ИНФОРМАТИКА

UDC 519.237.8

V. M. Bure, K. Yu. Staroverova

APPLYING CLUSTERING ANALYSIS FOR DISCOVERING TIME SERIES HETEROGENEITY USING SAINT PETERSBURG MORBIDITY RATE AS AN ILLUSTRATION

St. Petersburg State University, 7–9, Universitetskaya nab.,
St. Petersburg, 199034, Russian Federation

One of the machine learning approaches for unsupervised learning is clustering. Clustering has the task of exploring the structure of data with the aim of assigning a set of objects in such a way that objects belonging to the same group are more similar to each other than the objects drawn from different groups. Determining the number of clusters in a data set, searching for stable clusters, selection of dissimilarity measure and algorithm are significant tasks of cluster analysis. Multidimentional clustering is often used when an object is characterized by a vector. A dissimilarity measure or distance is selected with respect to the purpose and features of a certain task. But there are also such fields as economics, geology, medicine, sociology that are often presented by time series. Time series are random processes but not a random vector. That is why it is important to construct such a similarity (or dissimilarity) measure which would take into consideration that data are time–dependent. The research of morbidity rate of Saint Petersburg from 1999 to 2014 years and clustering of 18 districts are conducted. Several different similarity measures are used for clustering. Besides, an interesting aspect is clustering of multidimentional time series. There are two approaches. The first concept is to split multidimentional time series into several univariate time series, whilst the second one is to consider it as a whole unit that preserves the influence of data interdependence. Research is made with application of TSclust, tseries packages in R and missed algorithms are realised there. As a result of clustering of Saint Petersburg districts applying several similarity measures three stable clusters are found out but seven districts do not belong to any cluster. Refs 10. Figs 2.

Keywords: cluster analysis, clustering, time series similarity measure, stable clusters.

В. М. Буре, К. Ю. Староверова

МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА КАК СПОСОБ ВЫЯВЛЕНИЯ НЕОДНОРОДНОСТИ ВРЕМЕННЫХ РЯДОВ НА ПРИМЕРЕ ПОКАЗАТЕЛЯ ЗАБОЛЕВАЕМОСТИ В САНКТ-ПЕТЕРБУРГЕ

Санкт-Петербургский государственный университет, Российская Федерация,
199034, Санкт-Петербург, Университетская наб., 7–9

Кластеризация относится к методам машинного обучения без учителя и широко применяется при анализе данных для распределения объектов по группам (кластерам) таким

Bure Vladimir Mansurovich — doctor of technical sciences, professor; vlb310154@gmail.com
Staroverova Kseniya Yurievna — student; ksenygnirps@gmail.com

Буре Владимир Мансурович — доктор технических наук, профессор; vlb310154@gmail.com
Староверова Ксения Юрьевна — студент; ksenygnirps@gmail.com
© Санкт-Петербургский государственный университет, 2016

 DOI: 10.21638/11701/spbu10.2016.404

образом, чтобы объекты одной группы оказались более схожими, чем объекты разных групп. Важными вопросами в кластерном анализе являются определение числа кластеров, выделение устойчивых кластеров, выбор расстояния между объектами и подхода кластеризации. Часто производится кластеризация многомерных объектов, которые характеризуются вектором случайных величин, и их мера сходства подбирается исходя из условий и особенностей задачи. Но объектами исследования многих областей, таких как экономика, геология, медицина, социология, часто являются не вектора случайных величин, а случайные процессы, что вновь приводит исследователей к проблеме построения меры сходства, учитывающей зависимость данных от времени. Проведено исследование показателя общей заболеваемости в Санкт-Петербурге с 1999 по 2014 г. и построена кластеризация 18 районов города. Продемонстрированы результаты кластеризации с использованием нескольких мер сходства, в том числе рассмотрены и меры сходства многомерных временных рядов. Кластеризация многомерных временных рядов может происходить двумя способами: первый — представить многомерный временной ряд как несколько одномерных, второй состоит в кластеризации самих многомерных рядов и учитывает взаимосвязи, которые могут присутствовать между переменными ряда. Кластеризация произведена с помощью библиотек *TSclust, tseries* пакета R; недостающие алгоритмы реализованы также на языке R. В результате кластеризации районов Санкт-Петербурга с применением нескольких мер сходства выявлено три устойчивых кластера, и семь районов не были отнесены к определенному кластеру из-за того, что они меняли свое расположение в зависимости от выбора меры сходства. Библиогр. 10 назв. Ил. 2.

Ключевые слова: кластеризация, мера схожести временных рядов, устойчивость кластеров.

Introduction. It is a well-known fact that clustering is widely used in various fields for solving applied problems. Investigators utilise traditional approaches of cluster analysis and often modify and adopt methods for WEB-development, marketing, economics, archeology, physics, medicine, biology, sociology problems, etc. The key feature of clustering is that an investigator does not have a training set and any information about the structure of objects. That leads to such problems as determination of the number of clusters [1], verification of results that usually require presence of a specialist who is an expert in the research area. Unfortunately, such results are often ambiguous.

Cluster analysis of multivariate objects has been extensively studied for the last 80 years [2]. However, there is not so much work done on time series clustering in spite of the fact that a lot of fields have time dependent data. There is more meant than meets the eye, as it is important to consider both the behavior of time series and the distance between the objects.

Dealing with time series clustering while working on healthcare problems we have a question if districts of Saint Petersburg are different (in any mathematical sense) with respect to healthcare [3]. The morbidity rate was chosen as the reflection of the health of people living in the district. It is obvious that clustering of one observation (for example 2014 year) does not make any sense as morbidity rate is different in districts due to a variety of reasons. As morbidity is an annual rate, we can consider time series and notice that morbidity rate of some districts has been stable for the last 16 years while in others it has been coming up or going down. In some cases it can lead to the conclusion of a badly organized healthcare system in one district and a well-organized one in others [4, 5].

Review of several similarity measures is presented in the article and are used for clustering of morbidity rate. For the analysis library "TSclust" of R package [6] is used and lacking algorithms are released in R for this research.

As a result three stable clusters were obtained and seven districts were not included in any of the stable clusters.

Related work. This section contains a brief review of terms and methods that are used in the work.

Multivariate time series (MTS) is a series of observations $x_i(t)$ where t is a discrete or continuous value (interpretable as time moments) and i is an index of some process that is changing through time. It is obvious that observations of MTS are made sequentially through time. Let $t \in \{1, \ldots, n\}$, $i \in \{1, \ldots, m\}$, index i means that in time moment t we can do m measurements of different processes simultaneously. MTS could be considered as a union of m univariate time series.

Euclid distance is well-known conventional metrics used for determination of distance, or more precisely, dissimilarity between objects. For univariate time series X and Y Euclid distance is the following:

$$d_E(X,Y) = \sqrt{(X(1)-Y(1))^2 + (X(2)-Y(2))^2 + \cdots + (X(n)-Y(n))^2}. \tag{1}$$

Frechet distance is a measure of similarity between curves that takes into account not only location but also ordering of the points along the curves. For calculation of Frechet distance we should consider a set M which contains all possible sequences r of k pairs preserving the observations order in the form: $r = ((X(a_1), Y(b_1)), \ldots, (X(a_k), Y(b_k)))$ where $r \in M$. There are constraints on indexes

- $a_1, \ldots, a_k, b_1, \ldots, b_k \in \{1, \ldots, n\}$;
- $a_1 = b_1 = 1$;
- $a_k = b_k = n$;
- $a_{j+1} = a_j$ or $a_{j+1} = a_j + 1$;
- $b_{j+1} = b_j$ or $b_{j+1} = b_j + 1$.

Then the Frechet distance between X and Y is defined as

$$d_F(X,Y) = \min_{r \in M} \max_{j=1,\ldots,n} |X(a_j) - Y(b_j)|. \tag{2}$$

An adaptive dissimilarity index [7] is a measure that takes into account the behavior of curves and location of observations. The distance depends on two values: the first $\mathrm{CorT}(X,Y)$ shows if time series X and Y behave the same or different way, the second $\delta_{conv}(X,Y)$ stands for any conventional measure as Euclid, Minkowski, Manhattan distance etc. Formal definition is presented in following formulas:

$$\mathrm{CorT}(X,Y) = \frac{\sum\limits_{t=1}^{n-1} \big(X(t+1) - X(t)\big)\big(Y(t+1) - Y(t)\big)}{\sqrt{\sum\limits_{t=1}^{n-1} \big(X(t+1) - X(t)\big)^2} \sqrt{\sum\limits_{t=1}^{n-1} \big(Y(t+1) - Y(t)\big)^2}}, \tag{3}$$

$$d_{\mathrm{CorT}}(X,Y) = f\big(\mathrm{CorT}(X,Y)\big)\delta_{conv}(X,Y), \tag{4}$$

$$f(\alpha) = \frac{2}{1 + \exp k\alpha}, \tag{5}$$

where $k > 0$ is parameter that regulates the influence of time series behavior on the final dissimilarity value.

Eros distance metric [8] is a similarity measure for MTS. Consider MTS X as a matrix $T_x = [n_x \times m_x]$ and Y as a matrix $T_y = [n_y \times m_y]$. For matrices T_x and T_y covariance matrices M_x and M_y could be calculated. Applying singular value decomposition to M_x and M_y two right eigenvector matrices $V_x = [v_{x_1}, \ldots, v_{x_m}]$, $V_y = [v_{y_1}, \ldots, v_{y_m}]$ could be obtained. Eros (Extended Frobenius norm) is defined as

$$\mathrm{Eros}(V_x, V_y, w) = \sum_{i=1}^{m} w_i |\langle v_{x_i}, v_{y_i} \rangle| = \sum_{i=1}^{m} w_i |\cos\theta_i|, \tag{6}$$

where w is a weight vector based on the eigenvalues of the MTS dataset and θ_i is the angle between v_{x_i} and v_{y_i}. Formula (6) shows the similarity of objects but often it is necessary to know the distance (dissimilarity) between the objects, so we can modify formula (6) and get a new equation

$$D_{\text{Eros}}(V_x, V_y, w) = \sqrt{2 - 2\sum_{i=1}^{m} w_i |\langle v_{x_i}, v_{y_i} \rangle|}. \tag{7}$$

Weighted Borda method [9] is a modification of the Borda count which is a single-winner election method where voters rank candidates in order of preference. Then each candidate gets a number of points corresponding to the number of candidates ranked lower. The winner of the election is determined as a candidate with the highest total point. Weighted Borda method takes into account the similarity gap between the candidates. Let's note candidates $s_0, \ldots, s_k$ and distances between query candidate and others $d_j = d_j(s_0, s_j), j \in \{1, \ldots, k\}$. We assume that s_0 is query candidate that means we want to find out candidates that are closest to him.

Without loss of generality we suppose that $d_{j-1} < d_j \ \ \forall j \in \{1, \ldots, k\}$ as we always can change the order of candidates. As our goal is clustering of MTS we can imagine that every candidate has m dimensions.

Very often some dimensions appear to be much more important than the others, that is why we can compute weights of each dimension (for example as in previous method where weights are based on eigenvector matrices) and use it in calculation of total score

$$vs_i^j = w_i\big(1 + k(1 - \frac{d_j}{d_k})\big) \ \ \forall j \in 1, \ldots, k, \forall i \in 1, \ldots, m. \tag{8}$$

Accumulating the score of each item of each candidate we can find the nearest candidates to the query candidate.

Univariate clustering. We have statistics of morbidity rate from 1999 to 2014 of 18 districts of Saint Petersburg which is defined in 3 age-groups: children (0–14 years), teenagers (15–17 years) and adults (over 18). According to our notation we get 18 MTS where m is equal to 3 and n is equal to 16. First, cluster analysis of univariate time series was made where clusters were determined by dendrogram.

It is obvious that conventional similarity measures may be not be suitable for time series clustering because they take into account only location of the observation and results of clustering would not change if we mixed moments of time when observations were made. We choose several methods for determination of distance between objects:

1) Euclid distance;
2) Frechet distance;
3) an adaptive dissimilarity index with Euclid distance and $k = 1$;
4) an adaptive dissimilarity index with Euclid distance and $k = 2$;
5) an adaptive dissimilarity index with Frechet distance and $k = 1$;
6) an adaptive dissimilarity index with Frechet distance and $k = 2$.

We can notice that Euclid distance only considers the location of observations, Frechet distance — location, order and adaptive dissimilarity index — location, order and behavior.

Four clusters were determined. We compute center and corridor for each cluster. Width of corridor is equal to maximal standard deviation in cluster.

Compare the results of clustering on Figure 1 which were obtained with Euclid distance (1) and an adaptive dissimilarity index (3)–(5) where $\delta(X, Y)$ is Euclid distance

and k is equal to 1. That means that location of objects has a higher impact on dissimilarity value than their behavior does. Centers and corridors of only two clusters are presented on the Fig. 1 for better visualization. We can see there that corridors on Fig. 1, *a* are much thicker than on Fig. 1, *b*, while only two districts change their placement. We must notice that clustering with Frechet distance (2) for our problem gives worse results as corridors are very wide and the area of cluster intersection is too large.

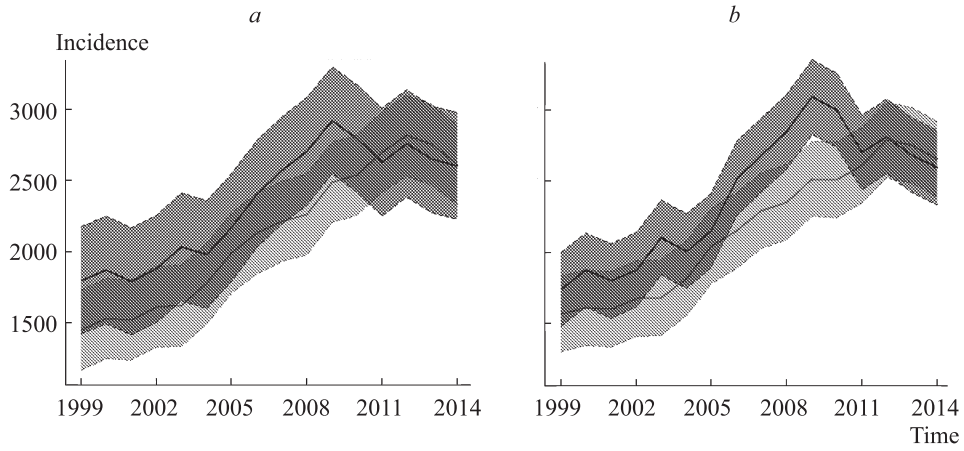

Fig. 1. Comparison of centers (—) and corridors (- - -) of two clusters obtained with Euclid distance (*a*) and adaptive dissimilarity index (*b*)

Results of clustering with adaptive dissimilarity index where $\delta(X, Y)$ is Euclid distance and k is equal to 1 are presented on Fig. 2, *a*. Every cluster has a definite color. Unfortunately, we cannot find any dependence of cluster location on geographical situation.

After clustering which was made for every dimension individually the aggregation of results had to be made, so we went ahead and utilized dissimilarity measure of match-by-dimension approach.

Match-by-dimension approach in multivariate time series clustering. Dealing with problem of multidimensionality we can consider every dimension as univariate time series and compute distances for it (as we do it in previous section). Correspondingly to this approach we have a 3-dimensional distance matrix. Weighted Borda count could be used for achieving the final result or other aggregating functions as mean, max and min could be used.

Weighted Borda count (8) has several advantages. Firstly, it considers the weight of every dimension, for example for our data it was found that children morbidity rate has a huge part of information. Secondly, the method takes into account a similarity gap while a usual Borda count does not. Thirdly, it is simple for computing and understanding.

Results of cluster analysis where Euclid distance was used as dissimilarity measure are shown on Fig. 2, *b*. Notice that Fig. 2, *a* has mapping of univariate clustering of children morbidity rate while Fig. 2, *b* shows the outcome for multivariate clustering.

The match-by-dimension approach let us utilize all the knowledge gained for univariate time series so it is simple enough. But at the same time, important correlations between dimensions could be lost because of breaking MTS into several univariate time series.

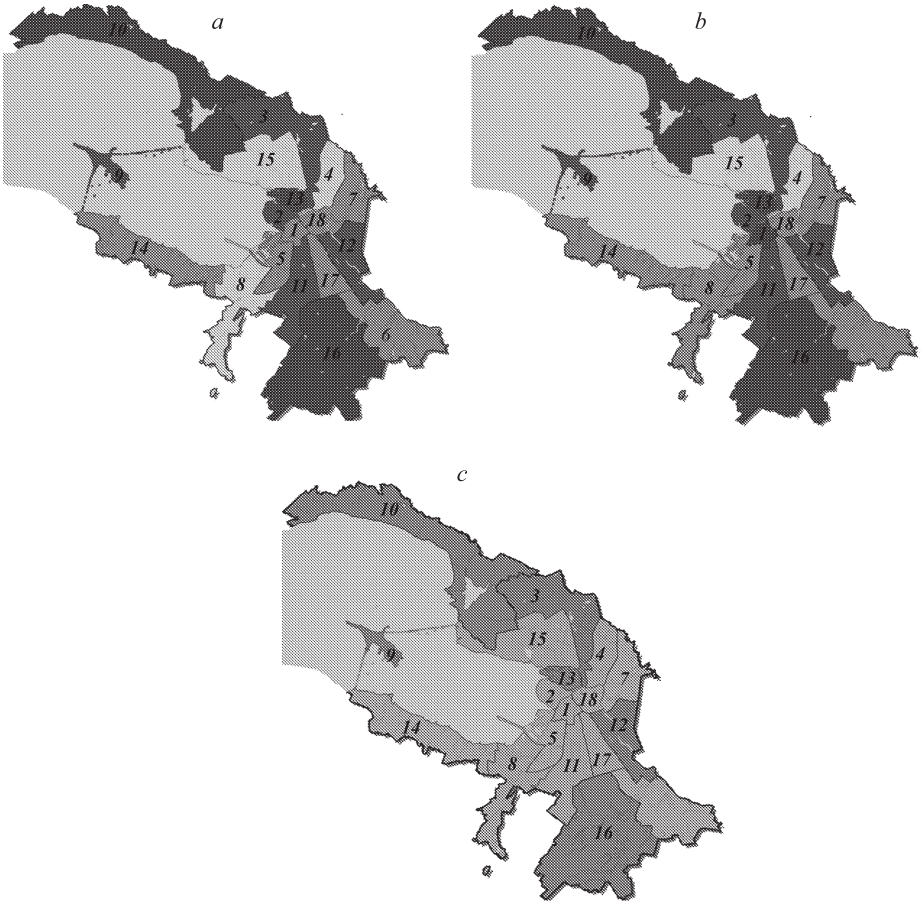

Fig. 2. Map of Saint Petersburg with results of univariate clustering (*a*), multivariate clustering (*b*) and map with stable clusters (*c*)

Districts: *1* — Admiralteysky, *2* — Vasileostrovsky, *3* — Vyborgsky, *4* — Kalininsky, *5* — Kirovsky, *6* — Kolpinsky, *7* — Krasnogvardeysky, *8* — Krasnoselsky, *9* — Kronshtadtsky, *10* — Kurortny, *11* — Moskovsky, *12* — Nevsky, *13* — Petrogradsky, *14* — Petrodvortsovy, *15* — Primorsky, *16* — Pushkinsky, *17* — Frunzensky, *18* — Tsentralny.

The overall matching approach in multivariate time series clustering. The disadvantage of the previous approach is eliminated in the overall matching methods. This approach considers MTS as a whole unit. Due to this fact correlations between dimensions are saved. The disadvantage of the method is a curse of dimensionality that is why all overall matching methods use some techniques to reduce data size [10].

For our problem to be resolved Eros distance metric (7) was chosen but the output was not so elegant as in the previous experiments. The dedrogram is too branchy and it is hard to determine clusters. Correspondingly to the dendrogram almost every cluster contains just one district while reducing the number of clusters we have a cluster that contains almost all districts. The reason of the problem could be small data size. Moreover, cluster analysis is not a strict algorithm and result of clustering highly depends on the method that is chosen by a researcher. Some methods suit to a problem while the others do not.

Conclusions. We have several mappings for different dissimilarity measures that we have used. As districts change their cluster depending on the way we compute distance we cannot make firm conclusions about exact structure of each cluster. We cannot claim that one mapping is more accurate than another one as we do not possess any information about the real structure. But we can determine stable clusters and conclude that these clusters are saved in different experiments and possibly districts from different clusters might really have some inequality in the dynamic of morbidity rate.

Three stable clusters are presented on Fig. 2, *c* where districts *1, 2, 4, 8, 11, 15, 17* do not belong to any stable cluster. Problems which can cause the differences in morbidity rate can be connected with a very rapid development of the district, the ageing of its inhabitants, poor organization of healthcare system etc.

It is necessary to develop new algorithms for multivariate time series clustering which would consider the behavior of time series as an adaptive dissimilarity index. Borda count showed good results for our problem but we could not make adequate interpretation of the results obtained with Eros method which takes into account correlations between dimensions. That is why for future research it would be great to construct such measure that would deal well with small data and consider MTS as a whole unit.

References

1. Lozkins A., Bure V. M. Veroyatnostniy podhod k opredeleniu lokalno-optimalnogo chisla klasterov [The probabilistic method of finding the local-optimum of clustering]. *Vestnik of Saint Petersburg University. Series 10. Applied mathematics. Computer science. Control processes*, 2016, issue 1, pp. 28–38. (In Russian)

2. Aldenderfer M. S., Blashfield R. K. Klasternyi analiz [Cluster analysis]. *Factornyi, diskriminantnyi i klasternyi analiz* [*Cluster, factor, discriminant function analysis*]. Pod red. I. S. Enukova. Moskow, Finansi i statistica Publ., 1989, 215 p. (In Russian)

3. Dmitriev A. P., Zubriyanova N. S. Statisticheskoe izuchenie dinamiki pervichnoj zabolevaemosti naseleniya Penzenskoj oblasti [Statistical research of dinamics of morbidity rate of Penza Region population]. *Izvestiya vysshih uchebnyh zavedenij. Povolzhskij region. Medicinskie nauki* [*University proceedings. Volga region. Medical sciences*], 2008, issue 2, pp. 89–98. (In Russian)

4. Staroverova K. U. Issledovanie dinamiki zabolevaemosti v Sankt Peterburge [Research of behavior of Saint Petersburg morbidity rate]. *Rezultatyi nauchnyih issledovaniy: sbornik statey mezhdunarodnoy nauchno-prakticheskoy konferentsii* [*Scientific research results: proceedings of the Intern. scientific and research conference*]. Tyumen, Aeterna Publ., 2016, pp. 11–14. (In Russian)

5. Staroverova K. U. Klasterizatsiya vremennyih ryadov s ispolzovaniem R [Time series clustering in R]. *Protsessyi upravleniya i ustoychivost* [*Control processes and stability*], 2016, vol. 3, issue 1, pp. 317–323. (In Russian)

6. Montero P., Vilar J. TSclust: An R package for time series clustering. *Journal of Statistical Software*, 2015, no. 62.1, pp. 1–43.

7. Douzal Chouakria A., Nagabhushan P. N. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, March 2007, vol. 1, issue 1, pp. 1–43.

8. Yang K., Shahabi C. A PCA-based similarity measure for multivariate time series. *MMDB '04 Proceedings of the 2nd ACM Intern. workshop on Multimedia databases*, 2004, pp. 65–74.

9. Li S. J. , Zhu Y. L., Zhang X. H., Wan D. BORDA counting method based similarity analysis of multivariate hydrological time series. *Journal of Hydraulic Engineering*, 2009, vol. 40, no. 3, pp. 378–384.

10. Wang J., Zhu Y., Li S., Wang D., Zhang P. Multivariate time series similarity searching. *The Scientific World Journal*, 2014, vol. 2014, article ID 851017, 8 p.

For citation: Bure V. M., Staroverova K. U. Applying clustering analysis for discovering time series heterogeneity using Saint Petersburg morbidity rate as an illustration. *Vestnik of Saint Petersburg University. Series 10. Applied mathematics. Computer science. Control processes*, 2016, issue 4, pp. 44–50. DOI: 10.21638/11701/spbu10.2016.404

Статья рекомендована к печати проф. Л. А. Петросяном.

Статья поступила в редакцию 17 августа 2016 г.

Статья принята к печати 29 сентября 2016 г.