Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in DISCRETE MATHEMATICS

Electronic Notes in Discrete Mathematics 47 (2015) 213–220 www.elsevier.com/locate/endm

Clustering cities based on their development dynamics and Variable neigborhood search

B. S. Zhikharevich [a,1] O. V. Rusetskaya [a,2] N. Mladenović [b,3]

[a] *Department of Urban and Regional Economics, National Research University, Higher School of Economics, Saint Petersburg, Russia*

[b] *LAMIH, University of Valenciennes, Valenciennes, France*

Abstract

Clustering cities based on their socio-economic development in long time period is an important issue and may be used in many ways, e.g., in strategic regional planning. In this paper we continue our recent study where cumulative attribute for each year replaces nine other attributes, called 'vector of dynamics'. In our previous paper some original ranking method was proposed. Using the same data set, here we try out some classical clustering models such as Minimum sum of squares and Harmonic means clustering. Results for the two last models are obtained using Variable neighborhood search based heuristics. A comparative study among old and new results on 120 Russian large cities are provided and analyzed.

Keywords: Socio-economic, Ranking, Clustering, k-Means, Variable neighborhood search.

[1] Email: zhikh@leontief.ru
[2] Email: olga@leontief.ru
[3] Email: nenad.mladenovic@univ-valenciennes.fr

http://dx.doi.org/10.1016/j.endm.2014.11.028

1 Introduction

Generally speaking, the purpose of our research is to evaluate fluctuations in the dynamics of socio-economic development of the major Russian cities. Those results are supposed to be used in the future to identify factors that affect change in the dynamics type, including factors related to the urban development strategy of the city. Research objects were 120 major Russian cities, including 77 regional centers of Russia (except Moscow, St. Petersburg, capital of Chechen Republic - Grozny city and capital of Ingushetia - Magas). Regional centers that have strategic planning documents known to us are included in sampling. The main source of information for the calculation of "dynamics index" was statistical data from "Regions of Russia: main social and economic indicators of Cities", years 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012.

In the recent paper [6] we propose ranking approach for grouping 120 Russian cities that use the dynamic data set, or "vector of dynamics". In fact, such data are obtained when a cumulative attribute for each year is considered as a substitute for nine other attributes. In that way each city is represented by point in each year. Therefore, if 10 years period is considered, then the size of an input matrix is 120×10.

In this paper we propose two clussical clustering methods on the same data set: (i) Minimum sum of squares clustering and (ii) Minimum harmonic means clustering. We wanted to check how ranking and clustering methods compare in solving this type of problem. Since both problems are NP-hard, we use heuristic approach for finding near optimal solution. Results obtained are compared and analyzed.

In the next section we describe a way of collecting and modifying original data in order to get 'vectors of dynamic'. In section 3 we briefly explain steps of the recent method for ranking of 120 Russian cities. Since full details are only available in Russian lanquage (i.e., in [6]), Section 2 and 3 does not contain new results. In section 4 we provide results obtained by two well known clustering paradigms: Minimum sum of squares and Harmonic means. Computational results for clustering models are obtained by known Variable neighborhood search based heuristics. Conclusions and suggestions for possible future work in given in section 5.

2 Vector of dynamics

First stage - defining attributes. In research project nine attributes (indicators) were used. the value growth of which can be almost certainly interpreted as indicator of positive city development: (1) population size; (2) annual average number of workers; (3) average monthly nominal wages; (4) average living area in square meter per citizen; (5) number of physicians per 10000 population; (6) volume index of industrial output or volume of shipped own produced goods. works performed and services rendered; (7) value of construction work done; (8) retail turnover; (9) investments in fixed assets. Observed period: 2002 - 2011.

Second stage - calculation of growth rates. For each indicator and each city the growth rates were calculated for the respective year to the base year 2002.

Third stage - setting growth rates. Each growth rate for each year is valuated for the set of cities by referring to the average growth rate for the set of cities. As a result, for each city for each year indicator "relative dynamic coecient" was obtained that show how the city on this indicator was growing (faster or slower) comparing to the set of cities average.

Forth stage - calculation of integral "dynamics index". For each city and each year the integral "dynamics index" is calculated as arithmetic mean value of nine indicators "relative dynamic coecients". This index shows for each year how the city grows (faster or slower) relative to the set. For example the index of the city 1.04 means that is was growing 4% faster that the set of cities on average. Aggregation of city's dynamics indexes in one nine dimensional vector (by the years of observation) gives "dynamics vector".

These "dynamic vectors" for 120 cities can be studied by using different methods of city's grouping depending on the character / value of city "dynamics vector". This research addressed following methods of grouping: ranking method; cluster analysis methods. In the next section we give methods used to cluster 120 Russian cities.

3 Ranking method

Set of cities is first rank by decreasing order of "dynamics index". As a result, in each of the periods, two groups of cities can be distinguished:

(i) group A - cities with the "dynamics index" greater than 1, which are growing faster than the average for the observed sample of cities;

(ii) group B - cities with the "dynamics index" less than, which are growing slower than the average for the observed sample of cities.

Cities from group A that grow faster than average compose 30% to 45% that is less than half of the sample. Minimum value of "dynamics index" for the sample during the observed period tended to reduce, i.e., cities appeared that fall behind the average. Thus, the "dynamics index" of the city Nakhodka in 2007 was 0.588 - minimum value in this period and for the entire period from 2003 to 2011. Maximum value of "dynamics index" for the sample during the observed period had a mixed trend: growth of this indicator was replaced by its fall. Maximum value of "dynamics index" - 5.444 in 2009 was in the city Naryan-Mar, i.e. this year it overtook the average level more than five times. Within the groups A and B two rank groups were distinguished in each:

(i) Group 1 - cities in which "dynamics index" is above average of group A;

(ii) Group 2 - cities in which "dynamics index" is below average of group A;

(iii) Group 3 - cities in which "dynamics index" above average of group B;

(iv) Group 4 - cities in which "dynamics index" below average of group B.

Once the rank groups were defined for every city in each of the observed periods, "rank dynamics vector" was built for every city - sequence of nine numbers corresponding to the type of the rank group, to which the city belonged in each year from 2003 to 2011. On the next step, grouping of cities was made by types of trajectories based on "rank dynamics vectors". For this purpose those types of city movement trajectories by rank groups were defined:

(i) **Steadily advancing** - consistently high places (1-2 group. possible one transition in group 3).

(ii) **Steadily falling behind cities** - consistently low places (3-4 group, possible one transition in group 2).

(iii) **Accelerating develop** - ment - inconsistent with positive dynamics - transition from low places to high (possible one reverse transition).

(iv) **Slowing down development** - unstable with negative dynamics - transition from high places to low (possible one reverse transition).

(v) **Fluctuating** with mixed trends.

In determining characteristics of the referring to the above trajectories, it was assumed that change in the quality of dynamics in urban development appears in the change of rank for at least two straight years, so the rules in brackets were introduced.

4 Clustering approach

In this section we try two classical clustering methods on Russian city data, generated as explained earlier. Input matrix X has $n = 120$ rows and 9 columns. Each row corresponds to a city, why columns correspond to years examined (i.e., from 2002 to 2011). Therefore the row i of matrix X gives dynamics vector of city i. In that way each city is represented as a point in 9-dimensional Euclidean space. An element d_{ij} of dissimilarity matrix D measures the dissimilarities between each two cities i and j. It is calculated using Euclidean distance in R^9.

The most common criterion to cluster objects in Euclidean space is Minimum sum of squares clustering (MSSC) criterion. The desired number of clusters, say k, is given in advance. One wants to make k clusters (or groups) of objects such that the sum of squares distances among objects from the same group is minimum. This problem is NP-hard [2]. The most popular heuristics for solving it is k-means heuristic. MSSC contains 2 types of variables, location variables (coordinates of the centroid of each cluster) and allocation (membership of each object to cluster). By k-means, one type of variables is fixed while the other type is found: for fixed centroid points, the best assignments of entities to clusters are found and then, for a given alloca- tions of entities to cluster, the best centroid points are found. Those fixation continues until there is no more improvement in sum of squares objective function. Another solution method we implement is Variable neighborhood search based heuristic [4] (see also [5] for the recent survey on VNS).

It has been observed that the clustering results are more stable and less dependent on initial solution if Harmonic means clustering criterion is used instead of the sum-of-squares criterion. Hence, we also apply so called $k-$ Harmonic means clustering (KHMC) criterion on our problem. Again, two methods are tested: well known k-Harmonic means and VNS based heuristic [1,3]. In this paper we use the value of $m = 2$ for the membership parameter of harmonic means clustering [3].

Table 1 summarizes results obtained by the two clustering methods (k-means and VNS) applied on two clustering models: MSSC and KHMC.

Number	MSSC		KHMC	
of clusters	k-means	VNS	k-HM	VNS
2	54.47	51.14	86.28	86.28
3	35.37	31.59	68.53	68.52
4	20.80	20.57	56.05	56.05
5	15.61	15.56	47.96	47.96
6	13.70	13.53	43.39	42.61
7	12.16	12.07	39.09	39.06
8	11.02	10.10	36.85	36.85
9	10.35	8.97	35.48	35.47
10	9.81	8.30	32.54	32.45

Table 1
Clustering 120 Russian cities: errors for MSSC and KHMC for different number of clusters

From Table 1 we may conclude the following:

(i) VNS outperforms k-means heuristics for both objective functions, but in some cases results are very close;

(ii) The advantage of using VNS based heuristic is larger in solving MSSC problem than in solving KHMC problem;

(iii) Probably the most appropriate number of clusters is $k = 4$, since the relatively largest drops in objective function values, for both objectives, are between k=3 and k=4 (see the drop from 31.59 to 20.57 for MSSC and the drop from 68.52 to 56.05 of KHMC).

(iv) We observed the high sensitivity of both clustering problems: a similar objective values could be very far from each other, i.e., very different k clusters could have similar MSS or KHM objective values.

In table 2 we give the number of entities obtained by VNS for both criteria. Column 1 gives the number od desired clusters k, while columns 2 and 3 report on objective function values for MSSC and KHMC respedtively. Columns 4 and 5 report on the number of entities in each cluster obtained by MSSC and KHMC respectively. For example, for $k = 2$, order pair (118,2) reports that there are 118 cities in the first cluster and only 2 in the second.

One can observe the following:

(i) Very different clusters are obtained by two criteria, although both use

	Error		Number of cities in each cluster	
k	MSSC	KHMC	MSSC	KHMC
2	51.14	86.28	(118, 2)	(114, 6)
3	31.59	68.52	(114,5,1)	(83,36,1)
4	20.57	56.05	(96, 21, 2, 1)	(74, 40, 4, 2)
5	15.56	47.96	(80, 34, 3, 2, 1)	(50, 39, 25, 4, 2)
6	13.53	42.61	(57, 38, 19, 3, 2, 1)	(42, 35, 24, 13, 4, 2)
7	12.07	39.06	(49, 36, 16, 13, 3, 2, 1)	(34, 31, 20, 18, 11, 4, 2)
8	10.10	36.85	(49, 36, 16, 13, 3, 1, 1,1)	(29, 29, 20, 14, 11, 11, 5, 1)
9	8.97	35.47	(35, 32, 17, 17, 13, 3, 1, 1, 1)	(26, 25, 18, 17, 17, 11, 3, 2, 1)
10	8.30	32.45	(35, 33, 13, 13, 11, 9, 3, 1, 1, 1)	(25, 20, 19, 18, 13, 11, 8, 3, 2, 1)

Table 2
Distribution of cities to different clusters by VNS and two objectives: MSS and KHM.

the square distances (m=2 for KHMC). This fact can be explained by the sensitivity of vector of dynamic data;

(ii) Results obtained by KHMC seems to be more reliable and more stable: they keep cities in the same group longer, when k is increased;

(iii) The number of small clusters and outliers are larger for MSSC;

(iv) Both methods clearly recognized 6 cities that are different that others: Serpukhov, Kaliningrad, Elista, Khanty-Mansiysk, Anadyr and Naryan-Mar.

We next compare clustering results with those obtained with ranking method from [6], where 5 groups of citeies are recognised. Since the vector of dynamic data are very sensitive, direct comparison among approaches is almost impossible. Even groups of cities obtained by two similar clustering criteria were very different. However, we found that the same outliers and small groups of cities are recognized by both approaches.

5 Conclusions

In this paper we suggest two clustering techniques for grouping 120 large Russian cities, based on their socio-economic attributes. We use nine socio-economic attributes to evaluate fluctations in dynamics of development of

cities, suggested in [6]. As in [6], data are collected from Statisctical Russian proceedings in period of 10 years (from 2002 to 2011). Clustering techniques we use are Minimum sum of squares clustering (MSSC) and minimum harmonic means clustering (KHMC). It appears that more stable results are obtained with KHMC method with less number of outliers. When compared with existing ranking method, results obtained by clustering approaches are similar, although some merging of small clusters occurre in ranking method. Note that data used are averages of all attributes for one year. The future work may consist of clustering of full data set without aggregation of the data set.

Acknowledgments

Work of Nenad Mladenovic is conducted at National Research University Higher School of Economics and supported by RSF grant 14-41-00039.

References

[1] Alguwaizani A., P. Hansen, N. Mladenović and E. Ngai, *Variable neighbourhood search for harmonic means clustering*, Applied Mathematical Modelling **35** (2011), pp. 2688–2694.

[2] Aloise D., A. Deshpande, P. Hansen and P. Popat, *NP-hardness of Euclidean sum-of-squares clustering*, Machine Learning **75** (2009), pp. 245–248.

[3] Carrizosa E., A. Al-Guwaizani, P. Hansen, N. Mladenović, *New heuristic for harmonic means clustering*, to appear in Journal of Global Optimization (2014).

[4] Hansen P. and N. Mladenović, *J-Means: A new local search heuristic for minimum sum-of-squares clustering*, Pattern Recognition **34** (2001), pp. 405–413.

[5] Hansen P., N. Mladenović and Pérez J. A. M., *Variable neighbourhood search: algorithms and applications*, Annals of Operations Research **175** (2010), pp. 367–407.

[6] Zhikharevich, B. S., O. V. Rusetskaya, *Fluctuations in the socio-economic development of cities of Russia: methodology and results of the calculation of the "vector of dynamics"* (in Russian), to appear in Journal of Russian Geographical Society (2014).