysis.

A Wavelet $\varphi(t)$ (the wavelet-mother), is similar to the window. Its mean equals zero and it preserves the energy (variance) of a time series. The $\varphi(t)$ function is translated by τ and dilated by s in order to extract frequency information from the series at a specific moment t of time. The information is stocked in the wavelet-family $\varphi_{s,\tau}(t)$, which collecting all the translated and dilated version of $\varphi(t)$.

A wavelet mathematically is defined by the following equation:

$$\varphi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\varphi\left(\frac{t-\tau}{s}\right) \quad (2)$$

The wavelets transform (or wavelets decomposition) project the function $x(t)$ on the wavelet family. This method indicate how the wavelet-mother generates the wavelets and it providing the wavelets coefficients $W(s, \tau)$. the coefficients reproduce the variations of the series in the neighborhood of $t \mp \tau$ with a frequency width s . By

varying τ and s we have ,consequently, the temporal variations of the series for a given frequency scale, justifying the name "Time-Frequency Analysis". By varying τ and s we have ,consequently, the temporal variations of the series for a given frequency scale, justifying the name "Time-Frequency Analysis".

Formally we have the following result ⁴ :

$$W(s, \tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} x(t) \varphi^*\left(\frac{t - \tau}{s}\right) dt \quad (3)$$

φ^* is the complex-conjugate of φ .

We can find again the series with the inverse wavelet transform. Its mathematical equation shows the admissibility condition C_φ of the wavelet.

$$x(t) = \frac{1}{C_\varphi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |W(s, \tau)| \varphi\left(\frac{t - \tau}{s}\right) d\tau \frac{ds}{s^2} \quad (4)$$

This condition ensures the nullity of the mean and the energy preserving during the decomposition.

$$C_\varphi = \int_{-\infty}^{+\infty} \frac{|\widehat{\varphi(f)}|}{|f|} df < \infty \quad (5)$$

$\widehat{\varphi(f)}$ is the Fourier transform of $\varphi(f)$

This relation must respect the following constraints:

$$\int_{-\infty}^{+\infty} \varphi(t) dt = 0 \quad (6)$$

$$\int_{-\infty}^{+\infty} |\varphi(t)|^2 dt = 1 \quad (7)$$

The Continuous Wavelet Transform (CWT) needs a large number of observations, for its implementation. Its discrete version (Discrete Wavelet Transform) reduce the number of frequencies by a lesser variation step of s and τ . By reducing the variation-step, we have

⁴ Cf. Mallat [9;10;11]

an optimal sampling for reduce the calculation. This discretization impose a dyadic scale (the number of observations N must be a multiple of 2), and an High-Pass and Low-Pass Filters determining the the order of decomposition J (depth) which indicate the number of frequency bands. In this case $J = \text{Log}_2(N)$, with N is the number of observation. This process, called Pyramidal Algorithm, is created by Mallat (the name "Mallat Algorithm" is also used). There is an alternative version of the DWT without the dyadic length constraint , called the Maximal Overlap Discrete Wavelets Transform (or MODWT)⁵.

The different wavelets transformations distinguish themselves by their wavelet-mother (and associated family) and its characteristics, defining also the wavelet filter properties. Each wavelets family have their own properties and specificities, as the orthogonal, the symmetry, etc. The Daubechies wavelet family is commonly used, or also the Morlet Wavelet (called Mexican-hat wavelet). In a family, there are differences between the wavelets about the number of vanish moments (equal to zero). For instance, the "D8" wavelet is a Daubechies wavelet with 8 moments equal to zero whereas the "La8", even if it has 8 null moments, are not in the same family but in th Least Asymmetric Daubechies family. More details about the wavelets properties are given by Farge(1992) and Daubechies (1992).

We use, in this paper, the CWT with the Morlet complex wavelet because it is a good balance between temporal and frequency localization. The CWT is, moreover, has a finer frequency mesh (the frequency step). Its mathematical equation is:

$$\varphi_{s,\tau}^M(t) = \pi^{-1/4} e^{(if_0 t)} e^{(-t^2/2)}. \quad (8)$$

f_0 is the center frequency, equal to 6 in our case in order to respect the admissibility condition.

The calculation softwares realize, in practice, a frequency sampling for calculate the CWT, because we are constrain by the calculation power of the computers concerning the infinite integral. In theory, the depth of the decomposition J is defined only by the frequency step δ_j .

⁵ Cf. Mallat [9;10;11] and Gencay et al [7] for more details about the MODWT

In the continuous case, we can arbitrarily choose it, but in practice, J is related to the size of the series (N). This order is important to define the graphical resolution and reduce the calculation time. Torrence et Compo (1997) give the following formula for calculate the maximum frequency scale level:

$$J = \delta_j^{-1} \text{Log}_2(N\delta_t/s_0) = \delta_j^{-1} \text{Log}_2(N/2) \quad (9)$$

δ_j is the frequency step, δ_t the time step, and $s_0 = 2\delta_t$ the smallest scale. We can use commonly $\delta_j = 1/8$, we have a good frequency resolution with reasonable calculation time.

Lau et Weng(1995) gives the following formula for calculate the set of frequencies for the decomposition until the J order.

$$s_j = s_0 2^{j\delta_j}; j = 1, \dots, J \quad (10)$$

We notice that the values of s_j are related, for each scale j , to a time horizon (in the same time unity of the series).

The wavelets coherence

The wavelets coherence between two functions (with the same size N) also called *Time-Varying Coherence*, is based on a CWT using the Morlet Wavelets. Similar to the Fourier Coherence; we have a measure of wavelet spectral covariance defined by the cross wavelets spectrum $SW_{xy}(s, \tau)$:

$$SW_{xy}(s, \tau) = W_x(\tau, s).W_y^*(\tau, s) \quad (11)$$

$x(t)$ et $y(t)$ two temporal functions, $W_x(\tau, s)$ is the wavelet coefficients of the CWT, and W_y^* the complex conjugate of $W_y(\tau, s)$.

$|W_{xy}(s, \tau)|$ cross power spectral between $x(t)$ and $y(t)$. Associated with the auto-power spectrums, we define the wavelet coherence:

$$WQ(f) = \frac{|\Im(s^{-1}.SW_{xy})|^2}{\Im(s^{-1}.|SW_x|^2).\Im(|s^{-1}.SW_y|^2)} \quad (12)$$

$WQ(f)$ formula is similar to the determination coefficient formula. For each frequency scales s and for each moment t , we have a coefficient between 0 and 1 corresponding to the greater or lesser

squared correlation between the two signals and of course the explanatory power of the model. These coefficients, coming from the wavelets decomposition by Morlet Wavelet, are complex in nature. Consequently, the wavelet coherence is equal to 1 in the real space, regardless the time τ . Requiring the use of a time-frequency smoothing \mathfrak{S} to get the true values of the coherence. The temporal smoothing for a frequency scale given is \mathfrak{S}_{time} , and the frequency smoothing for a given time t is \mathfrak{S}_{scale} .

The general smoothing operator is:

$$\mathfrak{S}(W) = \mathfrak{S}_{scale(\mathfrak{S}_{time}(W))} \quad (13)$$

The equations of \mathfrak{S}_{scale} and \mathfrak{S}_{time} are given by Torrence et Webster (1998):

$$\mathfrak{S}_{time}(W_N) = W_N \cdot c1^{-t^2/2s^2} \quad (14)$$

$$\mathfrak{S}_{scale}(\cdot) = W_N \cdot c2 \Pi(0.6s) \quad (15)$$

$c1$ and $c2$ are normalizing constants and Π is the rectangle function⁶.

The wavelets provide a measure of the phase between two functions, so we can appreciate the positive or the negative correlation between the frequency components or mutual interactions (causality). The wavelet phase function $\theta_W(f)$ is the ratio between the imaginary part C and the real part P of SW_{xy} :

$$\theta_W(f) = \arctan(C(SW_{xy}f)/P(SW_{xy}f)) \quad (16)$$

In practice x_t et y_t are samples of two random processes, with time-step and frequency-step define by the CWT previously described.

⁶ the rectangle function is a function equal to a in $[-1/2, 1/2]$ interval and equal to zero outside

2 Results and discussions

- The betas values, coming from the SML estimate by OLS, and the corresponding R^2 are given in the table 1:

Table 1: Estimates Results

Stocks	Beta	R^2
AXA	1,5	0.68
LVMH	1	0.62
Orange	0.73	0.43

The appendix 1 provides more details about these estimates: the variables are stationary, the residuals are homoscedastic and autocorrelation and no follow a Normal distribution. So the betas don't respect the required properties about the BLUE estimator, but we consider them as references.

The table 2 gathers the results in the frequency space about the cross-spectral analysis between the stocks and the *CAC40*. This table provides the mean and the standard-deviation of the coherence and also the mean of the absolute value of the phase and its standard-deviation.

Table 2: Characteristics of the coherence and the phase from the cross-spectral analysis

Stocks	Average Coherence	standard-deviation
AXA	0.71	0.17
LVMH	0.65	0.18
Orange	0.48	0.21

Stocks	Average Coherence	standard-deviation
AXA	0.21	0.17
LVMH	0.23	0.19
Orange	0.34	0.32

AXA and LVMH have an average coherence higher than Orange, in accordance with their R^2 (in table 2). In addition, the Orange coherence is a more erratic than the other two stocks (see their standard deviations), illustrating a greater variability of its link with the market over the frequencies. The appendix 3 confirms with the graphics these results. The absolute mean phase values are larger (and more volatile) for Orange, the phase-shift illustrates how Orange reacts(responds) to the market.

A stock with a low beta has a lower coefficient of determination and a lower average coherence than an stock with a strong beta. The explanatory power of the market (on the stock fluctuations) is therefore more important for a high stock beta. But we can not verify if this finding is valid whatever the period considered and the chosen investment horizon. This is a limit of the cross-spectral analysis (timelessness).

- We can improve this static method by using the Multidimensional Time-Frequency Analysis. The following figures illustrate, for each stock, the dynamic evolution of the coherence and the phase simultaneously.

The frequency scales (in days) are indicated in the y-axis and the Time in the x-axis starting from 0 (first observation) to 2869.

The wavelet coherence intensity is given by the following colors: red illustrates a high correlation between the stock and the CAC40 whereas blue indicates a weakest link. The bold lines delineate the areas for which the correlation (the R^2 in this case) is significant at the 5% risk level (using Monte-Carlo simulations). The white transparent surface is the "cone of influence", where edge effects can bias results about cross-spectrum because we have finite samples.

The arrows orientation simplifies the phase analysis: if an arrow points to the right then the series are *in-phase*, whereas if arrows point to the left then are *out-of-phase*.

The coherence and the phase give a simultaneous representation of Equity-Market relationship. The phase can be interpreted as the sign of the correlation, two series are positively correlated if it is in phase whereas it is negatively correlated when it is out-of-phase.

The phase can be interpreted as the sign of the correlation, two series are positively correlated if it is in phase whereas it is negatively correlated when it is out-of-phase. With the phase we can also appreciate the nature of the interdependencies between the series: an arrow pointing up indicates that x_t leads y_t to an high correlation, so x_t is the "Leader". Conversely, an arrow pointing down illustrates that y_t is the Leader (y_t leads x_t). The phase in a way constitutes a type of the causality measure.

Figures 1 : Wavelet Coherences-phases for the 3 stocks

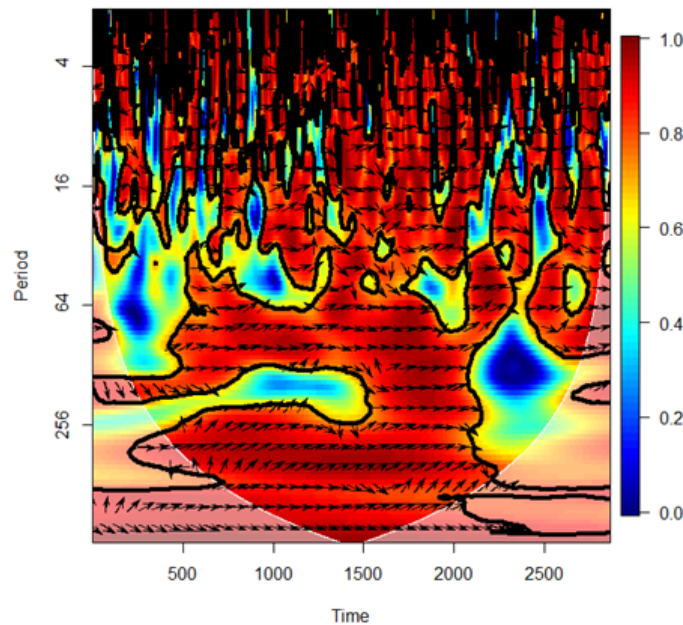


Figure 1.1 *AXA – CAC40*

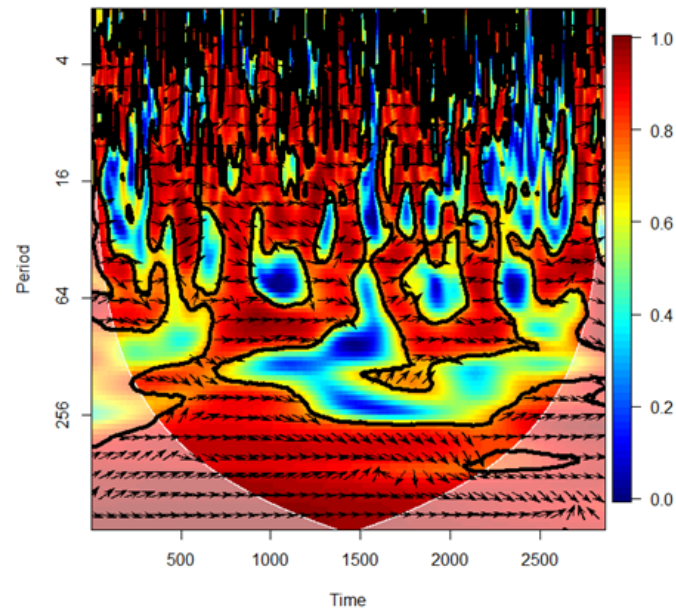


Figure 1.2 *LVMH – CAC40*

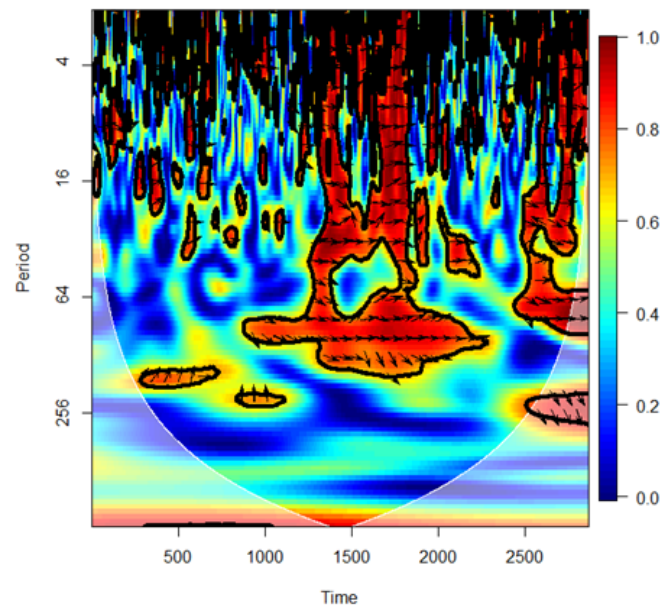


Figure 1.3 *Orange – CAC40*

This figures show a predominance of red in the coherences of AXA and , in a lesser extent, LVMH ,indicating an high correlation with the market. For the high frequencies, the coherence is not continuous because we can notice blue and red areas alternating. The blue is predominant in the Orange coherence but we remark a tight red area. Overall, AXA and LVMH are strongly correlated with the market whatever frequencies, to the contrary, Orange is not correlated. The high beta equities have, generally, a coherence with red predominance whatever the frequencies and the period. Conversely, a low beta equity has a blue predominance coherence, but with co-movement area on particular frequency and time.

To highlight and synthesize the main points of these results, we divide the global period in 5 sub periods:

- The ante crisis period extends from 2005 to the beginning of 2008 (around 800 days on the x-axis).
- The subprimes crisis in 2008 until the end 2010 (from 800 to 1300 days on the x-axis)
- The debt crises in 2011 – 2012 (from 1700 to 2000 days on the x-axis).
- The post-crisis period begin at the middle of 2012 until the end of our sample in 2015.

The Table 3 summarizes the intensity of the correlation (High, Medium,Low) for each stocks, the letter M indicates the market leadership and S for the Stock Leadership. If we can clearly determinate the Leader, we use the letter I to indicate the interaction between the market and the stock .

Table 3: Coherences Syntheses

Stocks	Frequencies	2005-2008	Subprimes Crisis 2008-2010	2010-2011	Debt crisis 2011-2012	2012-2015
AXA	HF	High-M	High-M	High-M	High-M	High-I
	MF	Medium-I	High-M	High-I	High-M	High-I
	LF	High-M	High-M	High-I	High-M	**
LVMH	HF	High-M	Medium-M	Medium-M	High-M	Medium-M
	MF	High-I	High-I	Medium-I	High-M	Medium-I
	LF	High-M	High-S	High-I	High-S	High-I
Orange	HF	Medium-M	Medium-M	Forte-M	Medium-M	Medium-High-M
	MF	Low-I	Medium-Low-M	High-I	Medium-I	Medium-High-I
	LF	Low	Low	Low	Low	Low

HF=High-Frequencies, MF=Medium-Frequencies and LF=Low-Frequencies.

We can confirm the CAPM hypothesis on the **High-Frequencies**: The market fluctuations cause the stock movements, but each of them have a different link over time. AXA is more often correlated than Orange, this is logical according to their beta value. We highlight the *Contagion* effects from the market to the stocks. A stock has a contra-cyclical fluctuations if it is out-of-phase with the market (not observed in this paper), whereas, it has a pro-cyclical if it is in-phase. However in short-run (**High-Frequencies**), a stock can not impact the market (or rarely).

On the **medium-frequencies**, we notice that the market and the stock lead themselves alternatively (bilateral interactions). The market is not systematically the Leader. So, the Equity-Market relationship is not uniform and homogeneous in the time-frequencies space, revealing the interest of the wavelets method compared to the previous traditional tools. During the two crisis periods, the three stocks are correlated with the market, but differently. For instance, AXA has a high beta before the sub-primes crisis (leads the market), it tend to amplify the market fluctuations (CAPM theory) and impact it. The market reacts to the stock movements, and he drop sharply. It is the Leader during the Crisis, and leads the stock. This result is observed for the other stocks.

On the **Low-Frequencies**(below 256 scale on the y-axis), Orange is not correlated with the *CAC40* contrary to AXA and LVMH. The latter stocks are in phase with the market with more complex interactions. We can't, indeed, define clearly the leadership, the arrows pointing up and down alternatively. We conclude the existence of a high and relevant interdependence. In long-run the stocks and the market mutually interact and lead the relationship alternatively, resulting of the price adjustment on the market.

Conclusion

The wavelet coherence-phase is an efficient tool to better understand the Equity-Market interactions. It appears that it is indispensable method to supplement the SML estimation and to solve the main limit of the co-spectral analysis (atemporality).

The stocks selected are considered as "major" assets of the CAC because of their index weight relatively large (LVMH weight is 8%, AXA 4 – 5% and Orange 2 – 3%). Consequently, it is not surprising LVMH influences more the market than Orange. This is one of the CAPM limits, theoretically the Market should include all the assets. In fact, the *CAC40* is a french portfolio composed by 40 assets, and the estimated beta is a measure of the asset sensitivity (to the market movement) , but it represents $x\%$ of this one. This problem is lesser, in our case, by the coherence providing the areas and the periods in which the stock leads the market, so we can analyze the CAPM results more precisely. Moreover, we have an obvious interest in the frequency decomposition because each agent can make a classification of his equities according his investment horizon (the scale in the y-axis). Orange, for instance, is more correlated with the cac in short-run during tumultuous market period than in long-run. Consequently, a pension fund (investing for 5 years) will be less exposed than a HFT (High Frequency Trading) investing in short-run. The HFT can spot, with this method, the shortcomings (or where the stock is less correlated with the market) and exploit its.

This study illustrates the frequency links (relations) between an equity and its market, a improve the stocks-assets profiling:

- On the High-Frequencies, we notice, the Contagion phenomenon describing how the market fluctuations lead the stocks volatility. However, the equities are not highly correlated with the market at the time t systematically; in this case the market lead it to greater co-movement area.
- On the Low-Frequencies, there is a predominance of the Interdependence phenomenon with bilateral interactions between equities and the market. The market influences the stock's returns but this one react to assets volatility. There is a response process between the two variables, illustrating a different and more complex systematic risk. This results is significant for AXA and LVMH, Orange being not correlated with the *CAC40* on this frequencies.
- On the Medium-Frequencies, there are a different combinations of this two phenomenons according the asset profile.

Ultimately, there is a plurality of asset profiles to build well-diversified portfolios over time with different investment horizons. The coherence-phase of a portfolio is a new indicator of the well-diversification (or not) according the investors profiles and risk appetites.

Appendixes

A1-Tests on the variables and regressions

Phillips-Peron test on the risk premium.

Stocks	Test Value	Critical Value at 1%
CAC	-56.11	-3.96
AXA	-51.22	-3.96
LVMH	-55.7	-3.96
Orange	-54.42	-3.96

OLS estimates and residuals tests

Actions	Beta	T-Stat	R ²	LB	ARCH	JB
AXA	1,5	31.74	0.68	21.07	62.13	41993.2
LVMH	1	36.81	0.62	13.24	38.34	10867.6
Orange	0.73	18.83	0.43	17.7	37.81	4480.43

At 5% risk level, column LB (Ljung-Bpx Test): $\chi^2(5) = 11.1$, column ARCH (ARCH-LM Test): $\chi^2(2) = 5.99$, column JB (Jarque-Bera Test): $\chi^2(2) = 5.99$.

A2-Cross-spectral analysis Coherences and phases

The coherence is on the y-axis and the frequencies on the x-axis (in days).

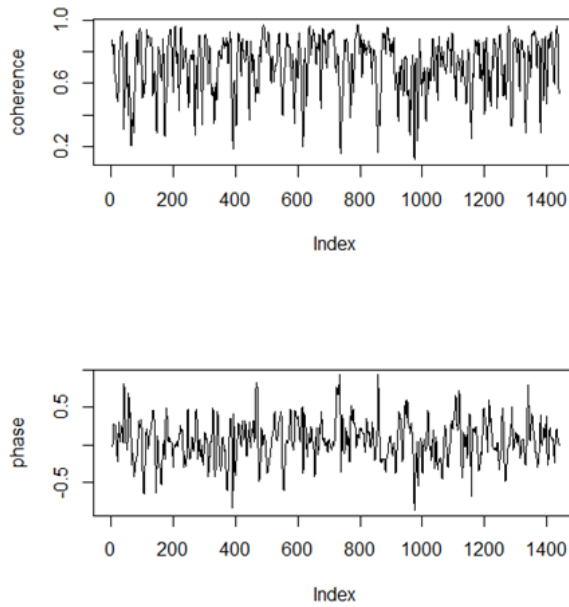


Figure 1.1 *AXA – CAC40*

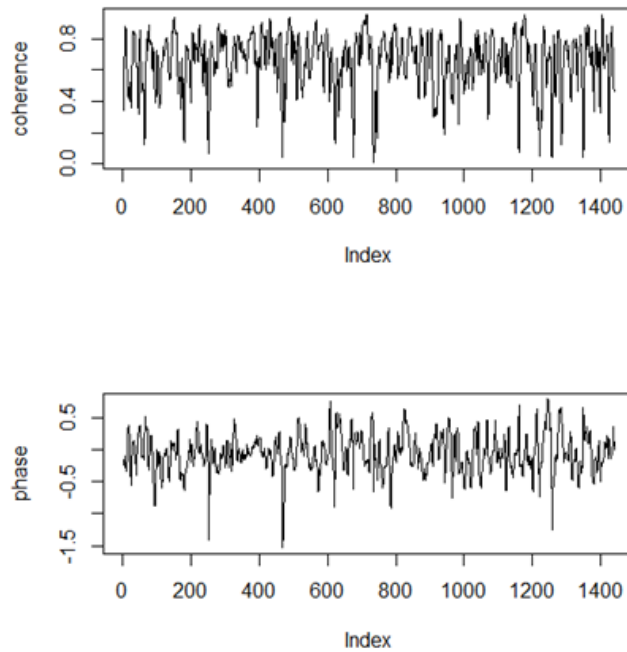


Figure 1.2 *LVMH – CAC40*

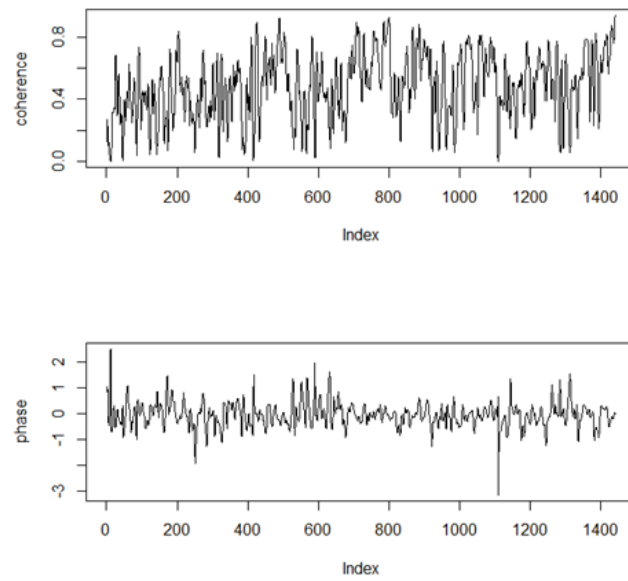


Figure 1.3 *Orange – CAC40*

References

- [1] Auth C., *Continuous Wavelet Transform and Wavelet Coherence-Implementation and Application to the Diversification Analysis of Hedge Funds Returns*, 2013.
- [2] Black F. and Jensen M. and Scholes M., *The Capital Asset Pricing Model: Some Empirical Test ; Studies in the Theory of Capital Markets*, M. Jensen ed., New York: Praeger Publishers, 1972, pp 79-121.
- [3] Daubechies I., *Ten lectures on wavelets*, Conference Series of Applied mathematics in Philadelphia, Society for industrial and applied mathematics, 1992.
- [4] , Grossmann A. and Morlet J., *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM Journal on mathematical analysis, vol 15, issue 4, 1984, pp 723-736.
- [5] Fama E. and French K., *The Cross-Section of Expected Stock Returns* , Journal of Finance, vol 47, issue 2, 1992, pp 427-465.
- [6] Farge M., *Wavelets transforms and their applications to turbulence*, Annual Review of fluid Mechanics, vol 47, issue 2, 1992, pp 427-465.
- [7] Gencay R. and Selcuk F. and Whitcher B., *Multiscale systematic risk*, Journal of internationnal Money and Finance, vol 24, 2005, pp 55-70.
- [8] Lau K.M Weng H., *Climate Signal detection using Wavelet transform: How to make a time series sing.*, Bulletin of the American Meteorological Society, vol 76, issue 12 1995, pp 2391-2402.
- [9] Mallat S., *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 11, issue 7, july 1989.
- [10] Mallat S., *Une exploration des signaux en ondelettes*, Ecole polytechnique, 2009.
- [11] Mallat S., *Wavelet tour of signal processing: the sparse way*, Academic Press, 2009.
- [12] Melard G., *Introduction a l analyse des series temporelle et a la prevision* , Revue MODULAD, 2006.
- [13] Melard G., *Examples of the evolutionary spectrum theory*, Journal of time series analysis, vol 6, issue 2, 1985, pp 81-90.
- [14] Mestre R. and Terraza M., *Time-frequencies analysis of CAPM-Application to the CAC 40*, MIC Conference Managing the Global Economy, Monastier di Treviso, Italy, 2017.
- [15] Meyer Y., Jaffard S., Rioul O., *L analyse par ondelettes*, 1986.
- [16] Sharpe W. , *Capital Asset Prices : a Theory of Market Equilibrium under risk*, Journal of Finance, Canberra. ed. Western Australia : Modelling and Simulation Society of Australia and New Zealand Inc. , vol 19, issue 3, Sept 1964, pp 425-442.
- [17] Torrence C. Compo G.P., *A practical guide to wavelet analysis*, Bulletin of the American Meteorological Society, vol 79, issue 1, 1998, pp 61-78.
- [18] Torrence C. Webster P.J., *Interdecadal Changes in the ENSO-Monsoon System*, Journal of Climate, vol 12, 1999, pp 2679-2690.

Prediction of High-Dimensional Time-Series with Exogenous Variables Using Extended Koopman Operator Framework in Reproducing Kernel Hilbert Space

Jia-Chen Hua^{1*}, Farzad Noorian², Philip H.W. Leong², Gemunu Gunaratne³,
and Jorge Gonçalves¹

¹ Luxembourg Centre for Systems Biomedicine,
University of Luxembourg, 4367 Belvaux, Luxembourg

² School of Electrical and Information Engineering,
University of Sydney, NSW 2006, Australia

³ Department of Physics,
University of Houston, Houston, Texas 77204, USA

Abstract. We propose a novel methodology to predict high-dimensional time-series with exogenous variables using Koopman operator framework, by assuming that the time series are generated by some underlying unknown dynamical system with input as exogenous variables. In order to do that, we first extend and generalize the definition of the original Koopman operator to allow for input to the underlying dynamical system. We then obtain a formulation of the extended Koopman operator in reproducing kernel Hilbert space (RKHS) and a new derivation of its numerical approximation methods. We also obtain a probabilistic interpretation of this numerical method developed for deterministic Koopman operator by using the connection between RKHS and Gaussian processes regression, and relate it to the stochastic Koopman and Perron-Frobenius operator. In applications, we found that the prediction performance of this methodology is very promising in forecasting real world high-dimensional time-series with exogenous variables, including energy consumption data and financial markets data. We believe that this methodology will be of interest to the community of scientists and engineers working on quantitative finance, econometrics, system biology, neurosciences, meteorology, oceanography, system identification and control, data mining, machine learning, computational intelligence, and many other fields involving high-dimensional time series and spatio-temporal data.

Keywords: High-dimensional time-series, Spatio-temporal dynamics, Complex system, Koopman operator, Perron-Frobenius operator, Dynamical system, Reproducing kernel Hilbert space, Gaussian Processes, Machine learning, Data mining, Econophysics, Financial markets modeling, Energy forecasting, Collective behavior

* The corresponding author would like to thank Dr. Alexandre Mauroy for insightful discussions on generalizing Koopman operator to systems with input.

1 Introduction

In many application fields, high-dimensional time-series are generated or sampled from some underlying dynamical system. Even if this is not apparently the case, assuming so could potentially enable methodologies developed in systems and control communities to be utilized for time series analysis and prediction purposes. The Koopman operator [8, 10, 1] is a good example. It enables characterization, reduced order modeling and dimensionality reduction, system identification, prediction, control, etc. of the underlying nonlinear dynamical system using linear theories and techniques, and since it has been developed as a data-driven framework [13, 15], most of its applications up to now are dealing with high-dimensional time series. Hence this framework fits well in the time series prediction context, especially the high dimensional ones [5]. There have been several major numerical methods developed to extract the spectral properties of Koopman operator from time series data, and utilizing these properties for time series prediction has several major advantages, which we will elaborate and re-emphasize in Sec. 2.2.

In this paper, we generalize the Koopman operator framework to system with input as exogenous variables. By using the simplest generalization trick [11], we found that the techniques and methods that we developed for Kernel KMR [5] methodology can be utilized almost directly with minimal modification. Hence we can generalize Kernel KMR to Kernel EKMRX (Kernel-based Extended Koopman Mode Regression with eXogenous variables) to predict high dimensional time series with exogenous variables. In theory part of this paper, we formulate the Koopman operator in reproducing kernel Hilbert space (RKHS), which is the most important function space in modern machine learning, and we obtain a new derivation of the EDMD [15] and its kernel-based extension [16] by exploiting the Dirac bra-ket notation [3]. Moreover, we obtain a probabilistic interpretation of these numerical methods developed for deterministic Koopman operator by exploiting the connection between RKHS and Gaussian processes regression, and relate it to the stochastic Koopman and Perron-Frobenius operator. In application part, we test our new prediction methodology for various types of data from different fields and found promising initial results.

2 Theory

2.1 Reproducing kernel Hilbert space and Gaussian processes regression

In this subsection, we briefly summarize the basic theory of reproducing kernel Hilbert space and its relation to Gaussian processes regression. For a more complete exposition of this topic with technical details, we refer the readers to Refs. [4, 12].

In many applications of statistical learning, a typical task is to approximate an unknown function $f(\mathbf{x})$ given some training data or observations $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_M, y_M)\}$, where $\mathbf{x}_i \in \mathbb{R}^N$ and $y_i \in \mathbb{R}$, such that the learned

function \hat{f} minimizes some regularized empirical risk function and can provide reasonably accurate prediction at a new data point $\hat{f}(\mathbf{x}_*)$. The unknown function f is usually chosen from a “reproducing kernel Hilbert space (RKHS)” \mathcal{H}_k . The representer theorem [14] states that this minimizer \hat{f} of a (special case of) regularized empirical risk function $J[f] = \sum_{i=1}^M (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2$ can be written as $\hat{f}(\cdot) = \sum_{i=1}^M \alpha_i k(\cdot, \mathbf{x}_i)$, where $k(\mathbf{x}, \mathbf{x}') : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ is the “reproducing kernel” function which is symmetric and positive (semi-)definite. Throughout this paper, we will often exploit the Dirac’s bra-ket notation [3] to write functions, functionals, and linear operators in a compact way, *e.g.*, the minimizer can be written as $|\hat{f}\rangle = \sum_{i=1}^M \alpha_i |k_{\mathbf{x}_i}\rangle$ in bra-ket notation.

The RKHS is a Hilbert space of functions equipped with inner product $\langle \cdot | \cdot \rangle_{\mathcal{H}_k}$ satisfying: (1) $\forall \mathbf{x}$ fixed, $k(\mathbf{x}, \mathbf{y}) = k_{\mathbf{x}}(\cdot) \in \mathcal{H}_k$ is a function of \mathbf{y} ; (2) $k(\cdot, \cdot)$ has the “reproducing” property: $\forall f \in \mathcal{H}_k$, $\langle f(\cdot) | k_{\mathbf{x}}(\cdot) \rangle_{\mathcal{H}_k} = f(\mathbf{x})$. It follows from (2) that $\langle k_{\mathbf{y}}(\cdot) | k_{\mathbf{x}}(\cdot) \rangle_{\mathcal{H}_k} = k_{\mathbf{y}}(\mathbf{x}) = k_{\mathbf{x}}(\mathbf{y}) = k(\mathbf{y}, \mathbf{x})$. Each RKHS has a unique k , and according to Moore-Aronszajn theorem, given any symmetric positive definite function $k(\mathbf{y}, \mathbf{x})$, there is a unique RKHS such that $k(\mathbf{y}, \mathbf{x})$ is the reproducing kernel. In fact, this theorem showed that this unique RKHS $\{f \in \mathcal{H}_k | f(\cdot) = \sum_{i=1}^{M \rightarrow \infty} \alpha_i k(\cdot, \mathbf{x}_i)\}$ can be built from defining the inner product $\langle f | g \rangle_{\mathcal{H}_k} = \sum_{j=1}^{M' \rightarrow \infty} \sum_{i=1}^{M \rightarrow \infty} \alpha_i \beta_j k(\mathbf{y}_j, \mathbf{x}_i)$, where $g(\cdot) = \sum_{j=1}^{M' \rightarrow \infty} \beta_j k(\cdot, \mathbf{y}_j)$. It satisfies the reproducing property $\langle f(\cdot) | k_{\mathbf{x}}(\cdot) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{M \rightarrow \infty} \langle \alpha_i k(\cdot, \mathbf{x}_i) | k_{\mathbf{x}}(\cdot) \rangle_{\mathcal{H}_k} = \sum_{i=1}^{M \rightarrow \infty} \alpha_i k(\mathbf{x}, \mathbf{x}_i) = f(\mathbf{x})$. The reproducing kernels can be considered as a basis of this RKHS, and they are also called “point evaluation functional”. As an analog, in L^2 , the Dirac delta is the point evaluation functional $\langle \delta_{\mathbf{x}} | f \rangle_{L^2} = \int f(\mathbf{x}') \delta(\mathbf{x} - \mathbf{x}') d\mathbf{x}' = f(\mathbf{x})$, but since $\delta_{\mathbf{x}}(\cdot) \notin L^2$, L^2 is not a RKHS.

Another representation of RKHS is from Mercer’s theorem, which states that a positive (semi-)definite function can be eigen-decomposed as $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{\infty} \sigma_i q_i(\mathbf{x}) q_i(\mathbf{x}')$, where $\{q_i(\cdot)\}$ are orthonormal in L^2 , and $\{\sigma_i\}_{i=1}^{M \rightarrow \infty}$ is a non-increasing sequence of eigenvalues with $\sigma_M \rightarrow 0$ when $M \rightarrow \infty$. It follows from this theorem that the unique RKHS associated to this $k(\mathbf{x}, \mathbf{x}')$ is $\{f \in L^2 | \sum_{i=1}^{\infty} \frac{\langle q_i | f \rangle_{L^2}^2}{\sigma_i} < \infty\}$, and the inner product is given by $\langle f | g \rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \langle f | q_i \rangle_{L^2} \frac{1}{\sigma_i} \langle q_i | g \rangle_{L^2}$. One consequence of this inner product is that the induced norm is $\|f\|_{\mathcal{H}_k}^2 = \langle f | f \rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{\langle q_i | f \rangle_{L^2}^2}{\sigma_i}$, and in order to be bounded, the components $f_i = \langle q_i | f \rangle_{L^2}$ must decay quickly when i increases, which effectively imposes a smoothness requirement on L^2 in order for it to become a RKHS. Another consequence of this inner product is that one can define $\{p_i(\cdot) = \sqrt{\sigma_i} q_i(\cdot)\}$ such that it is an orthonormal basis of this unique RKHS, and as an analogue to the Dirac delta which can be represented by $\delta_{\mathbf{x}}(\cdot) = \sum_{i=1}^{\infty} q_i(\mathbf{x}) q_i(\cdot)$, the reproducing kernel functions can be written as $k_{\mathbf{x}}(\cdot) = \sum_{i=1}^{\infty} p_i(\mathbf{x}) p_i(\cdot)$. One can easily check that the reproducing property holds with respect to this inner product.

Back to the regularized optimization problem $J[f] = \frac{1}{2\lambda^2_M} \sum_{i=1}^M (y_i - f(\mathbf{x}_i))^2 + \frac{1}{2} \|f\|_{\mathcal{H}_k}^2$, the representer theorem asserts that the minimizer $\hat{f}(\cdot) = \sum_{i=1}^M \alpha_i k(\cdot, \mathbf{x}_i)$, such that one can effectively minimize $J[\alpha_i]$ by setting the derivatives with respect to α_i equal to zeros, and then the α_i ’s can be solved as a column

vector $\alpha = (\mathbf{G} + \lambda_M^2 \mathbf{I})^{-1} \mathbf{y}$, where $\mathbf{y} = [y_1, \dots, y_M]^T$ are the training outputs, \mathbf{I} is the identity matrix, and \mathbf{G} is the kernel Gramian matrix where $\mathbf{G}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Given a new test data \mathbf{x}_* , the function output or prediction is $\hat{f}(\mathbf{x}_*) = \mathbf{k}(\mathbf{x}_*)^T (\mathbf{G} + \lambda_M^2 \mathbf{I})^{-1} \mathbf{y}$, where $\mathbf{k}(\mathbf{x}_*)^T = [k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_M)]$. This is the same as the posterior mean of Gaussian processes regression with i.i.d. noise variance λ_M^2 .

A more heuristic view of predicting the function output given a new test data is from the point evaluation at this new data. As an analogue to the point evaluation in L^2 using Dirac delta $f(\mathbf{x}_*) = \langle \delta_{\mathbf{x}_*} | f \rangle_{L^2} = \int f(\mathbf{x}) \delta(\mathbf{x} - \mathbf{x}_*) d\mathbf{x}$ (which is computationally infeasible using training data), one can work in the RKHS using the reproducing kernel function $k(\mathbf{x}, \mathbf{x}')$ as: $f(\mathbf{x}_*) = \langle k_{\mathbf{x}_*} | f \rangle_{\mathcal{H}_k} = \sum_{i=1}^M \langle k_{\mathbf{x}_*} | q_i \rangle_{L^2} \frac{1}{\sigma_i} \langle q_i | f \rangle_{L^2}$, where the inner products in L^2 can be approximated by summation using training data as $\langle g | f \rangle_{L^2} = \int g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \approx \sum_{i=1}^M g(\mathbf{x}_i) f(\mathbf{x}_i) = \sum_{i=1}^M \langle g | k_{\mathbf{x}_i} \rangle_{\mathcal{H}_k} \langle k_{\mathbf{x}_i} | f \rangle_{\mathcal{H}_k}$. Hence one can obtain $\langle k_{\mathbf{x}_*} | f \rangle_{\mathcal{H}_k} = \sum_{i=1}^M \langle k_{\mathbf{x}_*} | q_i \rangle_{L^2} \frac{1}{\sigma_i} \langle q_i | f \rangle_{L^2} \approx \sum_{i,j,l} \langle k_{\mathbf{x}_*} | k_{\mathbf{x}_j} \rangle_{\mathcal{H}_k} \langle k_{\mathbf{x}_j} | q_i \rangle_{\mathcal{H}_k} \frac{1}{\sigma_i} \langle q_i | k_{\mathbf{x}_l} \rangle_{\mathcal{H}_k} \langle k_{\mathbf{x}_l} | g \rangle_{\mathcal{H}_k}$. Notice that the kernel Gramian matrix has eigen-decomposition $\mathbf{G} = \mathbf{Q} \mathbf{\Sigma}^2 \mathbf{Q}^T$, where $\mathbf{Q}_{ij} = q_j(\mathbf{x}_i) = \langle k_{\mathbf{x}_i} | q_j \rangle_{\mathcal{H}_k}$ and $\mathbf{\Sigma}$ is diagonal with $\Sigma_{ii} = \sqrt{\sigma_i}$. Hence $\mathbf{G}^{-1} = \mathbf{Q} \mathbf{\Sigma}^{-2} \mathbf{Q}^T$ and $(\mathbf{G}^{-1})_{ij} = \sum_{l=1}^M \langle k_{\mathbf{x}_i} | q_l \rangle_{\mathcal{H}_k} \frac{1}{\sigma_l} \langle q_l | k_{\mathbf{x}_j} \rangle_{\mathcal{H}_k}$. Finally one arrives at

$$f(\mathbf{x}_*) = \langle k_{\mathbf{x}_*} | f \rangle_{\mathcal{H}_k} = \mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1} [f(\mathbf{x}_1), \dots, f(\mathbf{x}_M)]^T, \quad (1)$$

which is the same as the posterior mean in noiseless Gaussian processes regression. Replacing \mathbf{G}^{-1} by the Moore–Penrose pseudoinverse \mathbf{G}^+ will be equivalent to regularization, or additive noise in Gaussian processes regression. A typical way of regularization using \mathbf{G}^+ is to truncate out some small eigenvalues σ_i 's and the corresponding eigenvectors $q_i(\mathbf{x})$'s, although a more sophisticated way to perform this truncation is using a smooth cutoff, as developed in Ref. [5]. A useful result following the above derivation is that the inner product in RKHS can be approximated using training data as $\langle g | f \rangle_{\mathcal{H}_k} \approx \sum_{i,j} \langle g | k_{\mathbf{x}_i} \rangle_{\mathcal{H}_k} \mathbf{G}^{-1} \langle k_{\mathbf{x}_j} | f \rangle_{\mathcal{H}_k}$, which means that the “resolution of the identity” or the projection operator into this RKHS can be approximated by training data as $\mathbb{1}_{\mathcal{H}_k} = \sum_{i=1}^M |p_i\rangle \langle p_i| \approx \sum_{i,j} |k_{\mathbf{x}_i}\rangle_{\mathcal{H}_k} \mathbf{G}^{-1} \langle k_{\mathbf{x}_j}|$.

In summary, deterministic approximation of a function in RKHS, or point evaluation of a function on new data can have a probabilistic interpretation via Gaussian processes regression. Moreover, since $\mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1}$ is a row vector of weights on the training outputs $[f(\mathbf{x}_1), \dots, f(\mathbf{x}_M)]^T$, and if it sums up to 1 and if the amount of training data is sufficiently large, it may be considered as a density estimation for the posterior distribution of Gaussian processes, which will induce a density on the training data $[\mathbf{x}_1, \dots, \mathbf{x}_M]^T$. A special case is the point evaluation on training data $f(\mathbf{x}_i) = \mathbf{k}(\mathbf{x}_i)^T \mathbf{G}^{-1} [f(\mathbf{x}_1), \dots, f(\mathbf{x}_M)]^T$, where $\mathbf{k}(\mathbf{x}_i)^T \mathbf{G}^{-1}$ will become a row vector with every element equal to zero except for the i -th one equal to 1, which is a probability mass function concentrated on \mathbf{x}_i that approximates the Dirac delta distribution $\delta_{\mathbf{x}_i}(\cdot)$. Again, replacing \mathbf{G}^{-1} by the Moore–Penrose pseudoinverse \mathbf{G}^+ effectively corresponds

to Gaussian processes with additive noise such that the Dirac delta will become a narrow Gaussian centered at the training data.

2.2 Koopman operator of dynamical system and its generalization to systems with input

Consider a high dimensional time series $\{\mathbf{x}_n\}$ sampled from an underlying dynamical system $(\mathcal{M}, n, \mathbf{F})$, where $n \in \mathbb{Z}$ is discrete time, $\mathcal{M} \subset \mathbb{R}^N$ is the N -dimensional state space containing the $\{\mathbf{x}_n\}$, and $\mathbf{x}_i \mapsto \mathbf{F}(\mathbf{x}_i) = \mathbf{x}_{i+1}$ defines the evolution law. For continuous-time dynamical system $(\mathcal{M}, t, \mathbf{F}^t)$, the flow \mathbf{F}^t evolves the system state as $\mathbf{x}_0 \mapsto \mathbf{F}^t(\mathbf{x}_0) = \mathbf{x}_t$. Since time series data are often sampled with a fixed time gap τ , the adjacent two snapshots of the system are related by $\mathbf{F}^\tau(\mathbf{x}_t) = \mathbf{x}_{t+\tau}$. When the context is clear, we will drop the τ in \mathbf{F}^τ to denote either the discrete time map or continuous time flow of a fixed time gap τ . Here we restrict to stationary time series, or at least locally stationary time series, which can be considered as being sampled from autonomous dynamical systems. We will generalize the Koopman operator to systems with input later. The (deterministic) Koopman operator $\mathcal{K} : \mathcal{F} \rightarrow \mathcal{F}$, where \mathcal{F} consists of scalar observables or functions of state space $\phi : \mathcal{M} \rightarrow \mathbb{C}$, is defined as

$$(\mathcal{K}\phi)(\mathbf{x}) = (\phi \circ \mathbf{F})(\mathbf{x}) = \phi(\mathbf{F}(\mathbf{x})), \quad (2)$$

where \circ denotes the composition of ϕ with \mathbf{F} . Since $\mathcal{K}\phi$ is another element in \mathcal{F} , the Koopman operator defines a new dynamical system $(\mathcal{F}, n, \mathcal{K})$ where \mathcal{K} evolves observables $\phi \in \mathcal{F}$ to a new function $\mathcal{K}\phi$ that gives the value of ϕ at “one step in the future”. Unlike \mathbf{F} which is finite dimensional, \mathcal{K} is infinite dimensional because it acts on function space \mathcal{F} . However, it is also linear even when \mathbf{F} is nonlinear, and hence one can investigate its spectral properties, *i.e.*, eigenvalues and eigenfunctions, which we refer to as Koopman eigenvalues $\{\mu_k\}$ and eigenfunctions $\{\varphi_k\}$.

The dynamical systems $(\mathcal{M}, n, \mathbf{F})$ and $(\mathcal{F}, n, \mathcal{K})$ are two different representations of the same evolution. The link between them is the “full state observable” $\mathbf{g}(\mathbf{x}) = \mathbf{x}$, where $\mathbf{x} \mapsto \mathbf{F}(\mathbf{x})$, and $g_i \mapsto (\mathcal{K}g_i) = g_i \circ \mathbf{F}$ where $g_i \in \mathcal{F}$ is the i -th component of the *vector-valued observable* $\mathbf{g} : \mathcal{M} \rightarrow \mathbb{R}^N$. Assuming g_i is in the span of a set of K Koopman eigenfunctions $\{\varphi_k\}_{k=1}^K$, where K could (and often will) be infinite, then it can be projected as $g_i = \sum_{k=1}^K \xi_{ik} \varphi_k$ with $\xi_{ik} \in \mathbb{C}$. Hence \mathbf{g} can be obtained by “stacking” these weights into vectors (*i.e.*, $\boldsymbol{\xi}_j = [\xi_{1j}, \xi_{2j}, \dots, \xi_{Nj}]^T$). As a result,

$$\mathbf{x} = \mathbf{g}(\mathbf{x}) = \sum_{k=1}^K \boldsymbol{\xi}_k \varphi_k(\mathbf{x}), \quad (3)$$

where $\boldsymbol{\xi}_k$ is the k -th *Koopman mode* corresponding to the eigenfunction φ_k . To make prediction or arrive at the system state of “one step in the future”, one can either evolve \mathbf{x} through \mathbf{F} directly, or evolve the full state observable $\mathbf{g}(\mathbf{x})$

through the Koopman operator \mathcal{K} as $\mathbf{g}(\mathbf{F}(\mathbf{x})) = (\mathcal{K}\mathbf{g})(\mathbf{x}) = \sum_{k=1}^K \xi_k(\mathcal{K}\varphi_k)(\mathbf{x}) = \sum_{k=1}^K \mu_k \xi_k \varphi_k(\mathbf{x})$. Similarly, for continuous time case, we have

$$\mathbf{x}_{t+\tau} = \mathbf{F}^\tau(\mathbf{x}_t) = \mathbf{g}(\mathbf{F}^\tau(\mathbf{x}_t)) = (\mathcal{K}_\tau \mathbf{g})(\mathbf{x}_t) = \sum_{k=1}^K e^{\lambda_k \tau} \xi_k \varphi_k(\mathbf{x}_t), \quad (4)$$

where λ_k and φ_k are the k -th eigenvalue and eigenfunction of the infinitesimal generator $\hat{\mathcal{K}} \triangleq \frac{d}{dt}$ of the semi-group of Koopman operators $\{\mathcal{K}_t\}_{t \in \mathbb{R}^+}$, and $\mu_k = e^{\lambda_k \tau}$ is the k -th eigenvalue of finite-time Koopman operator $\mathcal{K}_\tau = e^{\tau \hat{\mathcal{K}}}$.

There are many available methods and techniques to approximate each component F_i of the unknown \mathbf{F} one-by-one by using training data in order to make predictions (*e.g.*, Gaussian processes regression). However, the Koopman operator framework is advantageous for these reasons: (a) the dynamics associated with each eigenfunction is determined by its corresponding eigenvalue, such that one can predict the system state at any time later (rather than a fixed time length only), by setting an *arbitrary* real number τ in Eq. (4), (b) \mathbf{F} is usually highly nonlinear and/or stochastic, whereas \mathcal{K} is linear, so it is easier to investigate and much more convenient to generate predictive models and utilize for other applications such as system identification and control, (c) because of the linearity, the high-dimensional time series generated by the system dynamics can be decomposed linearly using spectral properties of \mathcal{K} as Eq. (4), where by truncating out some noisy, irregular, or non-important terms in the summation, one can accomplish *both* dimensionality reduction and time series prediction simultaneously, (d) the state variables \mathbf{x}_t and many designed or learned features are extrinsic to the underlying dynamical system, which means that models and predictions could be dependent on specific extrinsic variables chosen or features designed to sample and describe the system dynamics, whereas the eigenfunctions $\{\varphi_k\}$ of \mathcal{K} are *intrinsic dynamic variables* [17] of the underlying system which are independent from particular experimental apparatus such as sensors or specific observations of the high dimensional time series, so they are able to extract the intrinsic features of the system dynamics that generates the time series and are more fundamental and physically meaningful, (e) Koopman modes $\{\xi_k\}$ and eigenfunctions $\{\varphi_k\}$ characterize the underlying *system* dynamics *collectively* in continuous time instead of a number of functions *individually* with each of them predicting a *single* variable at a fixed time length later, and hence they enable us to avoid over-fitting not only by regularization and cross-validation on parameters and/or model complexity in usual ways of statistics, but also by “physical” cross-validation on intrinsic dynamic features *at the system level* [5], and by identifying irregular and non-repeatable/non-predictable features and dropping them out in Eq. (4), one can achieve more reliable predictions.

In order to compute $\{(\mu_k, \varphi_k, \xi_k)\}_{k=1}^K$ of Koopman eigenvalues, eigenfunctions, and modes from data, one has to find a matrix representation of \mathcal{K} by projecting it into some subspace of \mathcal{F} spanned by a basis $\{\psi_k(\mathbf{x})\}_{k=1}^K$. For computational feasibility and convenience, we usually require $\psi_k(\cdot) \in L^2(\mathcal{M})$, such that we can compute inner products using training data $\{(\mathbf{x}_1, \mathbf{y}_1), \dots,$

$(\mathbf{x}_M, \mathbf{y}_M)\}$ where $\mathbf{y}_i = \mathbf{F}(\mathbf{x}_i)$, in order to require $\{\psi_k(\mathbf{x})\}_{k=1}^K$ to be orthonormal by computing the Moore–Penrose pseudoinverse of the data matrix Ψ_x^+ , where $[\Psi_x]_{ij} = \psi_j(\mathbf{x}_i)$. Utilizing Dirac bra–ket notation, we denote the i -th row of Ψ_x^+ as $\langle \psi_i |$ such that $\langle \psi_i | \psi_j \rangle_{L^2} = \delta_{ij}$, where δ_{ij} is the Kronecker delta. Hence in this “feature space” $\mathcal{F}_K \triangleq \text{span}\{\psi_k(\cdot)\}_{k=1}^K$, the identity operator can be written as $\mathbb{1}_{\mathcal{F}_K} = \sum_{k=1}^K |\psi_k\rangle\langle\psi_k|$, and \mathcal{K} projected to \mathcal{F}_K can be written as $\mathcal{K} = \mathcal{K}\mathbb{1}_{\mathcal{F}_K} = \mathcal{K} \sum_{k=1}^K |\psi_k\rangle\langle\psi_k| = \sum_{k=1}^K |\psi_k \circ \mathbf{F}\rangle\langle\psi_k|$. Therefore, the elements of matrix representation \mathbf{K} of \mathcal{K} is $\mathbf{K}_{ij} = \langle \psi_i | \mathcal{K} | \psi_j \rangle_{L^2} = \langle \psi_i | \psi_j \circ \mathbf{F} \rangle_{L^2}$, and $\mathbf{K} = \Psi_x^+ \Psi_y$, where the j -th column of Ψ_y is $|\psi_j \circ \mathbf{F}\rangle$ and $[\Psi_y]_{ij} = \psi_j \circ \mathbf{F}(\mathbf{x}_i) = \psi_j(\mathbf{y}_i)$. Eigenvalue problem $\mathcal{K}|\varphi_k\rangle = \mu_k|\varphi_k\rangle$ becomes eigenvalue equation of \mathbf{K} as $\mathbf{K}\mathbf{v}_k = \mu_k\mathbf{v}_k$, where the i -th component of \mathbf{v}_k is $(\mathbf{v}_k)_i = \langle \psi_i | \varphi_k \rangle_{L^2}$, so the eigenfunction $|\varphi_k\rangle = \sum_{i=1}^K |\psi_i\rangle(\mathbf{v}_k)_i$, or $\Phi_x = \Psi_x \mathbf{V}$ in matrix notation, where $[\Phi_x]_{ij} = \varphi_j(\mathbf{x}_i)$ and columns of \mathbf{V} are $\{\mathbf{v}_k\}$. The *continuous-time eigenvalue* can be computed as $\lambda_k \triangleq \log(\mu_k)/\tau$, and according to Eq. (3), Koopman modes $\{\xi_k\}$ can be computed by projecting $\mathbf{g}(\mathbf{x}) = \mathbf{x}$ onto $\{\varphi_k(\mathbf{x})\}$ as $\Xi = \Phi_x^+ \mathbf{X}$, where the i -th rows of Ξ and \mathbf{X} are ξ_i^T and \mathbf{x}_i^T , respectively. This procedure is called extended Dynamic Mode Decomposition (EDMD)[15] and it has become one of the most widely adopted numerical methods for data-driven Koopman spectral analysis, even outside the fluid dynamics community where the Koopman operator’s spectral properties was thoroughly investigated for the first time [10].

Koopman operator can be also defined as an integral operator [9, 2, 6, 7], which enables a better and uniform formulation of both deterministic and stochastic Koopman operator, and its Hermitian adjoint, namely the Perron-Frobenius operator $\mathcal{L} = \mathcal{K}^\dagger$, where the \dagger denotes Hermitian adjoint. Again, consider the dynamical system $(\mathcal{M}, t, \mathbf{F}^t)$, when \mathbf{F}^t is highly nonlinear and/or stochastic, starting from an initial point on \mathcal{M} and keeping track of its single trajectory along the time evolution will become meaningless, as any finite initial difference will blow up exponentially. Instead, a better strategy is to investigate the statistical behavior of a swarm of points’ time evolution, which leads to the investigation of (probability) measure/density on \mathcal{M} and its time evolution induced by \mathbf{F}^t . Consider a probability density function ρ defined on \mathcal{M} , and for computational convenience, we require $\rho \in \mathcal{F} \subseteq L^2(\mathcal{M})$. When \mathbf{F} evolves an arbitrary swarm of points of system states on \mathcal{M} , *i.e.*, evolves the pre-image $\mathbf{F}^{-1}(\mathbb{A})$ of any measurable domain $\mathbb{A} \subseteq \mathcal{M}$ to \mathbb{A} at time τ later, the density ρ on $\mathbf{F}^{-1}(\mathbb{A})$ will be evolved by a linear operator to a new density on \mathbb{A} as

$$\int_{\mathbb{A}} (\mathcal{L}_\tau \rho)(\mathbf{y}) d\mathbf{y} = \int_{\mathbf{F}^{-1}(\mathbb{A})} \rho(\mathbf{x}) d\mathbf{x}, \quad (5)$$

such that the probability measure is conserved, where the \mathcal{L}_τ is the Perron-Frobenius operator that evolves probability densities. If \mathbf{F} is stochastic, which means that $\mathbf{F}(\mathbf{x})$ follows a *transition probability density* $p_\tau(\mathbf{y}|\mathbf{x})$, the Perron-Frobenius operator can be also defined as

$$(\mathcal{L}_\tau \rho)(\mathbf{y}) = \int_{\mathbf{F}^{-1}(\mathbb{A})} \rho(\mathbf{x}) p_\tau(\mathbf{y}|\mathbf{x}) d\mathbf{x}. \quad (6)$$

A special case is the deterministic system, where $p_\tau(\mathbf{y}|\mathbf{x})$ will become a Dirac delta distribution $\delta_{\mathbf{F}(\mathbf{x})}(\mathbf{y}) = \delta(\mathbf{y} - \mathbf{F}(\mathbf{x}))$, such that the center of an initial Dirac delta distribution $\delta_{\mathbf{x}}$ will be moved in consistence with the dynamics as $\mathcal{L}_\tau \delta_{\mathbf{x}}(\mathbf{y}) = \int_{\mathbf{F}^{-1}(\mathbb{A})} \delta(\mathbf{x} - \mathbf{x}') \delta(\mathbf{y} - \mathbf{F}(\mathbf{x}')) d\mathbf{x}' = \delta_{\mathbf{F}(\mathbf{x})}(\mathbf{y})$. Analogous to this, notice that Koopman operator for deterministic system is defined as $(\mathcal{K}_\tau \phi)(\mathbf{x}) = (\phi \circ \mathbf{F})(\mathbf{x}) = \phi(\mathbf{F}(\mathbf{x}))$, it can be also written as

$$(\mathcal{K}_\tau \phi)(\mathbf{x}) = \int_{\mathbb{A}} \phi(\mathbf{y}) \delta(\mathbf{y} - \mathbf{F}(\mathbf{x})) d\mathbf{y}, \quad (7)$$

and following this idea, the Koopman operator for stochastic system should be defined as

$$(\mathcal{K}_\tau \phi)(\mathbf{x}) = \int_{\mathbb{A}} \phi(\mathbf{y}) p_\tau(\mathbf{y}|\mathbf{x}) d\mathbf{y} = \mathbb{E}[\phi(\mathbf{F}(\mathbf{x}))|\mathbf{x}], \quad (8)$$

which is the conditional expectation of observable ϕ 's value at time τ later. Using these definitions, one can check that the Koopman operator and Perron-Frobenius operator are adjoint to each other for both deterministic and stochastic systems, by considering how the expectation value of an observable over some region evolves in time:

$$\begin{aligned} \mathbb{E}[\phi(\mathbf{y})] &= \int_{\mathbb{A}} (\mathcal{L}_\tau \rho)(\mathbf{y}) \phi(\mathbf{y}) d\mathbf{y} = \langle \mathcal{L}_\tau \rho | \phi \rangle_{L^2} = \int_{\mathbb{A}} \int_{\mathbf{F}^{-1}(\mathbb{A})} \rho(\mathbf{x}) p_\tau(\mathbf{y}|\mathbf{x}) d\mathbf{x} \phi(\mathbf{y}) d\mathbf{y} \\ &= \int_{\mathbf{F}^{-1}(\mathbb{A})} \mathbb{E}[\phi(\mathbf{F}(\mathbf{x}))|\mathbf{x}] \rho(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{F}^{-1}(\mathbb{A})} (\mathcal{K}_\tau \phi)(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x} = \langle \rho | \mathcal{K}_\tau \phi \rangle_{L^2}, \end{aligned} \quad (9)$$

where \mathcal{K}_τ acting to the left on $\langle \rho |$ is $\langle \rho | \mathcal{K}_\tau | \phi \rangle_{L^2} = \langle \mathcal{K}_\tau^\dagger \rho | \phi \rangle_{L^2} = \langle \mathcal{L}_\tau \rho | \phi \rangle_{L^2}$. This formulation enables us to predict the expectation of a function's value at a later time when tracking and predicting a single trajectory is not meaningful due to high nonlinearity and/or stochasticity of \mathbf{F} , and we will relate this formulation to Koopman and Perron-Frobenius operators framework in reproducing kernel Hilbert space in the next subsection.

Finally, there are several ways to generalize Koopman operator to systems with input [11]. One of the simplest ways is to augment the system state \mathbf{x}_t with the current input $\mathbf{u}_t \in \mathbb{R}^{N'}$, such that the dimension of the extended system state $\tilde{\mathbf{x}}$ will be $N + N'$. The time evolution of the system will be extended as $\tilde{\mathbf{x}}_{t+\tau} = \tilde{\mathbf{F}}^\tau(\tilde{\mathbf{x}}_t) = \tilde{\mathbf{F}}^\tau(\mathbf{x}_t, \mathbf{u}_t)$, where the first N components of $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{x}}$ are $\mathbf{x}_{t+\tau} = \mathbf{F}^\tau(\mathbf{x}_t, \mathbf{u}_t)$, and we assume that there is a purely formal map or flow that “shifts” the input as $\mathbf{u}_{t+\tau} = \mathbf{S}^\tau(\mathbf{x}_t, \mathbf{u}_t)$, since there is not necessarily any “dynamics” of the input. The generalized Koopman operator can be defined on this extended system as before $\mathcal{K}\phi(\tilde{\mathbf{x}}_t) = \phi \circ \tilde{\mathbf{F}}^\tau(\tilde{\mathbf{x}}_t)$. For prediction purposes, we are only interested in the original system state \mathbf{x} , so there is no need to project N' dimensional full state observable of input $\mathbf{g}_u(\mathbf{x}_t, \mathbf{u}_t) = \mathbf{u}_t$ on Koopman eigenfunctions in order to compute the corresponding Koopman modes for input. Except for this trivial difference, all the available numerical procedures for Koopman spectral analysis and prediction can be applied with very little modification. Notice that this augmentation trick can be also applied to previous state

and input, such that one can investigate a system with finite amount of memory in the same way as investigating a system without memory. For simplicity, we only consider memoryless system in this paper, and this topic will be left for future investigation.

2.3 Koopman operator in reproducing kernel Hilbert space

Recall from Eq. (1) that point evaluation in RKHS is the same as computing some expectation value such as the posterior mean of Gaussian processes, for example, $\langle k_{\mathbf{x}_i} | f \rangle_{\mathcal{H}_k} = \mathbf{k}(\mathbf{x}_i)^T \mathbf{G}^{-1} [f(\mathbf{x}_1), \dots, f(\mathbf{x}_M)]^T$, where $\mathbf{k}(\mathbf{x}_i)^T \mathbf{G}^{-1}$ is a row vector with all zero elements except for the i -th equal to 1, which may be considered as discrete approximation to Dirac delta distribution $\delta_{\mathbf{x}_i}$. Replacing \mathbf{G}^{-1} by pseudo-inverse \mathbf{G}^+ will be equivalent to regularization or adding noise to the Gaussian processes, such that $\mathbf{k}(\mathbf{x}_i)^T \mathbf{G}^+$ can approximate some narrow Gaussian centered at \mathbf{x}_i . Similarly, consider the projection of Koopman operator in RKHS by point evaluation at a new state \mathbf{x}_* of a function $|h\rangle$ evolved by \mathcal{K} as $\langle k_{\mathbf{x}_*} | \mathcal{K} | h \rangle_{\mathcal{H}_k} = \mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1} [\mathcal{K}h(\mathbf{x}_1), \dots, \mathcal{K}h(\mathbf{x}_M)]^T$, where the $\mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1}$ is expected to approximate the initial density $\rho(\mathbf{x})$ before time evolution in Eq. (6), in the limit of infinite amount of training data, *i.e.*, $M \rightarrow \infty$.

Now, recall that the identity operator in RKHS $\mathbb{1}_{\mathcal{H}_k} = \sum_i |p_i\rangle \langle p_i| \approx \sum_{ij} |k_{\mathbf{x}_i}\rangle_{\mathcal{H}_k} \mathbf{G}^{-1}_{\mathcal{H}_k} \langle k_{\mathbf{x}_j}|$, and inner product can also be approximated as $\langle g | \mathbb{1}_{\mathcal{H}_k} | f \rangle_{\mathcal{H}_k} \approx \sum_{ij} \langle g | k_{\mathbf{x}_i} \rangle_{\mathcal{H}_k} \mathbf{Q} \mathbf{\Sigma}^{-2} \mathbf{Q}_y^T \langle k_{\mathbf{x}_j} | f \rangle_{\mathcal{H}_k}$, where $[\mathbf{Q}_y^T]_{ij} = \langle q_i | k_{\mathbf{y}_j} \rangle_{\mathcal{H}_k} = \sum_l \langle q_i | q_l \rangle_{L^2} \frac{1}{\sigma_l} \langle q_l | k_{\mathbf{y}_j} \rangle_{L^2} \approx \sum_l \frac{1}{\sigma_l} \langle q_i | k_{\mathbf{x}_l} \rangle_{\mathcal{H}_k} \langle k_{\mathbf{x}_l} | k_{\mathbf{y}_j} \rangle_{\mathcal{H}_k} = [\mathbf{\Sigma}^{-2} \mathbf{Q}^T \mathbf{K}^T]_{ij}$, and $\mathbf{K}_{ij} = \langle k_{\mathbf{x}_i} | \mathcal{K} | k_{\mathbf{x}_j} \rangle_{\mathcal{H}_k} = k_{\mathbf{x}_j}(\mathbf{F}(\mathbf{x}_i)) = k(\mathbf{y}_i, \mathbf{x}_j) = \langle k_{\mathbf{y}_i} | k_{\mathbf{x}_j} \rangle_{\mathcal{H}_k}$. It follows that $\mathbf{Q} \mathbf{\Sigma}^{-2} \mathbf{Q}_y^T = \mathbf{Q} \mathbf{\Sigma}^{-2} \mathbf{Q}^T \mathbf{Q} \mathbf{\Sigma}^{-2} \mathbf{Q}^T \mathbf{K}^T = \mathbf{G}^{-2} \mathbf{K}^T$, and hence $\mathbb{1}_{\mathcal{H}_k}$ can also be approximated by $\sum_{ij} |k_{\mathbf{x}_i}\rangle_{\mathcal{H}_k} \mathbf{G}^{-2} \mathbf{K}^T_{\mathcal{H}_k} \langle k_{\mathbf{y}_j}|$. After plugging the appropriate approximations of $\mathbb{1}_{\mathcal{H}_k}$ into $\langle k_{\mathbf{x}_*} | \mathbb{1}_{\mathcal{H}_k} \mathcal{K} | h \rangle_{\mathcal{H}_k}$, one can obtain

$$\langle k_{\mathbf{x}_*} | \mathcal{K} | h \rangle_{\mathcal{H}_k} \approx \mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1} \mathbf{K} \mathbf{G}^{-2} \mathbf{K}^T [h(\mathbf{y}_1), \dots, h(\mathbf{y}_M)]^T. \quad (10)$$

When the number of training snapshots pairs $M \rightarrow \infty$, we would expect that $\mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1}$ approximates $\rho(\mathbf{x})$, and $\mathbf{K} \mathbf{G}^{-2} \mathbf{K}^T$ approximates the transition density $p_\tau(\mathbf{y} | \mathbf{x})$, such that the multiplication between \mathbf{G}^{-1} and \mathbf{K} in Eq. (10) approximates the integral over \mathbf{x} in Eq. (6), and the multiplication between \mathbf{K}^T and $[h(\mathbf{y}_1), \dots, h(\mathbf{y}_M)]^T$ in Eq. (10) approximates the integral over \mathbf{y} in Eq. (8). Finally, we can consider Eq. (10) as an appropriate discrete approximation of Eq. (9) using training data, and the point evaluation of a function h evolved by Koopman operator in RKHS at a new data point $\langle k_{\mathbf{x}_*} | \mathcal{K} | h \rangle_{\mathcal{H}_k}$ is equivalent to predicting its expectation value $\mathbb{E}[h(\mathbf{y})]$ over training data at a later time, which avoids tracking and predicting a single trajectory unreliably on \mathcal{M} governed by a highly nonlinear and/or stochastic \mathbf{F} . Notice that during the derivation of Eq. (10), *we did not use the definition of stochastic Koopman operator*, but we can in fact approximate $(\mathcal{K}_\tau \phi)(\mathbf{x}) = \int_{\mathbb{A}} \phi(\mathbf{y}) p_\tau(\mathbf{y} | \mathbf{x}) d\mathbf{y} = \mathbb{E}[\phi(\mathbf{F}(\mathbf{x})) | \mathbf{x}]$ on training data by the rows of $\mathbf{K} \mathbf{G}^{-2} \mathbf{K}^T [h(\mathbf{y}_1), \dots, h(\mathbf{y}_M)]^T$, and approximate $(\mathcal{L}_\tau \rho)(\mathbf{y}) = \int_{\mathbf{F}^{-1}(\mathbb{A})} \rho(\mathbf{x}) p_\tau(\mathbf{y} | \mathbf{x}) d\mathbf{x}$ on training data by the columns of $\mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^{-1} \mathbf{K} \mathbf{G}^{-2} \mathbf{K}^T$. These nice relations are induced by the connection be-

tween deterministic approximation of a function in RKHS and Gaussian processes regression, and replacing \mathbf{G}^{-1} by \mathbf{G}^+ will turn these almost singular densities to narrow Gaussians, which have even better probabilistic interpretation and correspond to regularized optimization in RKHS and noisy Gaussian processes regression that usually have better prediction accuracy.

In order to predict the future state of the system in RKHS using the spectral properties of Koopman operator, we first need to obtain a matrix representation of \mathcal{K} projected in this space. Following the derivation of EDMD procedure in previous section, one can write $\mathbb{1}_{\mathcal{H}_k} = \sum_i |p_i\rangle\langle p_i| = \sum_i |q_i\rangle_{L^2} \frac{1}{\sigma_i} L^2 \langle q_i| = \sum_{ij} |q_j\rangle_{L^2} \frac{1}{\sigma_j} \langle q_j|p_i\rangle_{L^2} \langle p_i| \approx \sum_{il} |k_{\mathbf{x}_l}\rangle_{\mathcal{H}_k} \langle k_{\mathbf{x}_l}|q_i\rangle_{\mathcal{H}_k} \frac{1}{\sqrt{\sigma_i}} \langle p_i| = \sum_{il} |k_{\mathbf{x}_l}\rangle_{\mathcal{H}_k} [\mathbf{Q}\Sigma^+]_{li} \langle p_i| = \sum_{il} |p_i\rangle_{\mathcal{H}_k} [\Sigma^+ \mathbf{Q}^T]_{il} \langle k_{\mathbf{x}_l}|$, where $|p_i\rangle = \sqrt{\sigma_i} |q_i\rangle$ (in some literature they are called canonical features or Mercer's features due to Mercer's theorem). Then \mathcal{K} can be written as $\mathcal{K}\mathbb{1}_{\mathcal{H}_k} = \sum_k |p_k \circ \mathbf{F}\rangle\langle p_k|$, and its matrix representation is $\hat{\mathbf{K}}_{ij} = \langle p_i|\mathcal{K}|p_j\rangle_{\mathcal{H}_k} = \langle p_i|\mathbb{1}_{\mathcal{H}_k}\mathcal{K}|p_j\rangle_{\mathcal{H}_k} = [\Sigma^+ \mathbf{Q}^T \mathbf{K} \mathbf{Q} \Sigma^+]_{ij}$, where we plugged in the last two expressions of $\mathbb{1}_{\mathcal{H}_k}$ above, and $\mathbf{K}_{ij} = \langle k_{\mathbf{x}_i}|\mathcal{K}|k_{\mathbf{x}_j}\rangle_{\mathcal{H}_k} = k_{\mathbf{x}_j}(\mathbf{F}(\mathbf{x}_i)) = k(\mathbf{y}_i, \mathbf{x}_j) = \langle k_{\mathbf{y}_i}|k_{\mathbf{x}_j}\rangle_{\mathcal{H}_k}$ can be computed directly on training data. Similarly, the eigenvalue problem can be solved by computing eigenvalues and eigenvectors of $\hat{\mathbf{K}}$, where the i -th component of eigenvector \mathbf{v}_j is $(\mathbf{v}_j)_i = \langle p_i|\varphi_j\rangle_{\mathcal{H}_k}$, so the eigenfunction $|\varphi_j\rangle = \sum_i |p_i\rangle(\mathbf{v}_j)_i$. The point evaluation of an eigenfunction on training data is $\langle k_{\mathbf{x}_i}|\varphi_j\rangle_{\mathcal{H}_k} = \varphi_j(\mathbf{x}_i) = \sum_l \langle k_{\mathbf{x}_i}|p_l\rangle_{\mathcal{H}_k} (\mathbf{v}_j)_l = \sum_{nl} \langle k_{\mathbf{x}_i}|k_{\mathbf{x}_n}\rangle_{\mathcal{H}_k} \langle k_{\mathbf{x}_n}|q_l\rangle_{\mathcal{H}_k} \frac{1}{\sqrt{\sigma_l}} (\mathbf{v}_j)_l = [\mathbf{G} \mathbf{Q} \Sigma^+ \mathbf{V}]_{ij}$, where columns of \mathbf{V} are $\{\mathbf{v}_j\}$. By defining $[\Phi_x]_{ij} = \langle k_{\mathbf{x}_i}|\varphi_j\rangle_{\mathcal{H}_k}$ and $[\Phi_y]_{ij} = \langle k_{\mathbf{y}_i}|\varphi_j\rangle_{\mathcal{H}_k}$, we can write the matrix of eigenfunctions evaluated on training data in a compact form as $\Phi_x = \mathbf{G} \mathbf{Q} \Sigma^+ \mathbf{V}$ and $\Phi_y = \mathbf{K} \mathbf{Q} \Sigma^+ \mathbf{V}$. Following the same convention and notation in derivation of EDMD, the matrix of Koopman modes can be solved as $\Xi = \Phi_x^+ \mathbf{X} = \Phi_y^+ \mathbf{Y} = [\text{diag}(\mathbf{e}^{\lambda\tau})]^+ \Phi_x^+ \mathbf{Y}$, where rows in \mathbf{Y} are $\{\mathbf{y}^T\}$ and $[\text{diag}(\mathbf{e}^{\lambda\tau})]$ is the diagonal matrix containing the finite time eigenvalues $\mu_i = \mathbf{e}^{\lambda_i\tau}$. This procedure is called kernel-based Koopman spectral analysis [16] and it is currently being adopted as a better approach for other applications [5]. Finally, given a new system state \mathbf{x}_* , the prediction of the l -th component of system state $F_l(\mathbf{x}_*)$ will be a point evaluation of the Koopman operator evolved observable g_l at \mathbf{x}_* as

$$\begin{aligned} \langle k_{\mathbf{x}_*}|F_l\rangle_{\mathcal{H}_k} &= \langle k_{\mathbf{x}_*}|\mathcal{K}|g_l\rangle_{\mathcal{H}_k} = \sum_{i=1}^M \langle k_{\mathbf{x}_*}|\mathcal{K}|\varphi_i\rangle_{\mathcal{H}_k} \Xi_{il} = \sum_{i=1}^M \langle k_{\mathbf{x}_*}|\varphi_i\rangle_{\mathcal{H}_k} \mathbf{e}^{\lambda_i\tau} \Xi_{il} \\ &= \sum_{i=1}^M k(\mathbf{x}_*, \mathbf{x}_i) [\mathbf{Q} \Sigma^+ \mathbf{V} [\text{diag}(\mathbf{e}^{\lambda\tau})] \Xi]_{il}, \end{aligned} \quad (11)$$

where Ξ_{il} is the Koopman mode associated with the i -th eigenfunction when projecting $g_l(\mathbf{x})$ on Φ_x .

Another benefit of working in RKHS is that when properly choosing and/or designing the kernel functions (*e.g.*, Gaussian RBF kernel), the unique associated RKHS is dense in the space of continuous bounded functions, which means that these kernel functions are universal approximators to any function in this

very large and general function space, and hence they should achieve better approximation and prediction in most cases, especially in computing Koopman eigenfunctions via point evaluation $\varphi_j(\mathbf{x}_i) = \langle k_{\mathbf{x}_i} | \varphi_j \rangle_{\mathcal{H}_k}$.

3 Numerical algorithm

Recall Eq. (11), if one needs to predict all state variables at a future time, one can simply compute

$$\mathbf{F}(\mathbf{x}_*) = \mathbf{k}(\mathbf{x}_*)^T \mathbf{Q} \mathbf{\Sigma}^+ \mathbf{V} [\text{diag}(\mathbf{e}^{\lambda \tau})] \mathbf{\Xi}, \quad (12)$$

where $\mathbf{k}(\mathbf{x}_*)^T = [k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_M)]$. Notice that for system with input, all the \mathbf{x}_* , \mathbf{x}_i , and \mathbf{y}_i are extended states with input, but the Koopman modes $\mathbf{\Xi}$ will only contain N columns corresponding to the first N components of the extended state, which eliminates meaningless prediction on input. Another observation is that if we substitute $\mathbf{\Xi}$ in Eq. (12) by $\mathbf{\Xi} = [\text{diag}(\mathbf{e}^{\lambda \tau})]^+ \mathbf{\Phi}_x^+ \mathbf{Y}$, after some simplification, we will get $\mathbf{k}(\mathbf{x}_*)^T \mathbf{G}^+ \mathbf{Y}$, which is exactly the regularized optimization in RKHS or Gaussian processes regression on each state variable one-by-one. As we elaborated in Sec. 2.2, one of the major advantages of utilizing the spectral properties of Koopman operator is to linearly decompose the system dynamics as a summation over individual modes, such that it is possible to regularize, sort, perform more “physical” cross-validation, and optimize these modes in order to generate an ensemble of prediction models to achieve better prediction, as developed in Ref. [5]. When investigating time series with exogenous variables as a dynamical system with input, since the only major change on the numerical procedure is to neglect the Koopman modes associated with input, one can simply work with the remaining Koopman modes and all techniques and methods developed for Kernel-based Koopman modes regression (Kernel KMR)[5] can be employed almost unchanged. Hence we achieved a simple yet useful extension of Kernel KMR, which we refer to as *Kernel-based Extended Koopman mode regression with exogenous variables* (Kernel EKMRX). For more details on the techniques and methods constituting the Kernel KMR, we suggest referring to Ref. [5].

4 Numerical examples and applications

For the following reason, we will not include prediction results in the current paper, instead, we will present the complete results in the conference: (1) page limit rule, (2) we may need to obtain approvals to publish the prediction results on some dataset, due to customers privacy policies and others, (3) we are currently testing the algorithm on other data in order to better assess the capability of it and improve it if possible, hence we believe presenting more complete and comprehensive results in the conference will be more reasonable.

5 Conclusion and outlook

In this paper, we extended our previously developed Kernel KMR methodology to Kernel EKMRX (Kernel-based Extended Koopman Mode Regression with exogenous variables) for prediction of high dimensional time series with exogenous variables, by utilizing a simple yet useful generalization of Koopman operator to dynamical systems with input that generates the time series. We found that the techniques and methods that we developed for Kernel KMR can be employed in Kernel EKMRX with minimal modification. We re-emphasized the advantages of using spectral properties of Koopman operator for prediction purposes, and by formulating Koopman operator in reproducing kernel Hilbert space, we obtained a new derivation of the kernel-based EDMD and the original EDMD algorithms by using Dirac bra-ket notation. Moreover, we obtained a probabilistic interpretation of these numerical methods developed for deterministic Koopman operator by exploiting the connection between RKHS and Gaussian processes regression, and relate them to the stochastic Koopman and Perron-Frobenius operators. This connection and probabilistic interpretation are crucial to justify the application of existing data-driven deterministic Koopman spectral analysis to non-deterministic dynamical systems, and account for the advantage of kernel-based EDMD over original EDMD which relies on explicit choice of basis functions spanning the space where the Koopman operator is projected and approximated. In applications, we found that the prediction performance of this methodology is very promising in forecasting real world high-dimensional time-series with exogenous variables, especially on financial markets data and energy generation and consumption data.

This generalization of Koopman operator to systems with input is not unique, and we are keen to investigate other generalization for prediction purposes. Moreover, even the very simple trick in this generalization that we used in this paper can be developed further to investigate system with memory in the same way as for memoryless systems. These will be left for future work. Another possible improvement, which is still an open question, is the design of kernel functions. When utilizing Gaussian RBF kernels, it should be possible to optimize the kernel widths as hyper-parameters by some other more sophisticated techniques in machine learning. This, again, will be left for future investigation.

References

1. Budišić, M., Mohr, R.M., Mezić, I.: Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **22**(4), 047,510 (2012). DOI 10.1063/1.4772195
2. Cvitanović, P., Artuso, R., Mainieri, R., Tanner, G., Vattay, G.: *Chaos: Classical and Quantum*. Niels Bohr Inst., Copenhagen (2016)
3. Dirac, P.A.M.: A new notation for quantum mechanics. *Mathematical Proceedings of the Cambridge Philosophical Society* **35**(03), 416 (1939). DOI 10.1017/S0305004100021162
4. Hofmann, T., Schölkopf, B., Smola, A.J.: *A review of RKHS methods in machine learning* (2006)

5. Hua, J.C., Noorian, F., Moss, D., Leong, P.H., Gunaratne, G.H.: High Dimensional Time Series Prediction Using Kernel-Based Koopman Mode Regression. *Nonlinear Dynamics* (In Press) (2017). DOI 10.1007/s11071-017-3764-y
6. Klus, S., Koltai, P., Schütte, C.: On the numerical approximation of the Perron-Frobenius and Koopman operator. *Journal of Computational Dynamics* **3**(1), 1–12 (2016). DOI 10.3934/jcd.2016003
7. Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., Noé, F.: Data-driven model reduction and transfer operator approximation. *arXiv:1703.10112 [math]* (2017)
8. Koopman, B.O.: Hamiltonian Systems and Transformation in Hilbert Space. *PNAS* **17**(5), 315–318 (1931)
9. Lasota, A., Mackey, M.C.: *Chaos, Fractals, and Noise, Applied Mathematical Sciences*, vol. 97. Springer New York, New York, NY (1994). DOI 10.1007/978-1-4612-4286-4
10. Mezić, I.: Spectral Properties of Dynamical Systems, Model Reduction and Decompositions. *Nonlinear Dyn* **41**(1-3), 309–325 (2005). DOI 10.1007/s11071-005-2824-x
11. Proctor, J.L., Brunton, S.L., Kutz, J.N.: Generalizing Koopman Theory to allow for inputs and control. *arXiv:1602.07647 [math]* (2016)
12. Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning. Adaptive computation and machine learning*. MIT Press, Cambridge, Mass (2006)
13. Rowley, C.W., Mezić, I., Bagheri, S., Schlatter, P., Henningson, D.S.: Spectral analysis of nonlinear flows. *J. Fluid Mech.* **641**, 115–127 (2009). DOI 10.1017/S0022112009992059
14. Schölkopf, B., Herbrich, R., Smola, A.J.: A Generalized Representer Theorem. In: G. Goos, J. Hartmanis, J. van Leeuwen, D. Helmbold, B. Williamson (eds.) *Computational Learning Theory*, vol. 2111, pp. 416–426. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
15. Williams, M.O., Kevrekidis, I.G., Rowley, C.W.: A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition. *J Nonlinear Sci* pp. 1–40 (2015). DOI 10.1007/s00332-015-9258-5
16. Williams, M.O., Rowley, C.W., Kevrekidis, I.G.: A Kernel-Based Approach to Data-Driven Koopman Spectral Analysis. *arXiv:1411.2260 [math]* (2014)
17. Williams, M.O., Rowley, C.W., Mezić, I., Kevrekidis, I.G.: Data fusion via intrinsic dynamic variables: An application of data-driven Koopman spectral analysis. *EPL* **109**(4), 40,007 (2015). DOI 10.1209/0295-5075/109/40007

Nonlinear Dynamical Analysis of Twitter Time Series

Andrey Dmitriev, Vitaly Silchev, Victor Dmitriev, and Svetlana Maltseva

National Research University Higher School of Economics, Moscow, Russia
a.dmitriev@hse.ru

Abstract. In this paper we present the results of nonlinear dynamical analysis of Twitter time series. According to these results we compare nonlinear dynamical model and nonlinear random dynamical model of Twitter with observed data. From results of nonlinear analysis of observed Twitter time series and evaluation of their probability density functions we conclude, that the most adequate forecasting model of social network is nonlinear random dynamical system. We determine that observed TTS have q -exponential distribution with $1/f^\beta$ noise. Also we consider possible applications of Tsallis entropy and self-organized criticality for analysis of Twitter.

Keywords: Twitter time series·Fractal dimensions· q -exponential distribution· $1/f$ noise·Nonlinear dynamical system· Nonlinear random dynamical system

1 Introduction

Microblogging is one of the most important instruments of business development nowadays. It is actively used for promotion of goods or services, making the positive opinion about the company and allows organizing and supporting customer relationships processes. Corporate microblogging networks and services serve as a platform for business communications between the employees in companies on different scales.

Modeling of processes taking place in microblogging social networks (one of the well-known examples is Twitter) is a complicated, but at the same time theoretically and practically important scientific problem. Results and conclusions that can be made by using such models allow us to identify whether the social network is able to remain stable under the internal and external informational influence, to define different ways of local community formation and to find out the parametric terms of social network management. Such modeling may have a large variety of practical applications. Thus, it can be useful for decision-making processes during the development of short-term and long-term marketing strategies, development of recommender systems, demand forecasting, as well as tasks related to the national security.

There are a number of works in the field of physical modeling of social networks. The main physical models of the social networks are following: Ising model [1-3], Bose-Einstein condensate model [4, 5], quantum walk model [6], ground state and community detection[7], etc. The other relevant works in this area are those of refs. [8-12].

adfa, p. 1, 2011.

© Springer-Verlag Berlin Heidelberg 2011

The weak point of the observed papers is that they do not cover nonlinear dynamical analysis of aggregated twitter time series¹ (TTS). Results of such analysis can provide a possibility to select the most appropriate prediction methods for TTS and give a general idea about adequate models of social networks generating these signals.

Recently, more attention has been paid to the study of time series from the point of view of chaos theory. Research in this direction will reveal the nature and interconnections between the hidden processes occurring in microblogging social networks, which will enable the construction of more adequate forecasting models for TTS and a deeper understanding of social networks functioning.

Analysis of chaotic phenomena requires methods and techniques for identifying of time series that is chaotic or having a chaotic component, as well as for quantitative evaluation of chaotic characteristics and comparison of theoretical and experimental time series. Having these methods and techniques allows one to answer the following problems: 1) the number of variables essential for modeling of system dynamics; 2) relation between changes in characteristics and changes in dynamical behavior of the system.

These methods and techniques are grouped into two different, but connected approaches. The first approach focuses on dynamical characteristics of chaos: the Lyapunov exponents and entropy measures, power spectral density and autocorrelation function. The second approach represents the geometric nature of trajectories in the state space considering fractal and correlation dimensions.

These two approaches complete each other. It is intuitively expected that they are closely interconnected. However, theoretical proof of such connection has not been developed yet. That is why we used several criteria of chaotic nature of time series.

This paper is organized as follows. In section 2 we present the results of fractal analysis for empirical TTS with their interpretation. In section 3 we present the results of fractal analysis and probability density function (PDF) for a sample of 3-dimensional nonlinear dynamical model of Twitter network as an open nonequilibrium system[13], as well as comparison with empirical results. In section 4 we provide the results of fractal analysis and PDF for the model of Twitter network as nonlinear random dynamical system comparing them with empirical results and describe the possibilities of applying the Tsallis entropy and self-organized criticality for analysis of TTS. Section 5 contains the conclusions of this paper.

2 Analysis of an Empirical Twitter Time Series

For analysis of empirical TTS we chose the following time series obtained from the resource Mozdeh "BigDataTextAnalysis" (<http://mozdeh.wlv.ac.uk/>):

- bbc_breaking, from 16/05/29 to 17/05/26, step 1 hour;
- cnn_braeking, from 16/07/12 to 17/01/11, step 1 hour;
- nasa, from 16/09/26 to 17/05/26, step 1 hour.

¹ Series of tweet and retweet numbers indexed in time order, TR_t .

Figure 1 shows the corresponding time series.

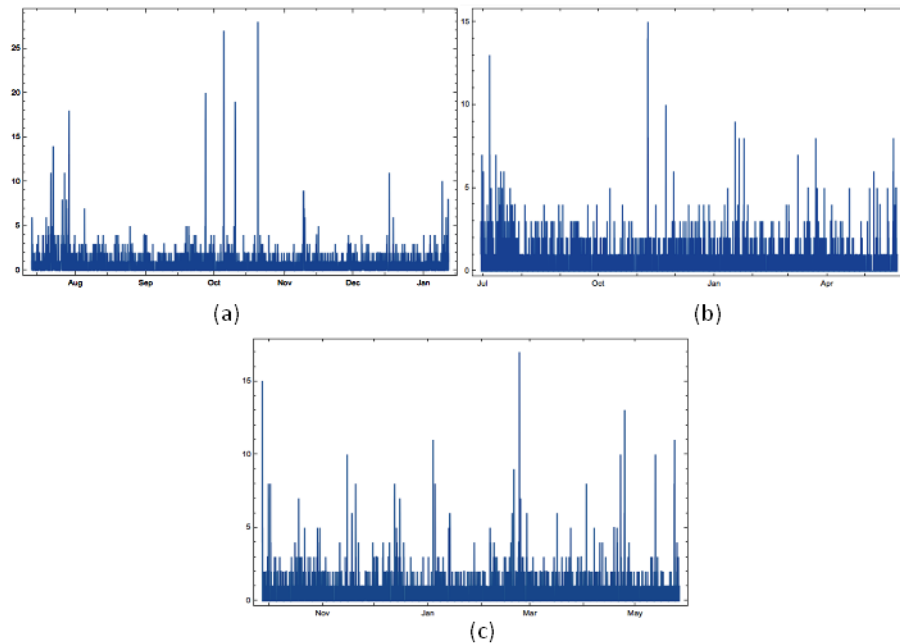


Fig. 1. Twitter time series: (a) bbc_breaking, (b) cnn-breaking, (c) nasa

It is clear, that these time series represent impulse-type signals with integer values.

The nonlinear analysis was conducted for all chosen TTS. Such measures as correlation dimension (D_2), embedding dimension (m), Hurst exponent (H) and fractal dimension (D_F) were calculated (table 1).

Table 1. Measures of chaos

Time series	D_2	m	H	D_F
bbc_breaking	3.732	6	0.7648	1.2352
cnn_breaking	3.984	6	0.8165	1.1835
nasa	4.202	6	0.7833	1.2167
Dynamical system	1.896	3	0.5328	1.4272
Random dynamical system	4.619	5	0.7872	1.2128

The determination of the correlation dimension [14] for a supposed chaotic process directly from experimental time series is often used to gain information about the nature of the underlying dynamics (see, for example, contributions in ref. [15]. In particular, such analysis has been made to support the hypothesis that the time series are generated from the inherently low-dimensional chaotic process [15]. The geome-

try of chaotic attractors can be complex and difficult to describe. It is therefore useful to understand quantitative characterizations of such geometrical objects. One of these characterizations is D_2 . D_2 has several advantages in comparison to the other dimensional measures:

- D_2 is easy to compute from the TTS;
- If D_2 is finite, then the TTS is a chaotic time series (generated by a dynamical system);
- If $D_2 \rightarrow \infty$, then the TTS is a stochastic time series (generated by a purely random process).

The correlation dimension of the attractor of dynamical system can be estimated using the Grassberger–Procaccia algorithm [14].

For calculation of D_F we used the algorithm, described in a paper [16]. If $D_F > d_T$ (d_T is a topological dimension of the TTS, that equals 1 for all time series), then the TTS is a random fractal. A value of $H = 2 - D_F$ characterizes the following features of the TTS:

- If $H > 0.5$, then the TTS represents a persistent process (a positive increment of a number of tweets and retweets in the past on the average means that there is a tendency to further increase in future, and vice versa);
- If $H < 0.5$, then the TTS represents an anti-persistent process (a positive increment in a number of tweets and retweets in the past on the average means that there is a tendency to decrease in future, and vice versa);
- If $H = 0.5$, then the TTS represents an intermediate state between the persistent and anti-persistent processes (the TTS is a stochastic time series).

In addition, the value of H allows to give a noise classification ($1/f$ -classification, where f is a signal frequency) of the TTS [17]:

- If $0 < H \leq 0.5$, then the TTS represents a process with the negative memory, $1/f$ noise or a pink noise (if there has been the positive increment in a number of tweets or retweets, then there is a high probability of appearance of the negative increment in future, and vice versa);
- If $0.5 < H \leq 1$, then the TTS represents a process with a positive memory, $1/f^\beta$ ($\beta > 2$) noise or a brown noise (if there has been the positive increment in a number of tweets or retweets, then there is a high probability of appearance of the positive increment in future, and vice versa);
- If $H = 0.5$, then TTS represents a process with the absence of memory, $1/f^2$ noise or brown noise (the next increment in the number of tweets and retweets doesn't depend on the previous increments).

Thus, according to the point values of measures, shown in a table 1, the following conclusions can be made:

- TTS is a chaotic time series, i.e. it is generated by dynamical systems in a phase space dimension that equals 6;

- TTS has a fractal structure;
- TTS represents processes with the positive memory;
- TTS represents the persistent process;
- TTS is a signal with the $1/f^\beta$ noise (in support of that, fig. 2 provides the spectral power density plots in log-log scale for corresponding TTS).

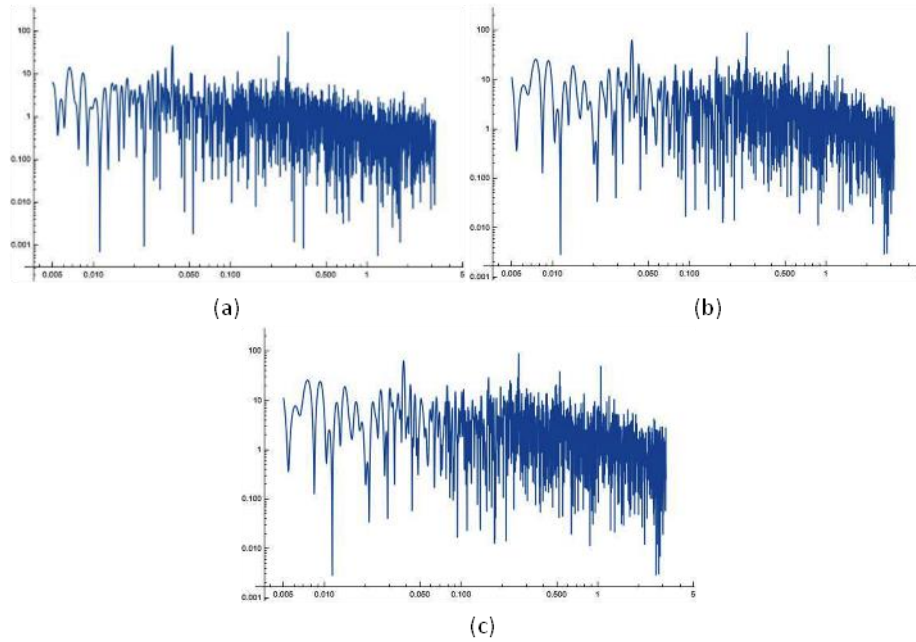


Fig. 2.Power spectral density for TTS: (a) bbs_breaking, (b) cnn_breaking, (c) nasa

3 Twitter Time Series as a Realization of the Nonlinear Dynamical System

Paper [13] proposes a model of Twitter social network as an open nonequilibrium system. Omitting the detailed construction of dynamical system, the model of Twitter is described by well-known Lorenz–Haken equations:

$$\dot{x}_1 = -\alpha x_1 + \beta x_2, \quad \dot{x}_2 = -\gamma x_2 + c x_2 x_3, \quad \dot{x}_3 = \varepsilon(I_0 - x_3) + k x_1 x_2 \quad (1)$$

In equation (1) $x = TR(t) - TR_{eq}$ represents the scaled deviation of number of tweets and retweets ($TR(t)$) from equilibrium value TR_{eq} ; $x_2(t) = I(t) - I_{eq}$ is the scaled deviation of aggregated internal amount of information ($I(t)$) from equilibrium value I_{eq} ; $x_3(t) = N_{|u\rangle}(t) - N_{|l\rangle}(t)$ is instantaneous difference in number of users between state $|u\rangle$ and state $|l\rangle$. According to the model, a particular user, being $|u\rangle$ -state, has enough information for sending tweet or retweet. If the user is in $|l\rangle$ -state (so, he

or she does not have enough amount of information), then he or she will not send any tweets or retweets. Control parameter I_0 is the intensity of external information flow.

The most important conclusions from model implementation are: 1) impossibility of social network being in equilibrium state and occurrence of low-dimensional chaos [18] in social network for specific values of I_0 . Fig. 3 as hows integral trajectory of dynamical system (1) ($x_1(t)$), demonstrating the existence of chaotic dynamics in case of significant intensity of external information flow.

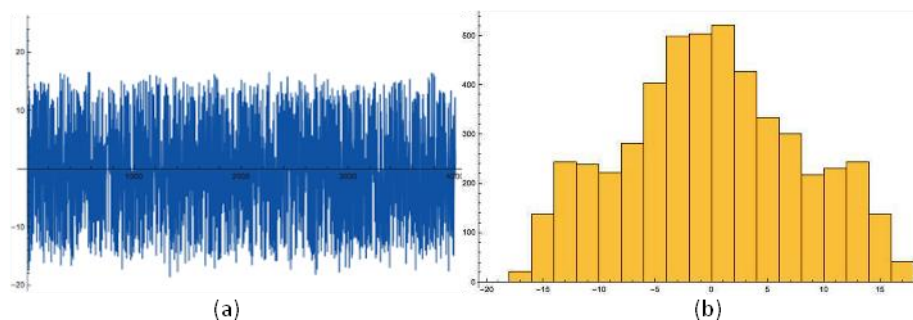


Fig. 3. Integral trajectory (a) and its histogram (b)

Except for values of higher Lyapunov exponent [19] as one of the measures of low-dimensional chaos, paper [13] does not contain calculated fractal dimensions for observed TTS.

Estimations of measures of the chaos for theoretical TTS (fig. 3a). Table 1 contains the estimated values for measures of chaos for the theoretical TTS (see dynamical system). Thus, 3-dimensional dynamical model of Twitter as open nonequilibrium system [13] explains some properties of social network functioning such as fractality, chaotic nature, persistency and positive memory of TTS.

The weakness of this model lies in significant discrepancy between empirical (fig. 1) and theoretical (fig. 3a) trajectories of TTS. Moreover, it is impossible to fit theoretical trajectories to observed data by varying control parameters (in range of chaotic state) of dynamical system [13]. As it shown on fig. 3b, this dynamical system has 3 stable equilibrium points (three maxima of the histogram) for any values of control parameters in range of chaotic state.

There are at least two possible ways to achieve the fitness between empirical and theoretical TTS: by adding specific noise to dynamical system [13] or by using one-dimensional nonlinear random dynamical system [20] as a model of Twitter network. According to table 1 at $n = 6$ the estimated value of correlation dimension reaches its "saturation point" and stops changing significantly. Because of that, the actual number of variables for constructing an adequate model is 6, but not 3 as it is for model [13]. We do not rule out, that six-dimensional model of Twitter network could explain existing experimental characteristics, including empirical PDF of Twitter time series.

4 Twitter Time Series as a Realization of the Nonlinear Random Dynamical System

In such autonomous dynamical systems as $\dot{\mathbf{X}} = \mathbf{F}(\mathbf{X})$, low-dimensional chaos can appear only at $n \geq 3$ [18]. Therefore, the one of opportunity to build an adequate model of a microblogging network is to consider it as a random dynamical system (RDS). In this case, the observable TTS is one of the realizations of $x(t)$ of a stochastic differential equation of the following kind:

$$dx = f(x, t)dt + g(x, t)dW \quad (2)$$

where $W(t)$ is a standard Wiener process.

One of the ways to solve the equation (2) is to find its solution in a form of a probability density function (PDF) $p(x, t)$. In this case, the equation (2) can be transformed into the Fokker-Planck equation [21], that represents a differential equation in partial derivatives of the following kind:

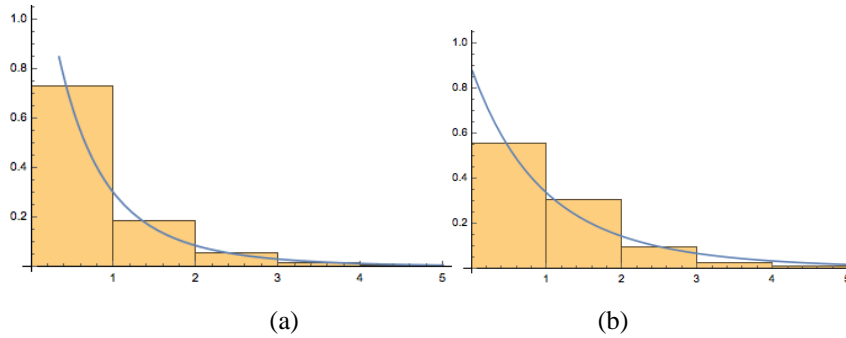
$$\frac{\partial p(x, t)}{\partial t} = -\frac{\partial}{\partial x}(f(x)p(x, t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(g^2(x)p(x, t)) \quad (3)$$

In this case it is necessary to define the PDF for the empirical TTS (a stationary solution of (3)). Having found out the explicit kind of PDF, we shall be able to find out the explicit kind of (1), describing the realizations of the empirical TTS.

Figure 4 provides PDFs for empirical TTS, which form point to the fact that it is q -exponential distribution [22-24]:

$$p(x) = (2 - q)\lambda \exp_q(-\lambda x) \quad (4)$$

where $\exp_q(x) = [1 + (1 - q)x]^{\frac{1}{1-q}}$.



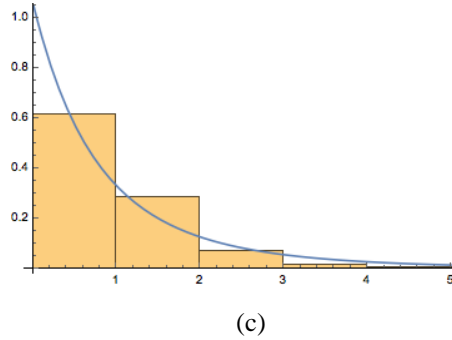


Fig. 4. Histograms of TTS: (a) bbc_breaking, (b) cnn_breaking, (c) nasa

The distribution (4) is a two-parameter generalization ($q < 2$ is a shape parameter, $\lambda > 0$ is a rate parameter) of a one-parameter exponential distribution. Table 2 contains the estimated values for parameters of PDF (4) obtained by maximum likelihood method [25].

Table 2. Point and interval estimations of the PDF (4) parameters

User	q	λ
bbc_breaking	1.202±0.005	1.980±0.074
cnn_breaking	1.155±0.025	1.482±0.086
nasa	1.184±0.038	1.362±0.069

From table 2 we conclude that empirical PDF corresponds to q -exponential distribution.

Going back to the equation (2): a stationary probability density function of the TTS looks as (4) with the numerical parameter values shown in a table 2 and is a stationary solution of the equation (3). Therefore, the equation (3) should be of such kind, that gives the distribution (4) for all realizations of the random dynamical system.

A group of researchers [26-28] has suggested the RDS in a view of a nonlinear stochastic differential equation:

$$dx = \sigma^2 \left(\eta - \frac{1}{2} \lambda \right) (x + x_0)^{2\eta-1} dt + \sigma (x + x_0)^\eta dW \quad (5)$$

where $x(t) \geq 0$ is a signal, $\eta \neq 1$ is a power-law exponent of the multiplicative noise, $\lambda > 0$ is a parameter, defining the behavior of stationary probability distribution, W is a standard Wiener process, σ is a parameter of the multiplicative noise. Parameter x_0 limits the divergence of the power-series distribution $x(t)$ by $x(t) \rightarrow 0$. If $x \ll x_0$, then (5) generates a linear additive stochastic process (Brownian movement with the stable drift); if $x \gg x_0$, then (5) generates a multiplicative process [27].

If $x_0 = 1$, then the stationary solution of the equation (3) takes the form of an q -exponential distribution (4) by $q = 1 + 1/\lambda$. Besides, some of realizations of the process (5) give a power spectral density in a form of $1/f^\beta$.

We have calculated estimations of the measures of chaos for some realizations of RDS (5). Table 1 contains the estimated values for measures of chaos for the theoretical TTS (see random dynamical system).

Thus, the realizations of the RDS (5) have not only close measures to the observable fractal measures of the TTS (table 1) in comparison to the realizations of the dynamical system [13], but they also have an observable (table 2) q -exponential distribution. Therefore, the RDS (5) is more adequate model in comparison to the model in a form of the dynamical system [13].

q -exponential distribution takes place by the maximization of the Tsallis entropy [29] considering definite limitations. Tsallis entropy as a non-additive generalization of the Boltzmann-Gibbs entropy has the following form:

$$T_q = \frac{1}{q-1} (1 - \sum_{i=1}^N p_i^q) \quad (6)$$

The probability $p_i = N_i/N(\varepsilon)$ can be estimated in much the same way as that one used in the Renyi entropy: N_i is a number of system elements for the i -element of the ε -partition; $N(\varepsilon)$ – is a full number of elements of the given ε -cover. If $q \rightarrow 1$, then the entropy (6) transforms into the well-known Shannon entropy.

In contrast to all entropy types, the Tsallis entropy is nonadditive. Being applied to the microblogging network (such as, for example, Twitter) it gives a possibility to correctly describe a social network, where any user interacts not only with the nearest user or several nearest users, but also with the whole network or some of its parts. Besides, from (5) it follows that T_q is concave by $q > 0$ and convex by $q < 0$.

Thus, entropy description of Twitter based on Tsallis statistics is appropriate for studying of evolution of social network that contains large amount of users who interact with each other in a particular way and, specifically, every user can interact not only with his or her nearest neighbors but also with remote users.

There are a lot of practical application of Tsallis theory. Among them there are studies on the anomalous diffusion [30, 31], uniqueness theorem [32], sensitivity to initial conditions and entropy production at the edge of chaos [33] and many others (see ref. [34]).

The fact, that the RDS (5) generates a signal with the power-series distribution (4) and with the occurrence of the $1/f^\beta$ noise [35], is the important feature of the RDS (5). It is determined by the existence of the degree $2\eta - 1$ in the drift term and degree η in the noise term. The same fact is observable for the empirical TTS as well.

The existence of the power laws of signal distribution with the presence of the $1/f^\beta$ noises (see fig. 2) is a necessary condition of system complexity, its nontrivial behavior or presence of the catastrophic events (unexpected and/or extraordinary). There is a relatively new field in non-linear dynamics – a theory of the self-organized criticality [36]. It was created to explain similar phenomena in systems with the power-series distributions and $1/f^\beta$ noises.

The existence of the $1/f^\beta$ noise in a system means the internal tendency to the catastrophic cases in a system. The theory of the self-organized criticality studies the dynamical dissipative systems with the high range of discretion, which operate in the neighborhood of the critical point without the smallest external influence. If the sys-

tem is in a critical configuration, than small fluctuations can lead to a random event of any “size” with the power-series distribution similar to (4):

$$p(s) \sim s^{-\tau} \quad (7)$$

Twitter as a self-organizing system generates signals with $1/f$ noise, since the lifetime of events is related to its scale according to [36]:

$$t^{1+\gamma} \approx s \quad (8)$$

where γ is the speed of event distribution in the system.

5 Conclusion

The main contributions of the present paper look as follows:

- The three-dimensional model of the microblogging network [13] (such as, for example, Twitter) as an open non-equilibrium system explains some features of social networks functionality, such as the fractality, chaotic state, persistence, as well as the positive memory of the TTS. But, at the same time, the dimension test of such dynamical system gives the negative result: empirical embedding dimension of all TTS equals to 6 (by $n = 6$ the correlation dimension reaches the saturation and stops changing). This fact leads to the necessity of building a new model of a microblogging network in a form of nonlinear RDS.
- We have conducted a research into the empirical PDF of some TTS to build a model of the microblogging network in a form of one-dimensional non-linear RDS. As a result it has been recognized that at the significance level equal to 0.05 the observable PDF has a q -exponential distribution. For such distribution, the one-dimensional nonlinear RDS has been suggested. The fractal measures of its realizations are equivalent to the measures of the observable TTS.
- It has been shown, that in contrast to all entropy types, the Tsallis entropy gives a possibility to correctly describe a network, where any user interacts not only with the nearest user or several nearest users, but also with the whole network or some of its parts. Use of the Tsallis entropy also allows to describe the macroscopic stability of a microblogging network.
- It has also been mentioned, that because of the existence of the $1/f^\beta$ noise and power series distribution, a social network may have a tendency to catastrophic events. If a social network keeps staying in a critical configuration, then small fluctuations may lead to the random event of any scale.

Despite the fact, that the results of the present study can be useful for the research into the fundamentals of the network functionality, we haven't yet defined the physical meaning of parameters of the one-dimensional nonlinear RDS. That is the question of our further research.

References

1. Grabowski A, Kosinski RA (2006) Ising-based model of opinion formation in a complex network of interpersonal interactions. *Physica A* 361: 651–664.
2. Dasgupta S, Pan RK, Sinha S (2009) Phase of Ising spins on modular networks analogous to social polarization. *Physical Review E* 80: 025101-1.
3. Bianconi G (2002) Mean field solution of the Ising model on a Barabasi-Albert network. *Phys. Lett. A* 303, 166-204.
4. Bianconi G, Barabasi AL (2001) Bose-Einstein Condensation in Complex Networks. *Phys. Rev. Lett.* 86: 5632-5635.
5. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74: 47–97.
6. Faccin M, Johnson T, Biamonte J, Kais S (2013) Degree Distribution in Quantum Walks on Complex Networks. *Phys. Rev.* 3: 041007.
7. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys. Rev. E* 74: 016110.
8. Mendes V (2005) Tools for network dynamics. *J. Bifurcation Chaos* 15: 1185.
9. Ebel H, Davidsen J, Bornholdt S (2003) Dynamics of Social Networks. *Complexity* 8: 24–27.
10. Toivonen R, Onnela JP, Saramaki J, Hyvonen J, Kaski K (2006) A model for social networks. *Physica A* 371: 851–860.
11. Toivonen R, Onnela JP, Saramaki J, Hyvonen J, Kaski K (2009) A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks* 31: 240–254.
12. Skaza J, Blais B (2017) Modeling the infectiousness of Twitter hashtags. *Physica A* 465: 289–296.
13. Dmitriev AV, Tsukanova OA, Maltseva SV (2016) Investigation into the Regular and Chaotic states of Microblogging Networks as Applied to Social Media Monitoring. In: 13th IEEE International Conference on e-Business Engineering. IEEE Press, 293–298.
14. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. *Physica D* 9: 189–208.
15. Ding M, Grebogi C, Ott E, Sauer T, Yorke J (1993) Estimating correlation dimension from a chaotic time series: when does plateau onset occur? *Physica D* 69: 404–424.
16. Dubovikov MM, Starchenko NS, Dubovikov MS (2004) Dimension of the minimal cover and fractal analysis of time series. *Physica A* 339: 591–608.
17. Mandelbrot BB, Ness V (1968) Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* 10: 422–437.
18. Hilborn RC (2000) *Chaos and nonlinear dynamics: an introduction for scientists and engineers*. Oxford University Press, United Kingdom.
19. Kuznetsov NV, Leonov GA (2005) On stability by the first approximation for discrete systems. In: Proceedings of the International Conference on Physics and Control, 596–599.
20. Arnold L (1998) *Random Dynamical Systems*. Springer-Verlag, Berlin.
21. Risken H (1984) *The Fokker – Planck Equation: Methods of Solutions and Applications*. Springer.
22. Tsallis C (1994) What are the numbers that experiments provide? *Quimica Nova* 17: 68–471.
23. Tsallis C (2009) Nonadditive entropy and nonextensive statistical mechanics—an overview after 20 years. *Braz. J. Phys.* 39: 337–356.

24. Picoli S, Mendes RS, Malacarne LC, Santos RPB (2009) q-distributions in complex systems: a brief review. *Braz. J. Phys.* 39: 468-474.
25. Zhang F, Shi Y, Ng H, Wang R (2016) Tsallis statistics in reliability analysis: Theory and methods. *Eur. Phys. J. Plus* 131: 379.
26. Ruseckas J, Gontis V, Kaulakys B (2012) Nonextensive statistical mechanics distributions and dynamics of financial observables from the nonlinear stochastic differential equations. *Advances in Complex Systems* 15: 1250073.
27. Ruseckas J, Kaulakys B (2011) Tsallis distributions and $1/f$ noise from nonlinear stochastic differential equations. *Phys. Rev. E* 84: 0511125.
28. Kaulakys B, Alaburda M, Gontis V, Ruseckas J (2009) Modeling long-memory processes by stochastic difference equations and superstatistical approach. *Braz. J. Phys.* 39: 453-456.
29. Tsallis C (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* 52: 479-487.
30. Plastino AR, Plastino A (1995) Non-extensive statistical mechanics and generalized Fokker-Planck equation. *Physica A* 222: 347-354.
31. Tsallis C, Bukman D (1996) Anomalous diffusion in the presence of external forces: Exact time-dependent solutions and their thermostistical basis. *Physical Review E* 54: R2197.
32. Abe S (2000) Axioms and uniqueness theorem for Tsallis entropy. *Physics Letters A* 271: 74-79.
33. Baldovin F, Robledo A (2004) Nonextensive Pesin identity: Exact renormalization group analytical results for the dynamics at the edge of chaos of the logistic map. *Physical Review E* 69: 045202-1.
34. Tsallis C (1999) Nonextensive Statistics: Theoretical, Experimental and Computational Evidences and Connections. *Braz. J. Phys.* 29: 1-35.
35. Kaulakys B, Alaburda M (2009) Modeling scaled processes and $1/f^\beta$ noise using nonlinear stochastic differential equations. *J. Stat. Mech.* P02051.
36. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of $1/f$ noise. *Physical Review Letters* 59: 381-384.

Interpolation of ARMA processes with infinitely divisible white noise

Argimiro Arratia, Alejandra Cabaña, and Enrique M. Cabaña

¹ Barcelona Tech (UPC), Dept. of Computer Science, Barcelona, Spain
`argimiro@cs.upc.edu`

² Departament de Matemàtiques, Universitat Autònoma de Barcelona (UAB) ,
Spain. `acabana@mat.uab.cat`

³ IMRL, Universidad de la República (Udelar), Montevideo, Uruguay.
`ecabana@fing.edu.uy`

Abstract. In this short presentation we sketch the following fact: any stationary time series satisfying an ARMA(p, q) model, for arbitrary p and q , can be embedded in a continuous time stationary process. Specifically we show that, given the stationary process X_t , $t \in \mathbb{Z}$, that satisfies the ARMA model $X_t = \sum_{j=1}^p \phi_j X_{t-j} + \sum_{k=0}^q \theta_k \epsilon_{t-k}$ where ϵ_t is an infinitely divisible white noise with finite variance, there exist a centred second-order Lévy process Λ_s , $s \in \mathbb{R}$, and a square integrable function $l(t)$, $t \in \mathbb{R}^+$ that vanishes exponentially at infinity, such that the restriction to $t \in \mathbb{Z}$ of the stationary process $Z_t = \int_{-\infty}^t l(t-s) d\Lambda_s$, $t \in \mathbb{R}$ has the law of the given ARMA(p, q) model.

Keywords: Continuous ARMA, Lévy process, embedding

1 Introduction

Given the stationary process X_t , $t \in \mathbb{Z}$, that satisfies the discrete ARMA model (DARMA)

$$X_t = \sum_{j=1}^p \phi_j X_{t-j} + \sum_{k=0}^q \theta_k \epsilon_{t-k} \quad (1)$$

where ϵ_t is white noise with finite variance, the problem of obtaining a process satisfying a continuous version of the DARMA model (a CARMA), such that when sampled at discrete times has the same autocovariance function as $\{X_t\}$ has been studied by several authors and termed the *embedding problem*. The works by [8], [10], [3] and [5] established embeddings of some DARMA(p, q) processes in continuous ARMA(p, q), for $0 \leq q < p$. [11] gave necessary and sufficient conditions for a DARMA process to be embedded in a CARMA process.

Brockwell [3, 4] proposes to define CARMA processes via a state space representation of the formal equation

$$a(D)Y(t) = \sigma b(D) d\Lambda(t)$$

where $\sigma > 0$ is a scale parameter, D denotes differentiation w.r.to t , Λ is a second-order Lévy process, $a(z) = z^p + a_1 z^{p-1} + \dots + a_p$ is a polynomial of order p and $b(z) = b_0 + b_1 + \dots + b_q z^q$ is a polynomial of order q . The resulting CARMA is a linear function of a CVAR Markovian process.

This formalism has some limitations:

- If q is not smaller than p , it requires the use of generalised processes [9].
- Even for $q < p$, not every DARMA processes are embeddable.

All these approaches to the embedding problem are only concerned with the covariance structure of the processes involved, not with their probability distributions besides the fact that, if the processes are Gaussian, the equality of the first- and second-order moments entails the equality of the probability laws. In general, the discretised version of the CARMA will not necessarily have the same law as the original DARMA. We propose in this work a different approach to construct for any DARMA(p, q) a continuous stationary *embedding in law*. The precise statement is the following:

Theorem 1. *Given the stationary DARMA(p, q) X_t that satisfies (1) with infinitely divisible innovations ϵ_t , there exists at least one function $L : \mathbf{R}^+ \rightarrow \mathbf{R}$ decaying exponentially at infinity and a Lévy process Λ on \mathbf{R} , such that for each real number a the stationary processes $x_t = \int_{-\infty}^t L(t-s)d\Lambda(s)$, $t \in \mathbf{R}$, sampled at times $a+t$, $t \in \mathbf{Z}$, have the same joint law as X_t .*

In the following sections we sketch the construction of the processes x_t . Details and proofs can be found in the extended version of this conference paper [2].

2 A stationary embedding

We show in the sequel that, given the stationary DARMA X_t , $t \in \mathbf{Z}$, there exist a centred second-order Lévy process Λ_s , $s \in \mathbf{R}$, and square integrable functions $\ell(t)$, $t \in \mathbf{R}^+$ that vanish exponentially at infinity, such that the restriction to $t \in \mathbf{Z}$ of the stationary processes

$$\int_{-\infty}^t \ell(t-s)d\Lambda_s, t \in \mathbf{R} \quad (2)$$

satisfy the given ARMA(p, q) model.

The construction makes use of a similar embedding for vectorial autoregressive (VAR) processes, driven by an infinitely divisible white noise, and an optimisation procedure for choosing ℓ . The function ℓ can be chosen so as to satisfy certain optimisation criteria, as shown in §4. This means that any stationary time series satisfying a DARMA(p, q) model for arbitrary p and q can be embedded in a continuous parameter stationary process.

One possible optimisation criterion, not always feasible, depending on the DARMA model and on the distribution of the noise, leads to the well known

CARMA models as in [5, 7], or Lévy driven generalisations of the processes in [1]. With a different criterion, we derive other interpolations, with no constraints on the model, nor on the nature of the noise.

The construction of the embedding is made in several steps:

1. A DARMA(p, q) in \mathbf{R} can be expressed as a DVAR(1) in \mathbf{R}^r , $r = p \vee (q + 1)$.
2. DVAR(1) in \mathbf{R}^r can be expressed as several canonical J-DVAR(1) in \mathbf{R}^{r_i} , $\sum r_i = r$ as a result of a Jordan Canonical decomposition.
3. Each J-DVAR(1) in \mathbf{R}^{r_i} can be embedded in a continuous time process J-CVAR in \mathbf{R}^{r_i} .
4. Finally, the J-CVAR in \mathbf{R}^{r_i} , $i = 0, 1, 2, \dots$ can be joined to get the continuous embedding of DARMA.

2.1 From DARMA(p, q) to DVAR(1) in \mathbf{R}^r , $r = p \vee (q + 1)$

Let $(\epsilon_t)_{t \in \mathbf{Z}}$ denote a standardized white noise, that is, the ϵ_k are i.i.d. with $\mathbf{E}\epsilon_1 = 0$, $\mathbf{E}\epsilon_1^2 = 1$, and D the $r \times r$ matrix

$$D = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \dots & \phi_{r-1} & \phi_r \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix}, \text{ and } \boldsymbol{\eta}_t = \mathbf{u}\epsilon_t, \mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

a vectorial white noise.

Then it is well known (Example 8.3.2. in [6]) that the stationary series satisfying the DARMA(p, q) model (1) can be expressed as the linear function

$$X_t = \boldsymbol{\theta}^{\text{tr}} \boldsymbol{\xi}_t = \sum_{k=0}^q \theta_k \xi_{t,k}$$

of the DVAR(1)

$$\boldsymbol{\xi}_t = D\boldsymbol{\xi}_{t-1} + \boldsymbol{\eta}_t \quad \boldsymbol{\xi}_t = (\xi_{t,1}, \xi_{t,2}, \dots, \xi_{t,r})^{\text{tr}} \quad (3)$$

where $r = \max\{p, q + 1\}$, $\phi_j = 0$ for $j > p$, and $\boldsymbol{\theta}^{\text{tr}} = (\theta_0, \theta_1, \dots, \theta_{r-1})$, with $\theta_k = 0$ for $k > q$.

2.2 From DVAR(1) to J-DVAR(1)

Let C denote the matrix that carries D to its Jordan canonical form $J = C^{-1}DC$ where

$$J = \begin{pmatrix} J_{\rho_0} & 0 & 0 & \dots & 0 \\ 0 & J_{\rho_1} & 0 & \dots & 0 \\ 0 & 0 & J_{\rho_2} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & J_{\rho_k} \end{pmatrix}.$$

The block $J_{\rho_h} = \rho_h I_{m_h} + I_{1,m_h}$ is associated to the eigenvalue ρ_h with multiplicity m_h . For each m , I_m is the $m \times m$ identity matrix and $I_{1,m}$ is the $m \times m$ matrix with the first sub-diagonal of ones and all other entries equal to zero.

Then the change $\xi = C\zeta$ leads to express the DVAR(1)

$$\xi_t = D\xi_{t-1} + u\epsilon_t$$

in the canonical form

$$\zeta_t = J\zeta_{t-1} + C^{-1}u\epsilon_t \quad (4)$$

equivalent to the $k+1$ canonical equations to be treated separately

$$\zeta_{h,t} = J_{\rho_h}\zeta_{h,t-1} + c_h\epsilon_t \text{ with } \zeta = \begin{pmatrix} \zeta_0 \\ \zeta_1 \\ \vdots \\ \zeta_k \end{pmatrix}, \quad C^{-1}u = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_k \end{pmatrix}$$

(with vectors partitioned in blocks of sizes m_0, m_1, \dots, m_k).

Our goal is to extend the domain of $\zeta_t, t \in \mathbf{Z}$ to all \mathbf{R} , maintaining the stationarity. After obtaining $\zeta_t, t \in \mathbf{R}$, the embedding $x_t = \theta^{\text{tr}} C \zeta_t$ is computed. This new scalar process is stationary and for $t \in \mathbf{Z}$, $x_t = X_t$. The eigenvalues ρ_h of the matrix J are the roots of the polynomial equation $\rho^r = \sum_{j=1}^r \phi_j \rho^{r-j}$ and the sizes m_h of the blocks in J are their respective multiplicities. In other words, the eigenvalues are the inverses of the roots of the polynomial $\phi(z) = 1 - \sum_{j=1}^p \phi_j z^j$ associated to the AR-coefficients of the ARMA(p, q) model, with their algebraic multiplicities, and also $\rho_0 = 0$ with multiplicity $r - p$ that may vanish.

The geometrical multiplicity of each ρ_h is one, and the matrix C is partitioned in $k+1$ blocks $C^{(h)}$ of $r \times m_h$ with possible values (the solution is not unique)

$$C^{(h)} = \left(c_{i,j}^{(h)} \right)_{\substack{i=1,2,\dots,r \\ j=1,2,\dots,m_h}}, \quad c_{i,j}^{(h)} = \sum_{n=0}^{m_h-j} \binom{r-i}{n} \rho_h^n$$

in correspondence with each of the blocks J_{ρ_h} , $h = 0, 1, \dots, k$, that satisfy the conditions $DC^{(h)} = C^{(h)}J_{\rho_h}$.

2.3 Solving one canonical equation

Find the \mathbf{R}^m -valued function $L(t), t \in \mathbf{R}^+$ such that $\zeta_t = \int_{-\infty}^t L(t-s) dA_s$ is a solution of

$$\zeta_t = J_{\rho}\zeta_{t-1} + c\epsilon_t, \quad J_{\rho} = \rho I + I_1$$

Let $SL(t) = L(t+1)$. Then

$$\begin{aligned} \int_{-\infty}^t L(t-s) dA_s &= \int_{-\infty}^{t-1} SL(t-1-s) dA_s + \int_{t-1}^t L(t-s) dA_s \\ &= J_{\rho} \int_{-\infty}^{t-1} L(t-1-s) dA_s + c\epsilon_t \end{aligned}$$

is satisfied when $SL = J_\rho L$, $\epsilon_t = \int_0^1 l(1-s)d\Lambda_{s+t-1}$ and for $t \in [0, 1)$, $L(t) = cl(t)$.

For $n = 0, 1, 2, \dots$ and $t \in [0, 1)$, $L(n+t) = SL(n-1+t) = J_\rho L(n-1+t) = \dots = J_\rho^n L(t) = J_\rho^n cl(t)$, and therefore

$$L(t) = J_\rho^{[t]} cl(\text{frac}(t))$$

with $[t]$ denoting the integral part of t and $\text{frac}(t) = t - [t]$ its fractional part.

Because $I_{1,m}^m = 0$, then $J_\rho^n = (\rho I_m + I_{1,m})^n = \sum_{j=0}^{n \wedge (m-1)} \binom{n}{j} \rho^{n-j} I_{1,m}^j$ and hence

$$L(t) = \sum_{j=0}^{[t] \wedge (m-1)} \binom{[t]}{j} \rho^{[t]-j} I_{1,m}^j cl(\text{frac}(t)).$$

In particular, if $l(t)$ is constant equal to one, then the i -th component of $L(t)$ is

$$L_i(t) = \sum_{j=0}^{[t] \wedge (i-1)} \binom{[t]}{j} \rho^{[t]-j} c_{i-j}.$$

The particular case $\rho = 0$ that necessarily applies when $q \geq p$, leads to the simpler expressions

$$L(t) = \mathbf{1}_{\{t < m\}} I_1^{[t]} cl(\text{frac}(t)).$$

for the vector L and, if $l(t)$ is constant equal one, the i -th component of $L(t)$ is

$$L_i(t) = \mathbf{1}_{\{t < i\}} c_{i-[t]}.$$

2.4 Joining the solutions corresponding to each Jordan block and retracing steps

The continuous parameter embedding of (4), that we denote by the same symbol ζ_t is composed by the juxtaposition of the processes

$$\zeta_{\rho_h, t} = \int_{-\infty}^t L_{\rho_h}(t-s) d\Lambda(s)$$

so that, for each selection of the function l , $L_{\rho_h}(t) = J_{\rho_h}^{[t]} c_{\rho_h} l(\text{frac}(t))$, and the matching Lévy process Λ , the continuous parameter stationary process $x_t = \theta^{\text{tr}} C \zeta_t$ is an embedding for X_t .

Now we proceed to show how to choose the Lévy integrator Λ and the function l .

3 Choosing the Lévy integrator Λ

Assume without loss of generality that $\mathbf{Var}\epsilon_t = \mathbf{Var}\Lambda_1 = \int_0^1 l^2(s)ds = 1$.

A necessary and sufficient condition to express the noise as an integral

$$\epsilon_t = \int_0^1 d\Lambda_{s+t-1}$$

is that the common law of $(\epsilon_t)_{t \in \mathbf{Z}}$ be infinitely divisible and, in that case, $\Lambda_1 = \epsilon_1$. Integrands $l \neq 1$ limit the family of noises that admit the representation

$$\epsilon_t = \int_0^1 l(1-s)d\Lambda_{s+t-1}.$$

On the other hand, Gaussian noises can be represented with any integrand l and Λ a Wiener process.

4 Choosing the integrand $l(t), t \in [0, 1]$

The optimisation criteria applied to select the integrand L refer to the covariances of the resulting continuous embedding

$$x_t = \boldsymbol{\theta}^{\text{tr}} \sum_{h=0}^k C_h \int_{-\infty}^t J_{\rho_h}^{[t-s]} c_h l_h(\text{frac}(t-s)) d\Lambda_s \quad (5)$$

A detailed computation of these covariances shows that $\mathbf{Cov}(x_t, x_0)$ is a product of a function of the parameters of the DARMA, independent of the restriction l of L to the domain $[0, 1]$, times a factor that depends exclusively of l .

5 Particular case $r = 1$ (one-dimensional AR(1))

We proceed to show how to choose the integrand $l(s), 0 \leq s < 1$, in the particular case $r = 1$. In fact, it is not hard to show that the results for dimension $r = 1$ of the state space corresponding to DARMA(1) extend to the general case.

Let us extend the domains \mathbf{R}^+ and $[0, 1]$ of the functions L and l by defining $L(s) = 0$ for $s < 0$, and $l(s) = 0$ for $s \notin [0, 1]$, employ the inner product notation $\langle f, g \rangle = \int_{-\infty}^{\infty} f(s)\bar{g}(s)ds$, $\|f\|^2 = \langle f, f \rangle$ and $\langle f, g \rangle_1 = \int_0^1 f(s)\bar{g}(s)ds$, $\|f\|_1^2 = \langle f, f \rangle_1$ for complex functions with domain \mathbf{R} or $[0, 1]$ and generalise the definition of the shift operator $\mathcal{S}f(s) = f(s+1)$ by introducing $\mathcal{S}^t f(s) = f(s+t)$.

Then the covariance of the process $\zeta_t = \int_{-\infty}^t L(t-s)d\Lambda_s$ is

$$\gamma_t = \langle \mathcal{S}^t L, L \rangle = \frac{1}{1 - |\rho|^2} \langle \mathcal{S}^t L, L \rangle_1 = \frac{\langle \mathcal{S}^t l, l \rangle + \rho \langle \mathcal{S}^{t-1} l, l \rangle}{1 - |\rho|^2}$$

Let \mathcal{F}_t denote the σ -field generated by $(\Lambda_s)_{s \leq t}$. The conditional expectation $\hat{\zeta}_t = \mathbf{E}(\zeta_t | \mathcal{F}_0) = \int_{-\infty}^0 L(t-s) d\Lambda_s$ is the \mathcal{F}_0 -measurable estimator of ζ_t with minimum error variance.

On the other hand, $\tilde{\zeta}_t = \gamma_t \zeta_0 / \gamma_0$ is the linear predictor of ζ_t given ζ_0 with minimum error variance.

Then, since $\tilde{\zeta}_t$ is \mathcal{F}_0 -measurable, the inequality

$$\mathbf{Var}(\zeta_t - \tilde{\zeta}_t) \geq \mathbf{Var}(\zeta_t - \hat{\zeta}_t)$$

must hold and also hold the equalities

$$\mathbf{Var}\zeta_t = \mathbf{Var}\hat{\zeta}_t + \mathbf{Var}(\zeta_t - \hat{\zeta}_t) = \mathbf{Var}\tilde{\zeta}_t + \mathbf{Var}(\zeta_t - \tilde{\zeta}_t)$$

because both decompositions of ζ_t as sum of the estimator and the error are orthogonal. Consequently

$$\mathbf{Var}\tilde{\zeta}_t \leq \mathbf{Var}\hat{\zeta}_t$$

Now compute

$$\begin{aligned} \mathbf{Var}(\hat{\zeta}_t - \tilde{\zeta}_t) &= \mathbf{Var} \int_{-\infty}^0 \left(L(t-s) - \frac{\gamma_t}{\gamma_0} L(-s) \right) d\Lambda_s \\ &= \frac{\|\mathcal{S}^t L - \frac{\gamma_t}{\gamma_0} L\|_1^2}{1 - |\rho|^2} = \frac{\|\mathcal{S}^t L\|_1^2 - \langle \mathcal{S}^t L, L \rangle_1^2}{1 - |\rho|^2}, \end{aligned}$$

because of the assumption $\|L\|_1^2 = \|l\|^2 = 1$.

This result attains its minimum value 0 when $\mathcal{S}^t L$ is proportional to L . This proportionality is achieved simultaneously for all t when $l(s) = \rho^s$, since this implies $\mathcal{S}^t L = \rho^t L$ and therefore

$$l(s) = \rho^s / \|\rho\| \text{ implies } \hat{\zeta}_t = \tilde{\zeta}_t$$

and also implies that ζ is Markovian, since the conditional law of $\zeta_t = \rho^t \zeta_0 + \int_0^t \rho^{t-s} d\Lambda_s$ given ζ_0 is the sum of a function of ζ_0 plus a term independent of \mathcal{F}_0 .

Comments about the selection $l(s) = \rho^s$:

1. The resulting

$$\zeta_t = \int_{-\infty}^t \rho^{t-s} d\Lambda_s = \int_{-\infty}^t e^{-\kappa(t-s)} d\Lambda_s$$

with $\kappa = -\log \rho$ is an Ornstein - Uhlenbeck process, with the Markov property.

2. Negative values of ρ lead to complex processes, and there is no solution for $\rho = 0$.
3. The admissible probability law of the noise ϵ_t is limited to the ones that can be represented as integrals $\int_0^1 \rho^{1-s} d\Lambda_s$.

5.1 Maximising the integrated covariance

Another possibility for choosing l is maximising the integrated correlation of the process, leading to the following results:

Theorem 2. *The embedding ζ_t for which the integrated correlation*

$$\int_{-\infty}^{\infty} \gamma_t = 2\Re \int_0^{\infty} \gamma_t dt = \frac{2}{1-|\rho|^2} \Re \int_0^1 \langle S^t L, L \rangle_1 dt,$$

of the process ζ_t is maximum and the integrated variance $\int_0^1 \mathbf{Var}(\zeta_t - \zeta_0) dt$ is minimum is obtained with $\ell(s) = 1$ for all $s \in [0, 1)$.

As for the covariances of the differences between the embedding and the polygonal interpolation of the original DARMA when $l(s) = 1, 0 \leq s < 1$, we have the following,

Theorem 3. 1. *For each integer n , denote*

$$S_{n,t} = \zeta_{n+t} - (1-t)\zeta_n - t\zeta_{n+1}, \quad 0 \leq t \leq 1$$

the difference between the stationary interpolation ζ_{n+t} and the segment joining (n, ζ_n) and $(n+1, \zeta_{n+1})$. Then

$$\mathbf{E}S_{n,s}S_{n,t} = \frac{2(1-\Re(\rho))}{1-|\rho|^2} (s \wedge t)(1 - (s \vee t)).$$

2. *For any integer m , $S_{n,t}$, $0 \leq t \leq 1$ and ζ_m are not correlated.*
3. *For integers $m < n$, the covariances between $S_{m,s}$ and $S_{n,t}$ are*

$$\mathbf{E}S_{m,s}S_{n,t} = \rho^{n-m} \frac{2(1-\Re(\rho))}{1-|\rho|^2} (s \wedge t)(1 - (s \vee t)).$$

Corollary 1. *As a particular case, consider a Gaussian AR(1) interpolated with $l = 1$, Assume*

1. *The stationary series $X_t, t \in \mathbf{Z}$, satisfies the AR(1) model $X_n = \rho X_{n-1} + \epsilon_n$ where ϵ_n is a Gaussian white noise,*
2. *The stationary sequence of processes B_n with domain $[0, 1]$, satisfies the model $B_n = \rho B_{n-1} + \beta_n$ where β_n is a sequence of Brownian bridges independent of the noise ϵ .*

Then the process $\zeta_t = X_{[t]} + (t - [t])(X_{[t]+1} - X_{[t]}) + B_{[t]}(t - [t])$, $t \in \mathbf{R}$ is a stationary interpolation of $X_n, n \in \mathbf{Z}$.

Acknowledgments. A. Arratia acknowledges support of MINECO project AP-COM (TIN2014-57226-P), and Gen. Cat. SGR2014-890 (MACDA). A. Cabaña acknowledges support of MINECo project MTM2015-69493-R

References

1. A. Arratia, A. Cabaña, E.M. Cabaña (2016) A construction of Continuous time ARMA models by iterations of Ornstein-Uhlenbeck processes, *SORT* **40**, 267–302.
2. A. Arratia, A. Cabaña, E.M. Cabaña (2017) Embedding in law of discrete time ARMA processes in continuous time stationary processes (*in preparation*).
3. Brockwell, P. J. (1995) A note on the embedding of discrete-time ARMA processes. *J. Time Ser. Anal.* **16**, 451–460.
4. Brockwell, P. J. (2004) Representations of continuous time ARMA processes. *J. Appl. Probab.* **41**, 375–382.
5. Brockwell, A. E. and Brockwell, P. J. (1999) A class of non-embeddable ARMA processes. *J. Time Ser. Anal.* **20**, 5, 483–486.
6. Brockwell, P. J. and Davis, R. A. *Introduction to Time Series and Forecasting*, Springer (2nd. Ed.), 2002.
7. Brockwell, P. J. and Hannig, J. (2010) CARMA(p, q) generalized random processes. *J. Statistical Planning and Inference* **140**, 3613–3618.
8. Chan, K. S. and Tong, H. (1987) A note on embedding a discrete parameter ARMA model in a continuous parameter ARMA model. *J. Time Ser. Anal.* **8**, 277–281.
9. Gel'fand, I. M. and Vilenkin, N. Ya. (1964) *Generalized Functions*, Academic Press, New York.
10. He, S. W. and Wang, J. G. (1989) On embedding a discrete-parameter ARMA model in a continuous-parameter ARMA model. *J. Time Ser. Anal.* **10**, 315–323.
11. Huzii, M. (2006) Embedding a Gaussian Discrete-time Autoregressive Moving Average process in a Gaussian Continuous time Autoregressive Moving Average process. *J. Time Ser. Anal.* **28**(4): 498–520.
12. Thornton, M. A. and Chambers, M. J. (2013) Continuous-time autoregressive moving average processes in discrete time: representation and embeddability. *J. Time Ser. Anal.* **34**, 552–561.

Analysis of time series of earthquake occurrence in Caucasus

T. Matcharashvili, N. Zhukova, E. Mepharidze, A. Sborshikov

M. Nodia Institute of Geophysics, 1 Alexidze str. Tbilisi, Georgia

matcharashvili@gtu.ge

Abstract. In this work the temporal features of earthquake time distribution in Caucasus is investigated. Several methods (power spectrum, wavelet and Hilbert-Huang transformation) are applied to earthquake time series. Our findings show that earthquakes hourly and daily occurrence is not characterized by the dominant frequencies.

It was shown that the variation of the power of cyclic components in the temporal features of earthquakes occurrence is not uniform, but their amplification corresponds to the decrease of released local seismic energy. Temporal distribution of the power of weak cyclic oscillatory modes is not uniform and varies significantly during certain periods.

Keywords: Seismicity, time series, time frequency analysis.

Introduction

Investigation of temporal features of earthquake occurrence remain among the most important scientific and practical tasks. It can be listed number of high level contemporary studies based on different conceptual frameworks which aimed at investigating earthquake temporal patterns using both field and laboratory data as well as numerical simulations [see e.g. 1, 2, 3, 4, etc.].

Most of such analyses agree that earthquake time dynamics is characterized by switching or intermittent behavior with periods of intense seismic activity interspersed with those of low seismicity. The details of such transition from one state (high seismic activity) to the other (low seismic activity) are still unclear. At the same time it is reasonable to presume that temporal variation of seismic processes should be caused by stress changes in the Earth's crust, which can be dynamically different and of both tectonic and non-tectonic origin [1, 5]. As a consequence, the question of earthquakes' temporal distribution is still an open problem. Nowadays, in scientific literature, it can be found controversial views on this question from earthquakes regular to completely random distribution [5].

Data and methods of analysis

In the present work, we aimed to continue investigation of seismic process in Caucasus on the presence of regular or irregular dynamical behaviors in the earthquake generation. For this purpose from the Caucasian earthquake catalogue spanning from 1970 to 2016 we compiled time series of frequency of earthquake occurrence (FEO). Exactly, from Caucasian earthquakes catalogue (further details about catalogue and study area can be found elsewhere [4]) we calculated number of earthquakes occurred in consecutive hours of observational period and divided them by the total number of yearly occurred events and normalized to zero mean and unit variance. FEO, as distinct from often used inter earthquakes time sequences, are evenly sampled time series enabling the correct using of different methods of frequency and time frequency analysis. In present research in order to investigate temporal characteristics of FEO time series we have used power spectrum calculation, wavelet analysis and Hilbert-Huang Transform (HHT) (for details see [5]). These three methods when used together enable to avoid restrictions typical for each of them separately, such as influence of non-stationarity, time-frequency uncertainty, etc. Next, in order to assess the robustness of the obtained results against the influence of possible noise, we filtered our FEO time series by using two different de-noising techniques, the Savitzky-Golay filtering and the Singular Spectrum Analysis (SSA) decomposition.

Results and discussion

The power spectrum of the original FEO series does not revealed prevalent cyclic components in the analyzed data obtained from declustered Caucasian earthquake catalogue. It is flat in the higher frequency range, and looks like the power spectrum of randomized by shuffling procedure FEO time series. At the same time, continuous wavelet transform shows lesser homogeneity in spectrum comparing to randomized FEO sequences. Difference from random processes is more visible when we analyzed smoothed by Savitzky-Golay filter and SSA decomposition FEO data time series. Randomness in the earthquakes time distribution makes even more questionable results of the HHT analysis of the hourly FEO time series indicating clear changes in Hilbert Energy Spectrum at different level of local seismic activity.

Based on the results of our analysis we concluded that the time series of frequency of earthquake occurrence does not reveal presence of leading cycles. At the same time the temporal distribution of the power of weak cyclic oscillatory modes is not uniform and varies significantly during certain periods. Our analysis indicates that the increase in the extent of regularity in FEO data sets is closely related with the amount of released local seismic energy.

Summary

We investigated features of time distribution of earthquakes in Caucasus. Methods of frequency and time frequency analysis as well as Hilbert-Huang Transform transformation have been used. It was shown that variation of cyclic components is not uniform and depends on the amount of released seismic energy.

Acknowledgements. The present study was supported by Shota Rustaveli National Foundation, grant "Investigation of dynamics of earthquake's temporal distribution" 217838.

References

1. Rundle, J., Turcotte, D., and Klein, W., GeoComplexity and the physics of earthquakes. AGU, Washington (2000).
2. Matcharashvili, T., Chelidze, T., Javakhishvili, Z., Nonlinear analysis of magnitude and interevent time interval sequences for earthquakes of Caucasian region. *Nonlinear Processes in Geophysics*, 7, 9-19(2000).
3. Ben-Zion, Y., Lyakhovsky, V., Accelerated Seismic Release and Related Aspects of Seismicity Patterns on Earthquake Faults. *Pure appl. geophys.*, 159, 2385–2412, (2002)
4. Telesca, L., Matcharashvili, T., Chelidze, T., Investigation of the temporal fluctuations of the 1960–2010 seismicity of Caucasus. *Nat. Hazards Earth Syst. Sci.*, 12, 1905–1909, (2012).
5. Matcharashvili, T., Telesca, L., Chelidze, T., Z. Javakhishvili, Zhukova, N. Analysis of temporal variation of earthquake occurrences in Caucasus from 1960 to 2011, *Tectonophysics*, 608, 857-865, (2013).

Author Index

Ait Hassou, Laila	1146
Akkaya, Aysen Dener	1175
Aknin, Noura	181
Alerini, Julien	343
Aljawazneh, Huthaifa	661
Alkhatib, Hamza	23, 1132
Amar, Amine	1146
Anh, Duong Tuan	355
Antoniou, Ioannis	129
Aoulad Abdelouarit, Karim	181
Ariza Villaverde, Ana Belén	114
Arkipov, Kirill	173
Arkipova, Marina	173
Arratia, Argimiro	585, 1231
Asencio-Cortés, Gualberto	786
Attoue, Nivine	939
Badaoui, Fadoua	1146
Baek, Thomas	11
Balodis, Janis	122
Baratashvili, Evgeni	452
Barba, Lida	922
Barbeito Cal, Inés	1089
Bardet, Jean-Marc	512
Batisani, Nnyaladzi	463
Bauer, André	444
Bayrak, Özlem Turker	1175
Beard, Joshua S.	400
Belashova, Inna	422
Benhmad, Francois	774
Benyacoub, Badreddine	877
Bhattacharyya, Malay	829
Bijleveld, Frits	59
Bochkarev, Vladimir	422
Bokde, Neeraj	786
Bondon, Pascal	47, 249, 319, 1073
Boteler, David	141
Bratsas, Charalampos	129
Bravo Caro, José Manuel	1108
Bringas, Carlos	195
Brink, Willie	865
Bueno Lopez, Maximiliano	474
Buniyamin, Norlida	432

Cabaña, Alejandra	1231
Cabaña, Enrique	1231
Cadahia, Pedro	1108
Cao, Ricardo	1089
Cardot, Hubert	740
Carità, Danilo	639
Caro, Eduardo	960
Castillo Valdivieso, Pedro	661
Cernuda, Paula	960
Chakrabarti, Amlan	548
Chakraborty, Basabi	536, 548
Chakraborty, Goutam	563, 597
Chau, Vo Thi Ngoc	355
Chelidze, Tamaz	452
Chen, Jingjie	719
Chinellato, Eris	379
Chou, Ray Yeutien	1047
Christidis, Panayotis	807
Claeys, Peter	1064
Clauter, Dean A.	400
Cohn, Anthony G.	379
Commandeur, Jacques J.F.	59
Congacha, Ana	922
Cosovic, Marijana	487
Cotta, Higor	47, 1073
Crone, Nathan	1075
Curtef, Valentin	444
Czechowski, Zbigniew	1089
de Franciscis, Sebastiano	1157
De Luca, Giovanni	639
de Oliveira, Manuela	609
de Souza, Juliana Bottoni	319
Delgado, Jorge	922
di Iorio, Francesca	686
Diependaele, Kevin	59
Dimitriou, Simone	76
Divins, Shaelyn G.	400
Dlask, Martin	316
Dmitriev, Andrey V.	1219
Dmitriev, Victor	1219
Durán Domínguez, Arturo	730
Ergun, Salih	321
Espinoza, Lady	922
Ezzahid, Elhadj	1146

Farn, Matteo	1050
Faure, Cynthia	512
Fernández González, Paula	752
Ferreira, Nuno	609
Fetisova, Nadezhda	1163
Finkenstadt, Barbel	899
Franaszczuk, Piotr	1075
Franco, Glaura C.	319
Fuertes, Walter	207
Fujimura, Shigeru	499
Gallo, Giampiero M.	639
García-Díaz, Juan Carlos	91
Garcia-Hiernaut, Alfredo	696
Garrido Haba, Rafael	1157
Gegundez-Arias, Manuel Emilio	1108
Gil, Antonio J.	146
Golpe, Antonio	1108
Gómez López, Juan María	114
Gómez-Losada, Álvaro	807
Gomez-Pulido, Juan A.	730
Gonçalves, Jorge	1206
Gonzalez Vasco, Maria Del Camino	616
Goswami, Saptarsi	548
Gunaratne, Gemunu	1206
Gutiérrez de Ravé Agüera, Eduardo	114
Gutierrez, German	820
Haber, Rana	400, 1120
Hammersland, Roger	1007
Hanel, Martin	162
Haritonova, Diana	122
Harker, Matthew	155
Hashimoto, Takako	575
Haßler, Marc	995
Haykal, Vanessa	740
Heggart, Callum	141
Herbst, Nikolas	444
Hoang, Dau Xuan	305
Hogg, David	379
Hua, Jia-Chen	1206
Huusko, Reino	795
Isogai, Keisuke	266
Ispány, Márton	319
Jalan, Arun	548
Janiashvili, Manana	452

Jeon, Yongdeok	390
Jeong, Heeyoung	390
Jeong, Kichang	1, 390
Jerez, Miguel	696
Jeschke, Sabina	995
Jiménez Hornero, Francisco José	114
Jowaheer, Vandna	985
Juan, Jesús	960
Junuz, Emina	487
Kamiyama, Takuya	563
Kara-Terki, Nesrine	1083
Kargoll, Boris	23, 1132
Kato, Taichi	254
Kawamura, Satoshi	331
Kehagias, Dionysios	129
Khettab, Zahira	367
Kielbik, Andrzej	764
Kim, Woo-Ram	1, 390
Kimura, Mariko	254
Kinnear, Ryan J.	841
Kinoshita, Tetsuo	563
Klemashev, Nikolay	117
Kon, Yukito	102
Konen, Wolfgang	11
Koponen, Pekka	795
Korzeniewska, Anna	1075
Kott, Marek	295
Kounev, Samuel	444
Krzemień, Alicia	649
Kubek, Daniel	232
Kukal, Jaromir	316
Kulat, Kishore	786
Kulia, Geir	474
Kumps, Diederik	1064
Lacaille, Jérôme	512
Landajo, Manuel	752
Le, Thi Ngc Anh	305
Lee, Jaeseob	1
Lee, Jae-Seob	390
Leong, Philip H.W.	1206
Lerner, Markus	59
Leulmi, Sara	1098
Lin, Amanda Yan	411
Liu, Lichun	964
Lukoseviciute, Kristina	974

Machado, Guillermo	922
Machete, Reason L.	463
Maciak, Matus	524
Maciejewska, Monika	764
Malaina, Iker	195
Maltseva, Svetlana	1219
Mamode Khan, Naushad	985
Manabe, Yusuke	503
Mandrikova, Oksana	1163, 1185
Mardia, Kanti	379
Martensen, Heike	59
Martínez de La Fuente, Ildefonso	195
Martinez, Luis	195
Martínez-álvarez, Francisco	786
Mat, Usamah	432
Matcharashvili, Tamar	452
Matcharashvili, Teimuraz	452, 1240
Mato, Fernando	207
Mazumdar, Ravi R.	841
Meisen, Tobias	995
Messaci, Fatiha	1098
Mestre, Roman	1187
Mijatovic, Nenad	400, 1120
Mitra, Arup	548
Molina, Lorena	922
Molinas, Marta	474
Montagnon, Chris	913
Montanari, Angela	1050
Moon, Hyosoo	809
Mora García, Antonio	661
Moravec, Vojtěch	162
Mourid, Tahar	467, 1083
Murakami, Takeshi	102, 331
Navickas, Zenonas	974
Nielsen, Mikkel Slot	707
Nikitina, Lidia	141
Nikolov, Ventsislav	1054
Niska, Harri	795
Noorian, Farzad	1206
Obradovic, Slobodan	487
Okou Guei, Cyrille	1146
O'Leary, Paul	155
Olteanu, Madalina	270, 343, 512
Omidalizarandi, Mohammad	1132
Osowski, Stanislaw	220

Paffenholz, Jens-André	23
Pandey, Manish Kumar	885
Park, Jinhong	390
Park, Moonseo	809
Pascual Granado, Javier	1157
Pavlidis, Efthymios	946
Pavón Domínguez, Pablo	114
Percebois, Jacques	774
Pérez Lopez, Cesar	616
Pérez Salinas, Juan Carlos	207
Pérez-Samartín, Alberto	195
Pesta, Michal	674
Pestova, Barbora	674
Peter, Adrian M.	400, 1120
Pino Angulo, Adrian	575
Politis, Dimitris	1089
Polozov, Yury	1163
Presno, María José	752
Ragot, Nicolas	740
Ragulskis, Minvydas	974
Randon-Furling, Julien	270
Rangarajan, Anand	1120
Ranjan, Ravi Prakash	829
Reisen, Valdério	47, 1073
Reisen, Valderio A.	319
Renedo, Martí	585
Ridall, Peter	282
Riesgo García, Maria Victoria	649
Ritt, Roland	155
Rodríguez Lozano, David	730
Rohde, Victor	707
Rothschedl, Christopher Josef	155
Ruíz-Cañadas, Carlos	74
Saastamoinen, Kalle	1035
Salamanis, Athanasios	129
Sanchez Lasheras, Fernando	649
Sanchis, Araceli	820
Santos, Jane Meri	319
Sasaki, Masatomo	331
Sbihi, Boubker	181
Scaglione, Miriam	76
Schepers, Andreas	59
Schuh, Wolf-Dieter	1132
Selpi, Selpi	411
Sesmero Lorente, M. Paz	820
Shahrour, Isam	939

Shananin, Alexander	117
Sheraz, Muhammad	37
Shigeki, Horie	597
Shimono, Shogo	331
Shin, Kilho	575
Silchev, Vitaly	1219
Sirotn, Viacheslav	173
Siwek, Krzysztof	220
Smit, Francois	865
Smith, Anthony O.	400, 1120
Smith, Leonard A.	463
Song, Wen	499
Steyn, Melise	865
Strnad, Filip	162
Stummer, Wolfgang	47
Suárez Gómez, Sergio Luis	649
Suárez Sánchez, Ana	649
Suárez, Juan Carlos	1157
Subbiah, Karthikeyan	885
Sunecher, Yuvraj	985
Suzuki, Norikazu	192
Szczurek, Andrzej	764
Tak, Hyungsuk	254
Takahashi, Hideyuki	563
Tapia, Santiago	207
Taş, Büşra	246
Telksnys, Tadas	974
Terraza, Michel	1187
Thatcher, Marcus	853
Thibault, Aymeric	249
Thill, Markus	11
Thuy, Huynh Thi Thu	355
Toulkeridis, Theofilos	207
Toulkeridis-Estrella, Katerina	207
Trapero, Juan R.	74
Triacca, Umberto	686
Trichtchenko, Larisa	141
Troncoso, Alicia	786
Trull, Oscar	91
Tzovaras, Dimitrios	129
Varna, Inese	122
Vizina, Adam	162
Wang, Weilun	597
Wiecek, Paweł	232
Williams, Trefor P.	809

Wilms, Josefine	853, 865
Xanthopoulou, Georgia	129
Yen, Tso-Jung	1047
Yen, Yu-Min	1047
Yoshida, Hitoaki	102, 331
Yoshida, Sho	536
Yozgatligil, Ceylan	246
Yue, Yading	964
Yusupova, Alisa	946
Zalyaev, Timur	1185
Zhang, Mengcheng	411
Zhang, Rong	964
Zhang, Yongping	719
Zhao, Jianchun	964
Zhuang, Guangan	964
Zhukova, Natalia	452
Zoglat, Abdelhak	1146
Zufle, Marwin	444