International Mathematical Forum, Vol. 12, 2017, no. 2, 77 - 85
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/imf.2017.611157

Research on Heterogeneity of Children Morbidity Rate in Saint Petersburg

Vladimir M. Bure, Kseniya U. Staroverova

Faculty of Applied Mathematics and Control Processes
Saint Petersburg State University, Saint Petersburg, Russia

Abstract

In the paper we present a research on heterogeneity of Saint Petersburg districts with respect to children morbidity rate. We used methods of cluster analysis for discovering groups of similar districts. As most of present distances work well with long time series but have biased results for short ones, the dissimilarity measure for short time series based on its characteristics is proposed. We constructed several clustering models to determine clusters which did not depend on the method. Five objects (districts) were distributed to several clusters in different models. For dealing with this problem an heuristic algorithm for inclusion of indeterminate objects to potential clusters was built. As a result we got three groups of districts which were similar in some sense.

Mathematics Subject Classification: 62-07

Keywords: cluster analysis, time series, applied statistics

1 Introduction

It is a problem to organize a healthcare system in megapolis. A hierarchical structure of system simplifies healthcare managment. Each district of Saint Petersburg has an executive body responsible for control and supervision of hospitals, clinics, ambulance activity. They collect the statistics and transmit it to higher level organizations. Health problems of inhabitants can be connected with different factors such as enviromental problems, age of population,

economic problems, activity of executive body of every district and the whole city and so on. To define the reason it is necessary to know if districts are different. That is why we study a problem of heterogeneity of Saint Petersburg districts. The morbidity rate was choosen as indicator of population health. This article contains the research on children morbidity rate as it is important factor in development of the city.

Methods of cluster analysis was utilised for detection of homogeneous groups of districts. Key feature is that data are time dependent so it is necessary to use special dissimilarity measures. There are a huge amount of studies of time series dissimilarity measures which are based on autocorrelation [1], spectral characteristics [2], assumptions of time series model [3,4], correlation [5], wavelet transformation [6,7] and others.

Three stable clusters were discovered by applying several methods, one of them is proposed by us in section 2. Five districts were assigned to several clusters so we built an algorithm based on Borda count, which determined each of controversial objects to a particular cluster [8].

2 Time series dissimilarity measures

Clustering is one of the machine learning approaches for unsupervised learning. We do not know the number of clusters or the distribution of objects to groups, so the model of clustering and distance define the results. That is why aplication of cluster analysis requires precision of dissimilarity measure selection. Our problem calls for methods that take into account that data are time dependent. **The Euclidean distance** could be relevant for the time series clustering only in the situation when dynamics of values is not important. The statement ensues the fact that after permutation of observations the distance between time series does not change.

In 1906 a French mathematician Maurice René Fréchet suggested an approach for measuring of similarity between curves. T. Eiter and H. Mannila proposed a **discrete Fréchet distance** and gave intuitive definition in [9].

Assume that time series belongs to the class of invertible and stationary **ARMA processes**. Then dissimilarity between two time series can be computed by testing whether or not two time series have significantly different generating processes. This method was introduced by E. A. Maharaj in 2000 [3].

Computation of the Euclidean distance between the **periodogram coefficients** of the time series was described by J. Caiado, D. Peña and N. Crato in 2006 [2]. If correlation structure of time series is essential then application of the Euclidean distance between the normalized periodogram ordinates is more appropriate.

A. D. Chouakria and P. N. Nagabhushan (2007) presented **an adaptive dissimilarity index** (ADI) for measuring time series proximity [10]. It covers both the conventional measure for the proximity with respect to values and the temporal correlation for the proximity with respect to behavior. Besides researcher can regulate the influence of both constituents.

Many present dissimilarity measures work well only with long time series. Unfortunately such fields as economics or demography are more presented by short time series. Therefore we propose a new **dissimilarity measure based on time series characteristics** (CBD — characteristics based distance). This section contains a brief review of the method.

Let $M = [n \times m]$ is a matrix where each row is a time series with m observations, then $M_k = \{M_{k,1}, M_{k,2}, \ldots, M_{k,m}\}$ is time series that corresponds to row $k \in \{1, 2, \ldots, n\}$. The distance depends on three addends

$$\begin{aligned} Dist_{CBD}(M_{k_1}, M_{k_2}) = \alpha\, D_1(M_{k_1}, M_{k_2}) + \beta\, D_2(M_{k_1}, M_{k_2}) + \\ + (1 - \alpha - \beta)\, D_3(M_{k_1}, M_{k_2}). \end{aligned} \tag{1}$$

The first addend of (1) shows dissimilarity with respect to values. To compute $\alpha\, D_1(M_{k_1}, M_{k_2})$, firstly data normalization is accomplished

$$\overline{m}_{k,t} = \frac{m_{k,t} - \min M}{m_{k,t} - \max M},$$

$$\overline{M} = \{\overline{m}_{k,t}\}, k \in \{1, 2, \ldots n\}, t \in \{1, 2, \ldots m\},$$

Then for each row a vector with five characteristics is computed. These are such values as mean, standard deviation, median, minimum and maximum values of $\overline{M}$. The Euclidean distance between the vectors is $D_1(M_{k_1}, M_{k_2})$. Coefficient α belongs to $[0, 1]$ and determines the influence of D_1 on the $Dist_{CBD}$.

The second addend $\beta D_2(M_{k_1}, M_{k_2})$ is dissimilarity with respect to dynamics where β is in $[0, 1]$. For every row of matrix M we construct four binary vectors. Elements of the first vector, which are equal to 1, correspond to increasing intervals. Ones of the second vector show the elements which are equal to mean of time series or higher. The third and the fourth vectors present fluctuations around $E(M_k) + sd(M_k)$ and $E(M_k) - sd(M_k)$.

The third addend of (1) is dissimilarity with respect to variability of time series. Firstly we normalize data in such a way that every time series has the maximal value equal to 1 and minimal value equal to 0, it is given by

$$\widetilde{m}_{k,t} = \frac{m_{k,t} - \min M_k}{\max M_k - \min M_k}, \tag{2}$$

$$\widetilde{M_k} = \{\widetilde{m}_{k,1}, \widetilde{m}_{k,2}, \ldots, \widetilde{m}_{k,m}\}, k \in \{1, 2, \ldots n\}.$$

Table 1: Evaluation of clustering models with different distances and number of clusters

Name of distance	Dunn index			Silhouette index		
	2	3	4	2	3	4
Fréchet	0.49	0.33	0.39	0.45	0.34	0.23
Euclidean	0.31	0.45	0.47	0.33	0.30	0.24
ADI(2)	0.16	0.17	0.17	0.32	0.32	0.23
ADI(3)	0.05	0.11	0.11*	0.25	0.33	0.29*
Periodogram coefficients	1.28*	0.75*	0.94*	0.72*	0.58*	0.51*
ARMA	0.32*	0.32*	0.31*	−0.43*	−0.43*	−0.42*
CBD(0.4, 0.4)	0.54	0.31	0.49	0.32	0.17	0.22
CBD(0.8, 0.1)	0.37	0.50	0.49	0.32	0.30	0.26
CBD(0.6, 0.1)	0.32	0.42	0.50	0.31	0.23	0.26

The transformation (2) allows us to ignore the differences in values and focus on differences in variability. We construct matrices of differences $\widetilde{FD}(1)$ and $\widetilde{FD}(2)$ of new matrix $\widetilde{M} = \{\widetilde{M}_1, \widetilde{M}_2, \ldots, \widetilde{M}_n\}^T$, whose elements are given by

$$\widetilde{FD}_{k,t}(l) = \widetilde{M}_{k,t+l} - \widetilde{M}_{k,t},$$

for $l \in \{1, 2\}, k \in \{1, 2, \ldots, n\}, t \in \{1, 2, \ldots, m-1\}$. Then a vector of variability characteristics $V_k = [1 \times 15]$ for each time series k is computed. The first three components, calculated for $\widetilde{FD}(1)$, correspond to average growth, decline and fluctuation. The next three elements are the same but they are computed for $\widetilde{FD}(2)$. Components $V_{k,7}$ and $V_{k,8}$ refer to the greatest growth and decline of $\widetilde{FD}(1)$, while $V_{k,9}$ and $V_{k,10}$ are the greatest growth and decline of $\widetilde{FD}(2)$. The last five elements of V_k are minimal value, quartiles and maximal value of $\widetilde{FD}(1)$.

We conducted experiments with 3 sets of synthetic data. A description of experiments is beyond the purpose of our paper. We just note that the results showed that application of CBD was appropriate as the similarity between the true cluster solution and the one obtained with CBD was higher than majority of present dissimilarity measures had.

3 Cluster analysis of children morbidity rate

We have annual children morbidity data from 1999 to 2014 for every district of Saint Petersburg. It is important for us to define homogeneous groups of districts. Values of morbidity rate are more important than the dynamics, moreover the last observations are more substantial than old ones.

Firstly we calculated the distance matrices using dissimilarity measures which were described in previous section. We used the k-medoids algorithm

Table 2: Distribution of districts in the clusters

Name of distance	Districts													
	1-2	3	4-6	7	8	9	10	11	12	13	14	15	16	17-18
Fréchet	3	2	1	2	2	3	2	2	3	3	2	3	1	2
Euclidean	1	2	1	2	1	3	1	2	1	3	2	3	1	2
ADI(2)	1	2	1	2	2	3	1	2	1	3	2	3	1	2
CBD(0.8, 0.1)	1	2	1	2	2	3	2	2	1	3	2	3	1	2

for breaking the dataset up into groups. *Table 1* with values of Dunn and Silhouette indexes helped us to determine that the number of clusters was 3. The asterisk above the value means that the cluster with only one object was produced by the procedure, apparently it was not desirable outcome. The reason of comparing models with only 2, 3 and 4 clusters is that greater number of groups is a source of single-object clusters. We can see that dissimilarity measures based on periodogram coefficients and ARMA-model have a single-object clusters in all cases (notice that the districts in single-object clusters are different), that is why we excluded them from further analysis. Moreover ADI(2) and ADI(3) represent the same method but with different values of parameter so correspondingly to *Table 1* we chose ADI(2). Similarly we chose parameters $\alpha = 0.8, \beta = 0.1$ for CBD. So we continued cluster analysis with four distances and considered that there were 3 clusters.

The next step was to find stable clusters. Stable cluster is a group of objects which are placed to the same cluster by all dissimilarity measures, which are applied in analysis. We noticed, that stable clusters were $\{4, 5, 6, 16\}$, $\{3, 7, 11, 14, 17, 18\}$, $\{9, 13, 15\}$ (*Table 2*) while objects $\{1, 2, 8, 10, 12\}$ were distributed to different clusters. For dealing with this problem we built heuristic algorithm. We considered the selection of the cluster as voting game where clusters were candidates.

Step 1. Make a list $L_{ind} = \{x_1, x_2, \ldots, x_l\}$ of indeterminate objects.
Step 2. Make a list $L_{cl}^k = \{c_1^k, c_2^k, \ldots, c_3^k\}$ of potential clusters for every indeterminate object $x_k \in L_{ind}$.
Step 3. Compose a key characteristic F_0 which evaluates the quality of cluster. We assume that when F_0 is going down (going up) the quality of clustering is increasing.
Step 4. Compose several other characteristics $F_1, F_2, \ldots, F_m$ which define the quality of clustering. These characteristics are voters. As on the previous step we assume that when $F_i, i \in 1, 2, \ldots, m$ is going down (going up) the quality of clustering is increasing.
ITERATION
Step 5. Assume that $x_1 \in L_{cl}^k$ is added to every $c_i^1 \in L_{cl}^1$. Compute F_0

Table 3: Iteration 1, step 5. The smallest value has a grey color

L_{ind}	1		2		8		10		12	
L_{cl}	1	3	1	3	1	2	1	2	1	3
F_0	1.76	1.83	1.58	1.66	1.76	1.55	2.06	1.50	1.88	1.53

for every new $c_i^1 \in L_{cl}^1$. Then do the same actions for every object in L_{ind}.

Step 6. Choose the candidate with the lowest (highest) F_0, let it be the cluster c_d^n which corresponds to the object x_n.

Step 7. Compute $F_1, F_2, \ldots, F_m$ for every candidate in L_{cl}^n.

Step 8. Rank the voters as in Borda count method. Give m points to candidate with the lowest (highest) F_i, $(m-1)$ point to candidate, who has the next lowest (highest) F_i and so on.

Step 9. Find a sum of points for every candidate in L_{cl}^n. Candidate with the highest sum is a winner c_e^n.

Step 10. If cluster c_e^n coincides with cluster c_d^n then continue, else go to the step 6, but exclude c_d^n from the game on this iteration.

Step 11. Include object x_n in cluster c_e^n. Remove x_n from L_{ind}. If L_{ind} is empty then finish the algorithm, else go to the Step 5 (new iteration).

The algorithm unambiguously determines the order of addition of elements from L_{ind} to potential clusters. If it is impossible to select the cluster for x_i from L_{ind} using Borda count method then inclusion or exclusion of F_i from $F_1, F_2, \ldots, F_m$ or application of modified Borda method are eligible.

We applied the algorithm presented above. On the steps 1-2 we got the lists $L_{ind} = \{1, 2, 8, 10, 12\}$ and $L_{cl} = \{\{1,3\}, \{1,3\}, \{1,2\}, \{1,2\}, \{1,3\}\}$.

We decided to lower the influence of statistical error by computing simple moving average (with lag equal to 3). Considering that absolute differences between smoothed time series are going down when the quality of cluster is going up we chose this value as F_0.

Characteristics for step 4 were calculated for every cluster,

- F_1: maximum absolute difference between values of last 3 periods;
- F_2: maximum coefficient of variation;
- F_3: ratio of the sum of absolute differences between last 3 observations in the cluster without a new element to the sum of absolute differences between last 3 observations in the cluster with a new element;
- F_4: ratio of the standard deviation in the cluster without a new element to the standard deviation in the cluster with a new element;

We show implementation of one iteration of the algorithm for clarity. Correspondingly to step 5 we calculate F_0 for every element in L_{ind} assuming that we add objects to potential clusters. The minimum of the sum of absolute differences between observations of smoothed time series is reached when dis-

Table 4: Iteration 1, steps 7-9

(a) *Characteristics values for clusters 1, 2 after inclusion of object 11*

Clus-ters	Characteristics			
	F_1	F_2	F_3	F_4
1	851.16	0.16	2.55	1.21
2	418.15	0.16	1.38	1.07

(b) *The ranking of candidates correspondingly to Table 4 (a)*

Candi-dates	Voters				
	F_1	F_2	F_3	F_4	Sum
1	1	2	1	1	5
2	2	2	2	2	8

trict 10 is included in the cluster 2 (*Table 3*). So we compute characteristics $F_1 - F_4$ for clusters $\{4, 5, 6, 16\} \cup 10$ and $\{3, 7, 11, 14, 17, 18\} \cup 10$ (*Table 4*).

As the winner is cluster 2 (*Table 4 (b)*) and it coincides with cluster with the lowest F_0 from step 5 (*Table 3*), we can add object 10 to cluster 2. Thus we get $L_{ind} = \{1, 2, 8, 12\}$ and $L_{cl} = \{\{1, 3\}, \{1, 3\}, \{1, 2\}, \{1, 3\}\}$. This is the end of the first iteration and we go to step 5. As a result we got 3 groups of districts (*Figure 1*).

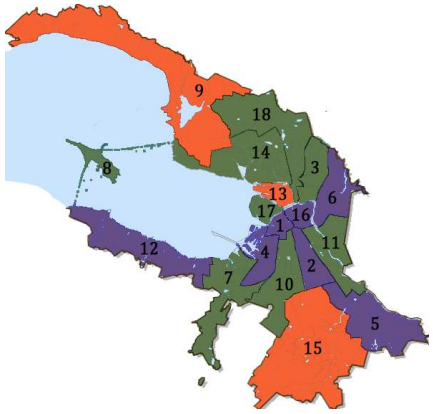

Figure 1: The map of Saint Petersburg. Districts from one cluster have the same color. Districts: *1* – Admiralteysky, *2* – Frunzensky, *3* – Kalininsky, *4* – Kirovsky, *5* – Kolpinsky, *6* – Krasnogvardeysky, *7* – Krasnoselsky, *8* – Kronshtadtsky, *9* – Kurortny, *10* – Moskovsky, *11* – Nevsky, *12* – Petrodvortsovy, *13* – Petrogradsky, *14* – Primorsky, *15* – Pushkinsky, *16* – Tsentralny, *17* – Vasileostrovsky, *18* – Vyborgsky.

4 Conclusion

Cluster analysis of 18 time series, which respond to the children morbidity rate in each district of Saint Petersburg, was conducted. We gave a brief review of dissimilarity measure based on time series characteristics that works well with short time series and several other distances in the second section and then we used them in the third section. Firstly we constructed 9 different variants of

clustering but the worst 5 ones were excluded. Applying 4 different methods we found 3 groups of objects which were included in the same cluster in all clustering models. Unfortunately 5 districts had several possible variants of distribution. That is why we proposed the algorithm for selection the only cluster for problematic objects. Following the algorithm we got such 3 clusters such that districts inside one of the clusters were more similar to each other than two districts from the different clusters were.

References

[1] P. Galeano, D. Peña, Multivariate analysis in vector time series , *Resenhas IME-USP,* **4** (2000), no. 4, 383 - 169.

[2] J. Caido, N. Crato and D. Peña, A periodogram-based metric for time series classification, *Computational Statistics & Data Analysis,* **50** (2006), 2668 - 2684. https://doi.org/10.1016/j.csda.2005.04.012

[3] E.A. Maharaj, Clusters of time series, *Journal of Classification,* **17** (2000), 297 - 314. https://doi.org/10.1007/s003570000023

[4] U. Triacca, Measuring the distance between sets of ARMA models, *Econometrics,* **4** (2016), no. 3, 32.
https://doi.org/10.3390/econometrics4030032

[5] X. Golay, S. Kollias, G. Stoll, D. Meier and A. Valavanis, A new correlation-based fuzzy logic clustering algorithm for FMRI, *Magnetic Resonance in Medicine,* **40** (1998), no. 2, 249-260.
https://doi.org/10.1002/mrm.1910400211

[6] A. Graps, An introduction to wavelets, *Journal IEEE Computational Science & Engineering,* **2** (1995), no. 2, 50-61.
https://doi.org/10.1109/99.388960

[7] H. Zhang, T. B. Ho, Unsupervised feature extraction for time series clustering using orthogonal wavelet transform, *Informatica,* **30** (2006), 305-319.

[8] D. Easeley, J. Kleinberg *Networks, Crowds, and Markets: Reasoning about a Highly Connected World,* Cambridge University Press, Cambridge, 2010.
https://doi.org/10.1017/cbo9780511761942

[9] T. Eiter, H. Mannila, Computing discrete Fréchet distance, *Technical Report,* Technische Universität Wien.

[10] A.D. Chouakria, P.N. Nagabhushan, Adaptive dissimilarity index for measuring time series proximity, *Advanced in Data Analysis and Classification,* **1** (2007), no. 1, 5-21. https://doi.org/10.1007/s11634-006-0004-6

Received: December 12, 2016; Published: January 16, 2017