The Early Booking Effect and Other Determinants of Hotel Room Prices in Europe

Anastasia Bezzubtseva
National Research University
Higher School of Economics
Pokrovsky boulevard, 11
109028 Moscow, Russia
Email: nstbezz@gmail.com

Dmitry I. Ignatov
National Research University
Higher School of Economics
Pokrovsky boulevard, 11
109028 Moscow, Russia
Email: dignatov@hse.ru

***Abstract*—This paper presents a preliminary analysis of hotel room prices in several European cities based on the data from Booking.com website. The main question raised in the study is whether early booking is advantageous indeed, and if so, how early should it be? First a script was developed to download more than 600 thousand hotel offers for reservations from 25 March 2013 to 17 March 2014. Then an attempt to discover more details concerning the early booking effect was made via basic statistics, graphical data representation and hedonic pricing analysis. It was revealed that making reservations in advance can be really gainful, although more data and research are needed to measure the exact numbers, as they depend on at least seasonality and city.**

I. Introduction

One of the most common tactics useful in traveling on a tight budget is booking everything in advance. Flights, train and coach tickets, tours, accommodation – having booked them a couple of months before the journey one might save substantial sums. The tactic is obviously successful for traffic fares, as discounts are often stated by transport companies: e.g., “save up to a dreamy 75% when you book no later than 6.00 pm the day before travel” [13] (naturally, cheap tickets are subject to availability and most likely will not be available the day before travel). But even with transport it is not clear when to book – does one need three weeks in advance to save greatly, or there is no need to sacrifice flexibility, and three days are enough?

Is it true that booking in advance is profitable and necessary in case of accommodation, hotels in particular? And what is “in advance”, i.e. how many weeks before the planned trip are room reservations to be made, so that savings are as high as they can be, whereas the number of weeks is at minimum?

In this paper we will try to gain a basic understanding of those problems via hedonic pricing analysis of data gathered from Booking.com. The hedonic price theory (Lancaster, 1966; Rosen, 1974) [9] [11] states that product price can be considered as a function of its features affecting consumer utility. Therefore room price is an additive function of room size, content (design, view, Internet-connection, household appliances), terms of service (board type, booking rules, check-in and check-out rules), but also of hotel parameters (proximity to the center/airport, neighborhood quality, spa/restaurant/fitness center availability etc.)

The existing studies are consistent with the theory. For example, Thrane (2006) [12] on a small sample of 74 Oslo hotels showed that several factors significantly affect double room rates, among which are hairdryer and minibar presence, free parking, distance to the city center, room service and total hotel bed number. Chen & Rothshild (2010) [3] considered another set of characteristics on 146 Taipei hotels and concluded that the significant ones are hotel location and chain, room area, TV and Internet access in the room, shuttle bus, conference room and fitness center availability. Other research papers found that hotel star rating and climatic features of the area are also important. The main disadvantages of most papers are small sample and narrow scope (one town, one hotel chain or one-two months in case of price dynamic analysis). From the studies we met only Clerides, Nearchow & Pashardes (2003) [4] attempted to compare prices and quality of resorts in several Mediterranean countries.

We will cover several countries, too, via the following European cities: Amsterdam, Brussels, Geneva, London, Madrid, Milan, Munich, Paris, Stockholm, Tallinn and Vienna. We will also examine some hotel and price formation differences in these cities, if they really exist. We do not intend to investigate special cases (particular price policy unusual for most hotels): sales due to hotel management change or director’s daughter birthday, discounts for the elderly or some other customer category etc.; thought they might help one to catch a fantastic offer, they are not part of some trend. The presented analysis is somewhat preliminary; we plan to extend it with hotel clustering, more precise regression analysis or, perhaps, on a new dataset in the future.

II. Data Collection

To answer some of the questions raised earlier we developed a script on Python, which downloads various data from one of the biggest hotel aggregator websites – Booking.com. We were able to scrape about 6 000 hotels out of more than 300 000 registered on the site. On the Download day the robot was requesting rooms available for a one-week stay, where weeks

978-0-7695-5109-8/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDMW.2013.32

IEEE computer society

iterated from 25 March 2013 to 17 March 2014, which resulted in almost 600 000 entries in the final csv-file. Naturally, some hotels and rooms were not available at all in any of those weeks, so the sample is slightly truncated (see Table I). Nevertheless, downloading week prices instead of one-night rates was advantageous for several reasons: we eliminated small price fluctuations caused by weekend surcharges/discounts or secret offers by dealing with average night rates for each week, we simulated a more natural travel situation (usually people stick to one accommodation per city), and we also gained in performance.

TABLE I: The number of prices, hotels and roomtypes

City	Prices (scraped)	Hotels (scraped)	Hotels (overall)	Roomtypes (scraped)
Amsterdam	54 215	634	686	623
Brussels	27 585	308	357	264
Geneva	7 825	78	98	122
London	161 661	1 362	1 429	1 065
Madrid	87 116	700	818	521
Milan	48 900	436	462	313
Munich	24 286	230	335	234
Paris	98 538	1 080	1 847	612
Stockholm	16 595	126	157	205
Tallinn	15 981	178	208	217
Vienna	53 974	514	615	541
Total	596 676	5 646	7 012	4 717

We only used the Booking search result pages for data retrieval, thus losing some important details (room area, Wi-Fi or swimming pool availability, parking options etc.), yet scraping all the basic information much faster. See Table II for the fields we obtained.

The 25 March loading was considered as the basic one for the following description and analysis. We also had another loading in four weeks after the first one (15 April), which was not discussed in that paper except one case in the next section (weekly histogram comparison, Figure 3).

The data scraping script itself is simple to develop when having enough knowledge of some convenient for this purpose language like Python. High changeability of web and Booking.com website code in particular prevents us from sharing our script (it stops working properly and needs updates every week or so as website developers make minor code revisions). Yet we can advise a few things we learned from scraping.

The general procedure is as follows: the script pretends to be a browser controlled by a human, exchanges some data with a website, parses a response page, extracts contents (or attributes) of some tags and saves the resulting table as a comma separated file.

To implement the procedure we intensively used the following Python libraries:

- *mechanize* – interacting with a website, emulating user actions,
- *cookielib* – handling cookies,
- *BeautifulSoup* – processing html-code, navigating tag trees,
- *re* – extracting data via regular expressions (finding patterns in strings),
- *threading* – making simultaneous queries (multiple threads) to scrape faster (waiting for a server response is the most time consuming operation).

TABLE II: Variable explanation

City	hotel location
Hotel	hotel name
Roomtype	roomtype name
Max_per	maximal number of room guests
Price	one-week stay price in euro
Date	week start date
Breakfast	= 1, if breakfast is included
Free_cancel	= 1, if free cancellation is available
Rating	total score based on customer reviews
Num_rev	number or reviews, from 5
Stars	number of stars, from 0 to 5
Lat	geographical latitude
Long	geographical longitude
Dist	distance to the city center in km

There are also other libraries (*lxml*, *urllib2* etc.) and even frameworks (*Scrapy*) which might as well be convenient.

Using proxy, changing user headers and establishing some waiting time between queries can be useful in avoiding ban from the target website. Or one can always explicitly ask a website for a permission to scrape.

III. Sample Properties

We imposed some limitations on the initial dataset to get a clearer picture of price formation.

- We included only double/twin rooms, which make up from 50 to 70 % of the initial sample.
- We excluded data on reservations after 31 December, as the number of available hotels drops dramatically in 2014 (see Figure 1). We believe, the reason of that is neither high demand nor script error, but a particular price-setting policy of most hotels.
- We left only one roomtype per week for each hotel, which was the cheapest of all available ones. That is, we have a single price time series for each hotel, and the type of the room may change in that series, so it reflects the price of staying in a specific hotel, not room. It is remarkable that after the removal of odd roomtypes the dynamics of housing supply almost doesn't change.
- We excluded very expensive rooms (price per night above 500 euro, 0.99 quantile).

The final sample had only 4 605 hotels, or 138 390 observations.

Almost one fifth of the hotels show small price variability (less than 4 euro) during 40 weeks included in the sample. However, more than a third of the hotels have a notable standard deviation of over 25 euro for a one night price, which can be significant for a customer. Besides, low variability might be spurious in series with a large proportion of missing values.

The most popular roomtype name in the sample (more than half of entries) is “double” or “twin” room (sometimes

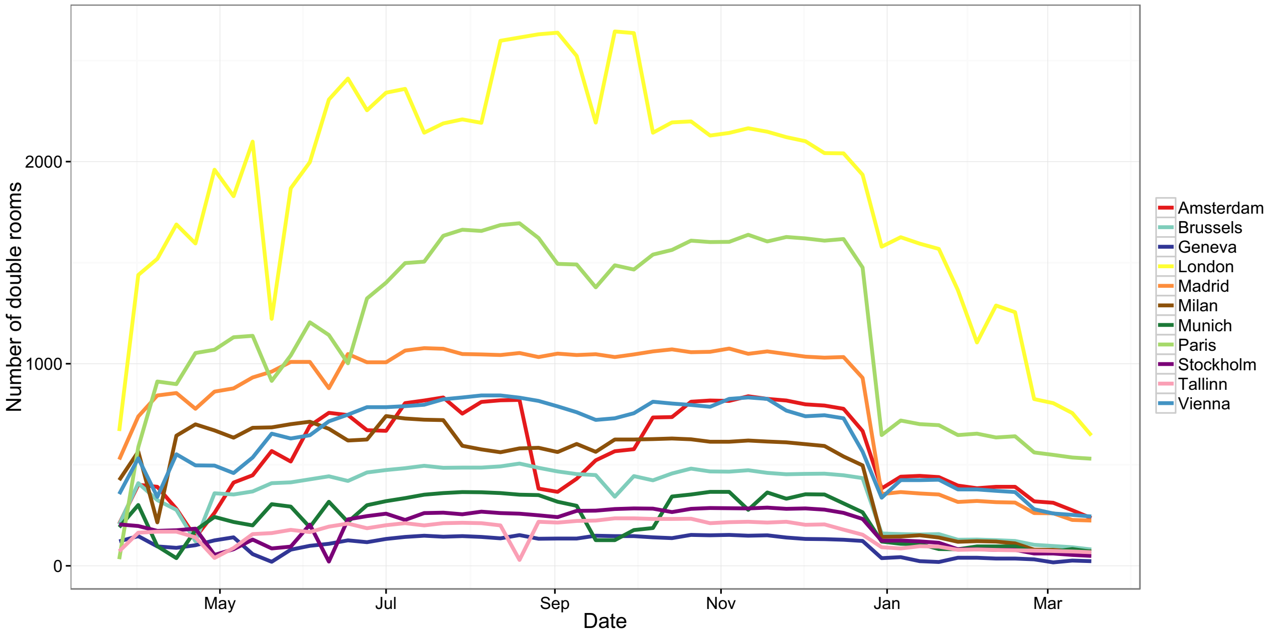

Fig. 1: Number of available rooms

TABLE III: Descriptive statistics

	Mean	Std. Dev.	Min	Median	Max
Price	131	76	20	111	500
Breakfast	0.3	0.5	0	0	1
Free_cancel	0.4	0.5	0	0	1
Rating	7.8	0.8	4.2	7.9	10
Num_rev	357	423	5	227	7 542
Stars	2.7	1.5	0	3	5
Num_rev	3.3	3.5	0	2.3	26

“standard” or “classic”), which agrees with our intention to investigate standard, preferably budget tourist housing opportunities.

Tables III and IV present additional statistics on the numeric variables. Generally the variables are weakly correlated, except for the expected connections between price and star rating, and between price and total score.

The 4 605 series of hotel room prices and mean series by cities are presented on Figure 2. It can be seen from the graphs that there is room shortage and price rise in Brussels and Amsterdam in late April (tulip blossom and Queen’s Birthday in Netherlands occur at the same period), and in Milan in mid-April (which is apparently an impact of the Milan Fashion Week). Early September in Amsterdam is also successful for the hotel industry whereas a world-famous flower parade takes place, as well as October in Munich thanks to Oktoberfest.

If we transform hotel scores into categories (1^{st} category – $Rating$ less than 5, 2^{nd} – $Rating$ from 5 to 6 etc., 6 categories overall) and compare the main loading (25 March) with the additional one (15 April), we can see (Figure 3) that the total histogram area reduction in the second loading was caused mostly by the reduction of the “good” hotel share (“good” hotels are those in categories 4 to 6, with the total score above 7), especially in April, early May, and late June. So, hotel room reservation in advance does seem to be a good idea for money saving.

We can also convert prices into categories by quantiles 0.1, 0.25, 0.5, 0.75 and 0.9, and draw total score distributions by city (Figure 4). Besides general city price levels we may notice on the graphs that the quality of Estonian and Spanish hotels does not suffer from low rates, that it is almost impossible to find a good hotel cheaper than 80 euro per night in Paris and Geneva (and quite hard in London), and also that Sweden stands out with altogether great (yet sometimes expensive) accommodation.

IV. Empirical Results

We will consider the following loglinear hedonic model of price formation in 11 cities:

$$\begin{gathered}\ln Price = \beta_0 + \beta_1\, Breakfast + \beta_2\, Free\_cancel + \\ +\beta_3\, Num\_rev + \beta_{4-8}\, Stars + \beta_9\, Dist + \\ +\beta_{10}\, Lag + \beta_{11-20}\, City + \varepsilon\end{gathered}$$

The total hotel score is composed of customer evaluations by six parameters, one being value for money, which prevents us from including total score in the equation by virtue of possible endogeneity. We can avoid that problem in the future, when we have a more detailed dataset.

In our case some variables cannot be considered as factors affecting the customer utility; e.g., dummy variables for 10 cities are just indicators of city price levels compared to the baseline (price level in Amsterdam). We also included

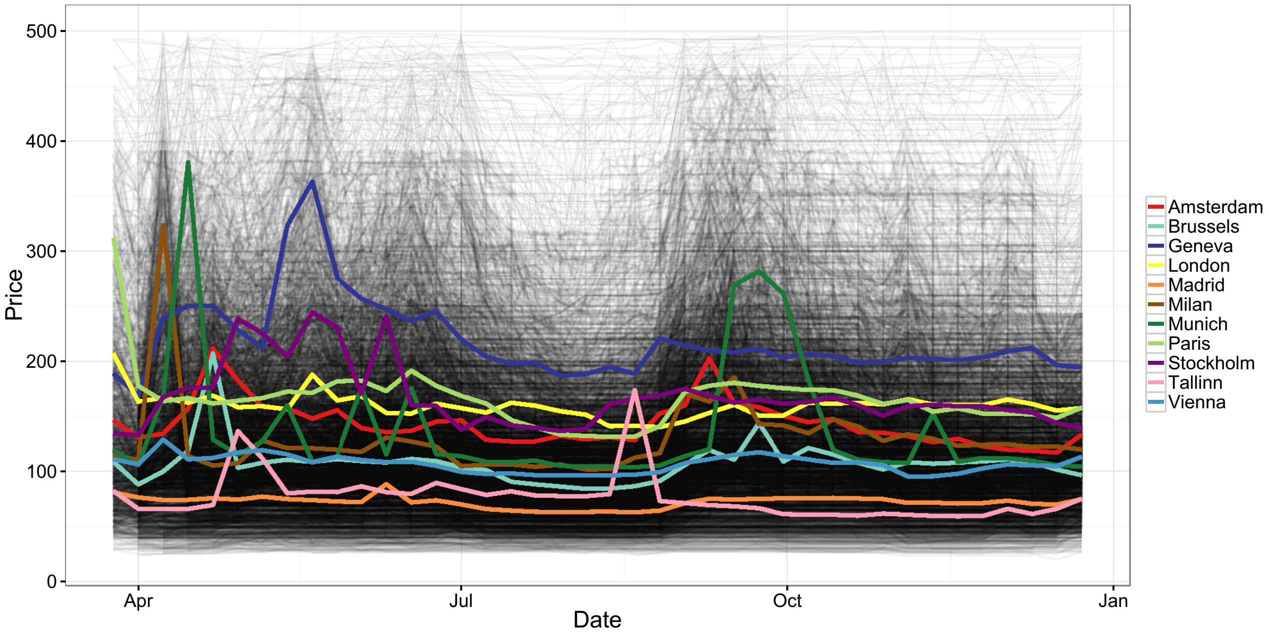

Fig. 2: Room price dynamics

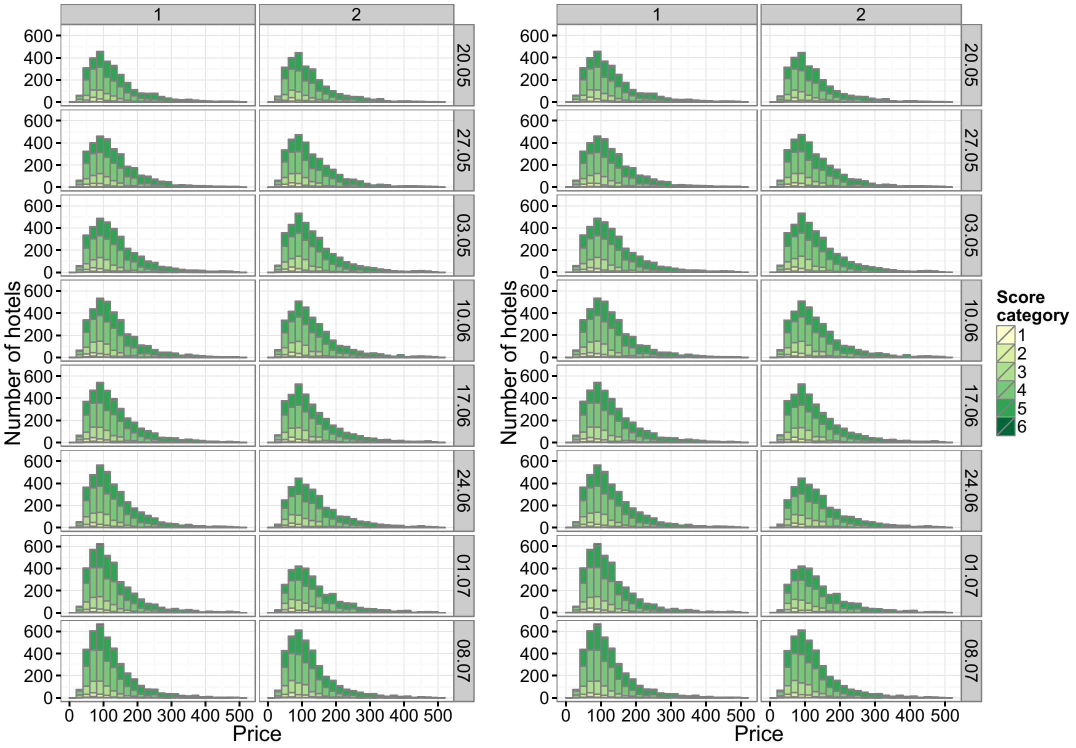

Fig. 3: Hotel room price distributions by week in two loadings. Column names (“1” and “2”) refer to the loadings, row names are week start dates.

TABLE IV: Variable correlations

	Price	Breakfast	Free_cancel	Rating	Num_rev	Stars	Dist
Price	1	−0.08	−0.01	0.33	−0.03	0.46	−0.13
Breakfast		1	0.04	−0.12	0.07	−0.03	0.05
Free_cancel			1	−0.06	−0.09	−0.17	0.03
Rating				1	0.03	0.26	−0.13
Num_rev					1	0.21	−0.07
Stars						1	0.04
Dist							1

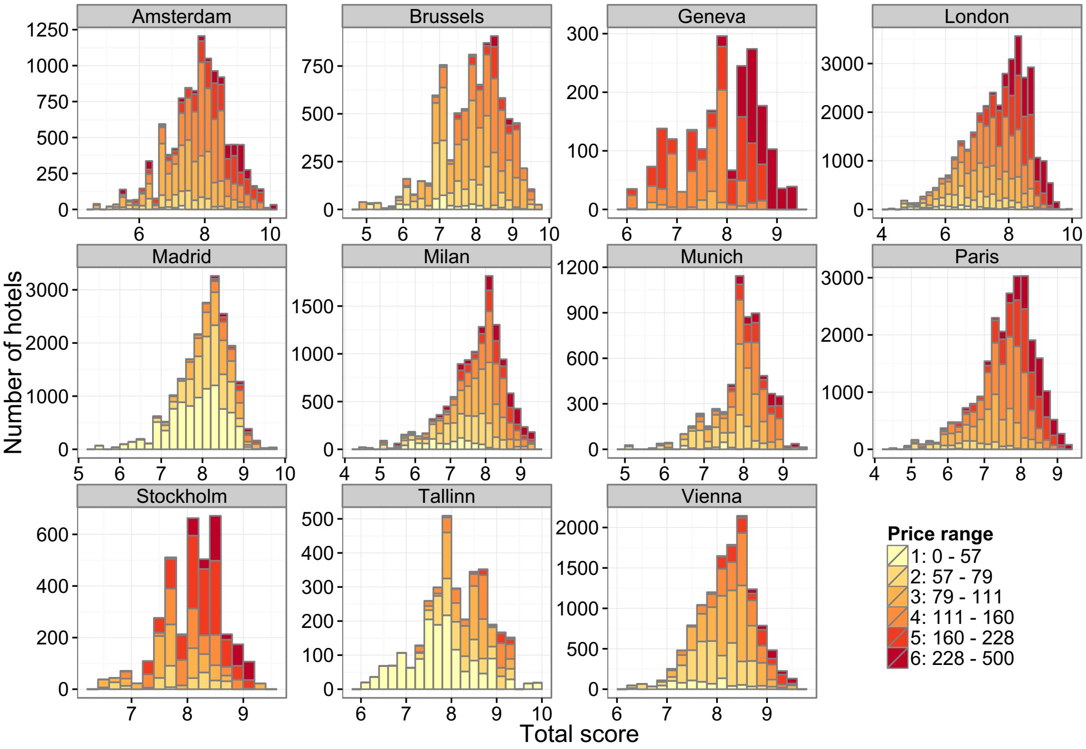

Fig. 4: Hotel total score distributions by city. Price range scales are purposely left unequal.

the Lag variable, which presents a number of days from 25 March to the implied arrival date, to check if early booking is advantageous. Table V shows the results of OLS estimation with robust standard errors (Breusch–Pagan test indicates the presence of heteroscedasticity).

All the coefficients are significant, and there is no multicollinearity in the model (see Table VI, all variance inflation factors are less than 10 or 5 or any other possible multicollinearity threshold valid for the given degrees of freedom), but R^2 is not as high as we would like it to be – presumable, because of lack of some important room characteristics. Price is obviously affected by breakfast inclusion ($+5.5\%$), free booking cancellation option ($+6.3\%$), distance in kilometers to the city center (-3.7%), hotel star rating (from -28% for one-star hotels to $+158\%$ for five-star ones as compared with no-star hotels). Also there is a connection with the hotel review number (100 reviews decrease price by 1%) and the number of days left to check-in date (if one books a room 100 days in advance, the price is 2% lower). The Lag variable coefficient turned out to be much less than we expected based on intuition and graphical evidence. It is probably caused by the actual nonlinearity of Lag contribution to price (for example, at $Lag = 0$ the price might be lower than at $Lag = 7$ both thanks to special offers or discounts on "last chance" reservations and because of a big holiday in two weeks), which most likely is also city-specific.

Let us estimate the same model in linear form for simplicity ($Price$ instead of $\ln Price$ as a dependent variable) for each city individually (see Table VII).

Judging by the vast range of coefficients of determination, not only same factors in different cities affect prices in different ways, but also the set of significant variables changes from town to town. Marginal effects of some factors vary surprisingly. For example, breakfast inclusion in London and Paris hotels reduces price by 7 euro, whereas in Vienna – increases by 22 euro. Tallinn is the only city where guests are

TABLE V: General loglinear model

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	4.836	0.004	0.000
Breakfast1	0.054	0.002	0.000
Free_cancel1	0.062	0.002	0.000
Num_rev	-0.0001	0.000	0.000
Stars1	-0.334	0.005	0.000
Stars2	-0.153	0.003	0.000
Stars3	0.12	0.003	0.000
Stars4	0.456	0.003	0.000
Stars5	0.956	0.005	0.000
Dist	-0.037	0.000	0.000
Lag	-0.0002	0.000	0.000
CityBrussels	-0.332	0.005	0.000
CityVienna	-0.413	0.004	0.000
CityGeneva	0.144	0.006	0.000
CityLondon	0.067	0.004	0.000
CityMadrid	-0.679	0.004	0.000
CityMilan	-0.241	0.005	0.000
CityMunich	-0.213	0.006	0.000
CityParis	0.138	0.004	0.000
CityStockholm	-0.039	0.006	0.000
CityTallinn	-0.797	0.008	0.000
$R^2 = 0.645$			
Num. obs.=138 390			

TABLE VI: VIFs by variables

	VIF	Df
Breakfast	1.31	1
Free_cancel	1.06	1
Num_rev	1.11	1
Stars	1.42	5
Dist	1.31	1
Lag	1.01	1
City	1.92	10

paid 13 euro for free cancellation option, while in Geneva that option is ruinous. It is unclear how to interpret star rating coefficients, considering national differences in star rating policy and various degrees of no-star hotel (the baseline category) prevalence and quality. Obviously, on average more stars means more expensive accommodation. However, it is not unusual for a no-star hotel to be as good as (and as pricey as) three- or four-star ones. We can conclude from the estimation results that this remark is rightful for all cities except Geneva and Stockholm.

Marginal effects of distances are generally adequate: in small towns yet another kilometer is noticeable, in big cities – not really important. Tallinn, Brussels and Milan slightly deviate from this rule, though. The effects of early booking decreased, even became not significant in some cities. It supports the hypothesis of nonlinearity and seasonality of price dynamics in all cities combined with a barely noticeable (also nonlinear) downward price tendency, which can be more clearly seen on the whole sample in the loglinear model. Adding high season indicators as dummies would be a proper way to solve the problem in the future.

Thus we have found some proof of differences in price formation mechanisms in different cities, and ascertained that

TABLE VII: Linear models by cities

	Amsterdam	Brussels	Geneva	London	Madrid	Milan	Munich	Paris	Stockholm	Tallinn	Vienna
(Intercept)	171.515*	98.181*	85.079*	146.954*	61.840*	141.707*	123.853*	149.358*	100.818*	80.380*	97.762*
Breakfast1	1.617	11.819*	4.943	−6.864*	2.354*	8.290*	8.200*	−6.929*	15.771*	10.338*	22.114*
Free_cancel1	7.188*	1.520*	30.780*	10.312*	8.587*	6.741*	4.991*	11.835*	16.256*	−12.990*	11.275*
Num_rev	−0.007*	−0.009*	−0.016*	−0.008*	−0.002*	−0.007*	−0.014*	−0.022*	−0.026*	−0.023*	−0.015*
Stars1	−61.916*	−27.456*		−39.419*	−15.491*	−32.011*	−17.868	−41.678*	2.314		−26.153*
Stars2	−40.858*	−12.781*	43.281*	−13.230*	−6.801*	−16.398*	−32.717*	−15.977*	5.346*	−8.306*	−26.075*
Stars3	−11.952*	6.402*	100.663*	12.166*	13.510*	0.650	3.114	26.409*	76.975*	8.309*	−3.399*
Stars4	18.377*	28.996*	144.443*	70.731*	37.096*	35.981*	38.710*	104.752*	107.123*	40.849*	27.646*
Stars5	123.371*	79.446*	313.340*	187.216*	131.823*	197.122*	158.003*	230.157*	191.237*	86.936*	130.593*
Dist	−9.087*	−3.466*	−16.941*	−4.427*	−2.688*	−15.076*	−4.087*	−6.104*	−13.619*	−2.619*	−5.107*
Lag	−0.083*	0.006	−0.001	−0.045*	−0.009*	0.031*	−0.028*	−0.036*	−0.008	−0.071*	−0.018*
R^2	0.349	0.323	0.762	0.652	0.462	0.405	0.282	0.506	0.660	0.480	0.523
Adj. R^2	0.349	0.322	0.761	0.652	0.462	0.405	0.281	0.506	0.659	0.478	0.523
Num. obs.	10 884	8 458	2 041	34 696	20 406	12 156	5 888	24 452	3 640	3 302	12 467

$^{*}p < 0.01$

dummy variables for seasonality are necessary to proceed with a meaningful investigation of the early booking effect. We only used regression analysis to gain a general understanding of the way hotel room features and hotel location influence price. We plan to extend the existing models further, particularly using more detailed or broad dataset.

V. Conclusion

In this paper we made an attempt of a comprehensive analysis of a broad European hotel dataset with prices a year ahead. We aimed at revealing hotel room price determinants, especially ones concerning city differences and early booking advantages. The biggest part of work was scraping data from Booking.com site, which left little time to answer some of the questions raised, and a huge (compared to other studies) sample, which, however, needs additions.

The preliminary hedonic pricing analysis showed that some hotel room features, not significant or not included in other studies (breakfast, free cancellation), substantially contribute to the room price according to our basic loglinear model. Individual linear models for cities seemed to be questionable (marginal effects variety is sometimes hard to explain) and thus gave us reasons to reconsider model specifications in the future. For one thing, problems in models and their interpretation could have been caused by the sample specifics: room price changes might be caused by room demand fluctuations, roomtype shifts, or both, and from the dataset we had it was difficult to identify each cause. We plan to avoid this issue further by downloading more details on rooms and including in the model at least some of the missing variables.

Three main conclusions made based on regression analysis and also some sample statistics and graphics are:

- There is a noticeable difference between cities not only regarding general hotel room prices and quality, but also in terms of price dynamics and seasonality.
- It might really be advantageous to book a room in advance because of a wider selection of cheap rooms in good hotels ("best offers").
- The nonlinearity of price trends prevents us from determining the precise effect of early booking, as it changes over time in a special way for each city; however, there is a slight tendency to price reduction even in the loglinear model.

We failed to calculate the exact number of days needed for effective early booking, yet we discovered that this number is most likely unique for each season in each city.

Further we plan to take in account our findings (add some variables, download more data) and refine our models. Dividing the hotels into some groups (other than natural groups by city) via time series clustering [6] or other techniques [2][7] might also be helpful in developing adequate hedonic price models or coming to some new conclusions in the future.

References

[1] Blanke J., Chiesa T. The Travel & Tourism Competitiveness Report 2013. // World Economic Forum 2013.
URL: http://www3.weforum.org/docs/WEF_TT_Competitiveness_Report_2013.pdf

[2] Brandmaier A.M. (2011) Permutation Distribution Clustering and Structural Equation Model Trees. // Saarland University, Ph.D. Thesis.

[3] Chen C., Rotschild R. (2010) An Application of Hedonic Pricing Analysis to the Case of Hotel Rooms in Taipei. // Tourism Economics, 2010, 16 (3).

[4] Clerides S.K., Nearchou P. and Pashardes P. (2003) Price and Quality in International Tourism. // Discussion Paper.

[5] Coenders G., Espinet J.M. and Saez M. (2003) Predicting Random Level and Seasonality of Hotel Prices. A Latent Growth Curve Approach. // Tourism Analysis, Vol. 8.

[6] Focardi S.M. (2001) Clustering Economic and Financial Time Series: Exploring the Existence of Stable Correlation Conditions. // Discussion Paper.

[7] Giorgino T. (2009) Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. // Journal of Statistical Software, 31(7), 1–24.

[8] ITB World Travel Trends Report 2012/2013. URL: http://www.itb-berlin.de/media/itbk/itbk_media/itbk_pdf/WTTR_Report_2013_web.pdf

[9] Lancaster K.J. (1966) A New Approach to Consumer Theory. // Journal of Political Economy 74, 132-157.

[10] Mattila A. S., ONeill J. W. (2003) Relationships between Hotel Room Pricing, Occupancy, and Guest Satisfaction: A Longitudinal Case of a Midscale Hotel in the United States. // Journal of Hospitality & Tourism Research.

[11] Rosen S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. // Journal of Political Economy 82(1), 34-55.

[12] Thrane C. (2006) Examining the Determinants of Room Rates for Hotels in Capital Cities: The Oslo Experience. // Journal of Revenue and Pricing Management, Vol. 5, 4 315–323.

[13] Scotland's Railway Fares and Tickets,
http://www.scotrail.co.uk/content/fares-and-tickets#advance-purchase