sciendo

APPLICATION OF MULTIVARIATE TIME SERIES CLUSTER ANALYSIS TO REGIONAL SOCIO-ECONOMIC INDICATORS OF MUNICIPALITIES

Valentas Gružauskas
Kaunas University of Technology
Digitalization Scientific Group
e-mail: valentas.gruzauskas@ktu.lt

Dalia Čalnerytė
Kaunas University of Technology
e-mail: dalia.calneryte@ktu.lt

Tautvydas Fyleris
Kaunas University of Technology
e-mail: tautvydas.fyleris@ktu.lt

Andrius Kriščiūnas
Kaunas University of Technology
e-mail: andrius.krisciunas@ktu.lt

Abstract

The socio-economic development of municipalities is defined by a set of indicators in a period of interest and can be analyzed as a multivariate time series. It is important to know which municipalities have similar socio-economic development trends when recommendations for policy makers are provided or datasets for real estate and insurance price evaluations are expanded. Usually, key indicators are derived from expert experience, however this publication implements a statistical approach to identify key trends. Unsupervised machine learning was performed by employing K-means clusterization and principal component analysis for a dataset of multivariate time series. After 100 runs, the result with minimal summing error was analyzed as the final clusterization. The dataset represented various socio-economic indicators in municipalities of Lithuania in the period from 2006 to 2018. The significant differences were noticed for the indicators of municipalities in the cluster which contained the 4 largest cities of Lithuania, and another one containing 3 districts of the 3 largest cities. A robust approach is proposed in this article, when identifying socio-economic differences between regions where real estate is allocated. For example, the evaluated distance matrix can be used for adjustment coefficients when applying the comparative method for real estate valuation.

Key words: *regional policy recommendations, machine learning, multivariate time series cluster analysis.*

JEL Classification: *R11, C38.*

Citation: Gružauskas, V., Čalnerytė, D., Fyleris, T., & Kriščiūnas, A. (2021). Application of multivariate time series cluster analysis to regional socio-economic indicators of municipalities. *Real Estate Management and Valuation, 29*(3), 39-51.

DOI: https://doi.org/10.2478/remav-2021-0020

1. Introduction

The socio-economic development of a country, municipality or region can be defined by a set of indicators. Although the socio-economic situation is usually evaluated annually according to the indicator values in the respective year, the variation of indicators can also be presented as multivariate time series. In the current research, the amount of analyzed data increased and consisted of indicator values, territorial units and periods. That is why machine learning methods are employed in the latest research. The multivariate time series clustering for socio-economic indicators of municipalities enables municipalities with similar development to be identified based on the relationships between the indicators and their trends.

Market trend monitoring and the quantification of a region's performance is essential in order to provide evidence-based policy recommendations or to apply adjustment coefficients when valuating real estate. The authors of this publication previously published a methodology for the comparative method for real estate valuation (Gružauskas, et al. 2020). The methodology provided in this publication provides more detailed information for the calculation of the location adjustment coefficient by using the distance matrix obtained from the clustering results. However, the proposed methodology can be applied for other reasons to analyze the real estate market. Belej and Kulesza (2014) analyzed real estate prices in Poland and quantified the high inertia of real estate markets, which manifested during rapid structural changes. Kokot (2020) analyzed the influence of socio-economic factors on housing prices in Poland and proposed a City Wealth Synthetic Measure. Usman et al. (2020) conducted an extensive literature analysis of property market segmentation into submarkets to improve the analysis of real estate prices. The publication mainly describes 3 types of methods to segment the market, i.e. predefined regions, statistically obtained regions and their combination. The reviewed methods focus mainly on spatial clustering approaches, and only briefly mention the importance of temporal analysis, thus the proposed methodological approach of our publication could provide a tool for market segmentation. Nugroho et al. (2020) analyzed regional economic growth to form regional clustering based on the speed of house price growth, so that the monetary policy in the housing sector achieves the target. However, real estate analysis is not only limited to price analysis. Asamoah et al. (2019) conducted an extensive literature analysis to determine the influence of economic indicators which are important for players in the construction industry including policy makers. The determination of economic factors f the construction industry could reduce the incidence of the high failure rate of construction firms. Kazak et al. (2017) analyzed the ageing society to determine which segments of real estate should focus on older people. Thus, the proposed methodological approach in the publication could be applied to determine patterns in time series data with robust insights. The proposed methodological approach in our publication could also help to identify other trends and form indicators of the regions. One of the major concerns in today's world is sustainability, and thus various indexes have been developed to analyze this phenomenon in the regions. For example, Manzhynski et al. (2016) measured the sustainability performance level of the Baltic region by 4 types of indicators, such as Adjusted Net National Income, Adjusted Net Savings, Environmental Performance Index, Human Development Index and Sustainable Value. The Vilnius Institute of Policy Analysis developed a municipality welfare index, which combines 5 categories, i.e.: evaluations of social security, physical safety, economic level, education level and demography (Vilnius Institute of Policy Analysis, 2019). Salvati and Carlucci (2014) proposed a sustainable development indicator for Italy; the research paper combined a wide range of variables and used a statistical tool to determine the composite indicator weights. A similar technique was used by Seidel et al. (2019), who developed an indicator to measure organic agriculture. Also, Senna et al. (2019) used a statistical tool to develop a water poverty index. The main difference between the proposed indicators is the approach which was used to determine the weights of the composite indicators. One category of indicators is based on the experts' decision as to what weights and indicators should be used in the composite indicator. Another category employs statistical tools, such as principal component analysis, factor analysis, clustering analysis and others. A detailed description of the available tools has been provided by the Organisation for Economic Co-operation and Development (Mattes & Sloane, 2015). A multivariate statistical analysis was used to analyze the spatial and socio-environmental consequences of applying general spatial planning in the municipalities of Catalonia (Serra, et al. 2014). Greco et al. (2019) provided a comprehensive comparison of the main existing approaches. The goal of these indicators is to quantify the performance of regions and to provide recommendations for policy makers. The described composite indicator approach uses statistical tools to derive the indicators for measuring the region's performance. Usually, it is experts who propose approaches to measure performance. Nowadays a large amount of data is used to describe the performance of the region and machine learning can be applied to obtain insights regarding the economic situation in the region. Einav and Levin (2013) analyze the value of big data and predictive modelling tools. They stated that large-scale administrative datasets and proprietary private sector data can greatly improve the measurement, tracking and description of economic activity. Athey and Luca (2019) indicated that technology companies employ economists more often than ever. This also results in developing machine learning approaches for public policy. In this publication, an example was provided to demonstrate how economists used statistical data to help explain gentrification in neighborhoods. Athey (2019) stated that unsupervised learning can be used to create new variables without human judgement in economic analyses.

When analyzing data on a regional level, the unsupervised learning approach could be used to identify similar regions and obtain insights which explain their performance levels. Municipality clusterization can be applied to predict the demand of economic stimulation, to identify similar social problems, to monitor social and economic development, to plan logistics, to extend the datasets of similar objects in realty price evaluation, insurance costs, etc. Cluster analysis of municipalities based on social and economic development indicators can be applied to identify the regions which are in the highest need of economic stimulation (Brauksa, 2013). Brauksa (2013) stated that one of the principal factors of the same group is the region the municipality is located in. A factorial and component analysis with hierarchical clustering was used to group the municipalities of Lithuania based on their socio-economic situation, in order to identify municipalities which are the most attractive for the foreign investment (Burinskienė & Rudzkiene, 2004). The K-means algorithm was applied to group municipalities of Slovenia in order to examine social-economic differences among municipalities (Rovan & Sambt, 2003). The clusterization indicated significant differences in socio-economic development and its result was one of the criteria for the approval of project funds. The comparison of regions in Visegrad Group Plus countries was performed according to the Human Development Index in (Majerova & Nevima, 2017). A combination of regression analysis, Ward and K-means methods was performed to cluster the regional labor and vocational training market in Germany (Kleinert, et al., 2018; Blien, et al. 2010). Rezankova (2014) applied a clustering algorithm on data of enterprise, macroeconomic and economic activity by age groups to European countries, including Lithuania. Majerova and Nevima (2017) applied a hierarchical clustering method by measuring the distance between the clusters as the squared Euclidian distance and Ward method on the human development index with a z-score normalization technique on the data. Augustysnski and Laskos-Grabowski (2018) compared various clustering algorithms and determined that the best results were achieved on their dataset by applying a compression-based dissimilarity measure. In most research, clusterization was performed for the indicators of a specific year. In the second step, the change between the clusters is usually analyzed (Burinskienė & Rudzkiene, 2004; Majerova & Nevima, 2017). The proposed approach enables clusters of similar municipalities to be determined with respect to the time series of socio-economic indicators and the relationships between the indicators.

In cluster analysis, the result depends on the applied method and the parameters that are considered in the model. Although different methods often give different results, even the same method can give different results in different runs. For example, the results of clusterization using K-means depend on the random initial distribution. However, in the clustering of municipalities, the aim is to find similar development trends between municipalities and a slightly different clusterization result is not essential. Although some research analyzes the development of the region, clusterization is usually performed for the annual data, and then the clustering results of different years are compared. In this paper, we consider the development of different municipalities based on the change of demographic and economic indicators provided as time series.

The aim of this paper is to demonstrate an application of the multi-variate time series clustering algorithm for real-life data. The data consisted of time series of social and economic indicators of the municipalities of Lithuania in the period between 2006 and 2018. The algorithm for multi-variate time series clusterization was presented in (Li, 2019) and includes common principal component analysis, multivariate time series analysis and k-means clusterization.

2. Data and Methods

2.1. Data

The multivariate time series are analyzed in various fields of the economy, including forecasting stock prices and interest rates, the development of regions, etc. Therefore, the general concept of multivariate time series is discussed in this section. The dataset **X** consists of N multivariate time series:

$$\boldsymbol{X} = \{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^N\} \tag{1}$$

here the element $\mathbf{X}^k$ represents the k-th object and defines the change of its parameters in time. $\mathbf{X}^k$ is a matrix of size $n_k xm$, here n_k is the length of $\mathbf{X}^k$ and m is the number of variables:

$$\mathbf{X}^k = \begin{bmatrix} x_{1,1}^k & \cdots & x_{1,m}^k \\ \vdots & \cdots & \vdots \\ x_{n_k,1}^k & \cdots & x_{n_k,m}^k \end{bmatrix} \tag{2}$$

here x_{ij}^k is the i-th element of the j-th variable of the k-th object.

As the objects of the same type are analyzed, they usually have the same variables which describe the development of the object. The length of the time series can be different for the object as objects can be observed in different intervals.

When the variables in time series represent values of various origins, they can obtain significantly different values. In order to use different variables with the same weight in the calculations, values of variables are standardized according to the following formula:

$$z_{ij}^k = \frac{x_{ij}^k - \mu_j}{s_j} \tag{3}$$

here z_{ij}^k is the i-th standardized item of the j-th variable of the k-th object, μ_j is the mean value of the j-th variable for the set of all objects, s_j is the standard deviation of the j-th variable for the set of all objects. The dataset **Z** of standardized multivariate time series is considered in the time series analysis:

$$\boldsymbol{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^N\} \tag{4}$$

This enables to analyze data of different variables with the same weight, as all values of one variable are standardized to the distribution with the mean value equal to 0 and standard deviation equal to 1.

2.2. Common Principal Component Analysis

The algorithm for multivariate time-series clusterization was originally designed for the field of engineering (Li, 2019). In this publication, the algorithm was adapted to clustering time series of economic data. The algorithm consists of a combination of principal component analyses for common space and k-means clusterization. The common principal component analysis is broadly explained in (Li, 2019). Firstly, the multivariate time series $\mathbf{Z}^k$ is transformed to covariance matrix:

$$\mathbf{V}^k = cov(\mathbf{Z}^k) \tag{5}$$

here $\mathbf{V}^k$ is the covariance matrix of the k-th object. Although the normalized values of the time series are used in the calculation of the covariance matrix by itself, it should be noted that, in this case, normalization is performed only with respect to the analyzed time series. Moreover, the deviation is not changed in this normalization. In general, variables can obtain values of significantly different magnitudes. That is why the standardization of time series should be performed before common principal analysis.

Secondly, the average covariance matrix **V** is calculated according to the following formula:

$$\mathbf{V} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{V}^k \tag{6}$$

This matrix generalizes the covariance of the variables for all objects of the dataset. The averaged covariance matrix **V** is decomposed using singular value decomposition (SVD). The result of the decomposition is a sorted vector $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$ of eigenvalues and matrix $\mathbf{U} = \{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_N\}$ of the respective eigenvectors, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. The importance of information stored in the eigenvector is defined by its respective eigenvalues. The most valuable information is provided in the first p components and, therefore, they are used in the analysis. This enables to the dimensionality of the data to be reduced and the most significant information to be retained. Although there are no general criteria to define the number of principal components, it is usually chosen based on the magnitude of eigenvalues. The common space $\mathbf{S}$ is constructed of the first p eigenvectors as $\mathbf{S} = \{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_p\}$. Each multivariate time series $\mathbf{Z}^k$ can be transformed to the common space $\mathbf{S}$ by the following transformation:

$$\mathbf{P}^k = \mathbf{Z}^k\mathbf{S} \tag{7}$$

2.3. K-means Clusterization

The clusterization of multivariate time series is based on the error, which is obtained after the projection of the multivariate time series $\mathbf{Z}^k$ to the common space $\mathbf{S}$, and its reconstruction to the initial space. The resulting time series $\mathbf{Y}^k$ is obtained by the formula (Li, 2019):

$$\mathbf{Y}^k = \mathbf{Z}^k\mathbf{S}\mathbf{S}^\mathbf{T} \tag{8}$$

here $\mathbf{Z}^k$ is the time series of the k-th object after standardization, $\mathbf{S}$ is the common space defined by the projection axes. As only the first p components are used in the analysis, the initial and resulting time series are not equal and the reconstruction error E^k of the k-th object is calculated as the Frobenius norm of the difference between the time series $\mathbf{Z}^k$ and $\mathbf{Y}^k$:

$$\mathrm{E}^k = \|\mathbf{Z}^k - \mathbf{Y}^k\| = \sqrt{\sum_{i=1}^{n_k}\sum_{j=1}^{m}\left(z_{ij}^k - y_{ij}^k\right)^2} \tag{9}$$

Here, n_k is the length of $\mathbf{Z}^k$ and m is the number of variables in the time series. Obviously, the magnitude of the error depends on the number of principal components. If all the components are used in the analysis ($p = N$), no information about the object is lost during transformation and the error is generated only because of numerical calculations; it is therefore close to 0. Similarly, if some variables are collinear in the multivariate time series, the respective eigenvalues can be close to 0. The elimination of such eigenvalues results in the insignificant loss of information and a small error after transformation and reconstruction. The error increases as the number of principal components decreases because more information is eliminated from the analysis.

The number of clusters N_C is chosen empirically based on the total error generated by the clusterization using different numbers of clusters. In the initial step, the set $\mathbf{C} = \{C_1, \dots, C_{N_C}\}$ of clusters is formed by assigning the objects to cluster C_i randomly. The common space of the i-th cluster $\mathbf{S}_i$ is constructed with respect to all objects assigned to the i-th cluster and defines the centroid of the cluster. The construction of this common space is described in the previous chapter. For all objects, the reconstruction error is calculated for all centroids of clusters. The object is assigned to the cluster for which the minimal reconstruction error is determined. Calculations are performed until clusters do not change, or iteration count is less than the maximum number of iterations I_{max}. In order to reduce the effect of random initialization, clusterization is performed for the determined number of runs and the result with the smallest summing error for all clusters is taken as the final result. It should be noted that a small number p of the principal components used in the analysis results in a relatively large error after projection to the common space and reconstruction. This fact may lead to situations where the only object of the cluster has a smaller reconstruction error calculated with respect to the centroid of the other cluster, and should be assigned to this cluster in the following iteration.

3. Empirical results

3.1. Data

Lithuania is divided into 60 municipalities. Some of them cover cities (Vilnius, Kaunas, Klaipeda, etc.), with their main characteristic being a high population density. These municipalities are surrounded by regional municipalities including the suburbs of the cities. Obviously, the development of the cities has an impact on the development of the regional municipalities around them. In other cases, municipalities cover average-sized towns, with respect to their population and surroundings.

In total, annual values of 27 indicators covering the period from 2006 to 2018 were used in this research. The indicators can be classified into 3 categories, such as demographic, housing and economic indicators. The main demographic indicators consist of population size, population density, ageing index, births and deaths. Additionally, indicators which describe the movement of residents are added. The additional indicators consist of newcomers, emigrants, immigrants, departures, net migration, net inside migration, and their combination. These indicators are important to describe the population tendency and needs. For example, people of a specific age group have a tendency to go to university, to rent flats, to purchase housing, to travel and so on. Thus, housing indicators are important when analyzing the change of demographic indicators in the region. The housing indicators consist of number of dwellings, number of houses, number of none-resident estate, residential fund, useful area per person and average dwelling size. The last category is economic indicators, which consist of the number of unemployed people, number of employees, employment rate, ratio of unemployed to those of an active working age, monthly wage, direct foreign investments, municipality income and municipality expenditures. The economic indicators are important to describe the potential of income and employment in the region. The integration of population, housing and economic indicators provides a better understatement of the region's social-economic tendencies.

In order to use different indicators with the same weight in the calculations, values of indicators are standardized according to Eq. *3*. To maintain the tendencies of indicators for the analyzed period, the mean value and standard deviation are calculated for all municipalities and years of the analyzed period. This standardization enables the change of indicators which have values of significantly different magnitudes or are given in percentages to be compared, as the values are standardized to a distribution with the mean value of 0 and standard deviation of 1.

3.2. Determining Number of Common Principal Components for the Analysis

Clusterization results depend on the number of the common principal components used in the analysis. To extract the most important features of the time series, only several first principal components should be used. The eigenvalue analysis has been used to determine the number of the components which should be used in the analysis. The eigenvalues of the covariance matrix for all municipalities in one cluster is provided in Fig. 1. The eigenvector of the largest eigenvalue accounts for the axis with widest distribution among data. Small values determine that the data is concentrated in this axis.

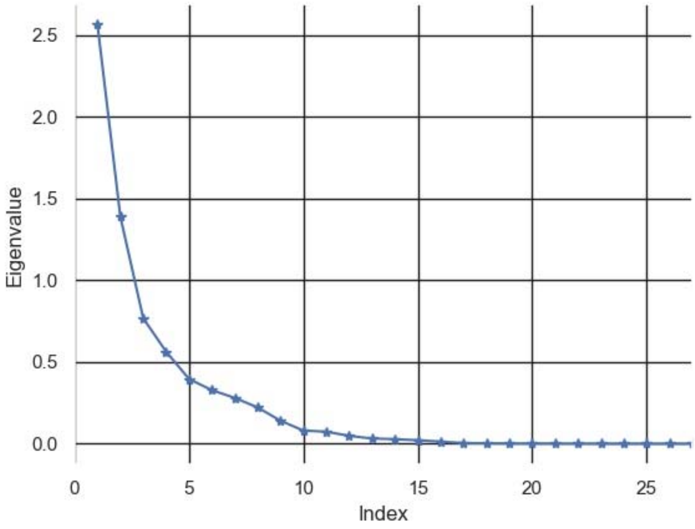


Fig. 1. Eigenvalues of the common space of all municipalities. *Source*: own study.

In the numerical example, 3 first principal components are employed. Obviously, these components can represent different features in different clusters. However, this is one of the advantages of the common principal analysis, as it combines objects with features which are typical for the same cluster.

3.3. Clusterization Results

To determine the number of clusters which should be applied in the determination of distinct clusters, clusterization was performed for various numbers of clusters. The total error was calculated as the sum distance of the objects to their centroid. It should be noted that, contrary to the standard K-means method, if a cluster consists of only one municipality, the distance to the cluster center (error) can be greater than zero. This is due to the fact that only the specific number of principal components is used in the transformation and reconstruction of the multivariate time series and some information can be lost resulting in the error. The total error was calculated after 100 iterations or if the cluster assignments do not change. Obviously, the dependency of the total error on the number of clusters can change if another initial distribution is used in the initialization step of K-means. That is why 10 cases were performed to determine the trend of dependency of total error on the number of clusters. The errors of individual runs are provided in blue (Fig. 2); the black curve represents the mean value of the error which was calculated for the respective number of clusters. The 5 clusters were used in successive calculations as the value of the averaged total error demonstrates a steep fall up to this number (Fig. 2).

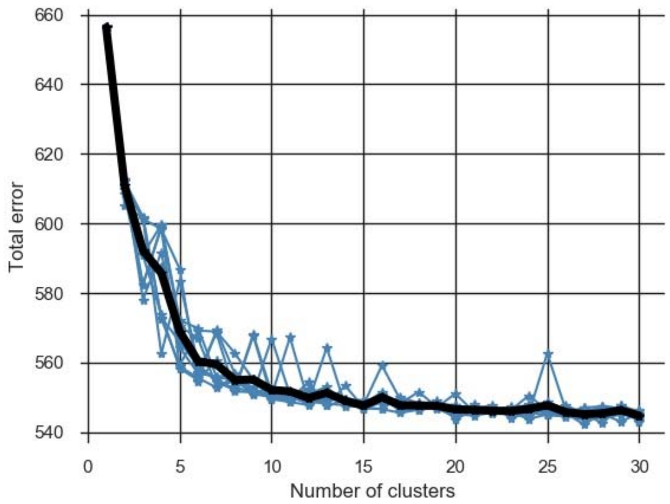
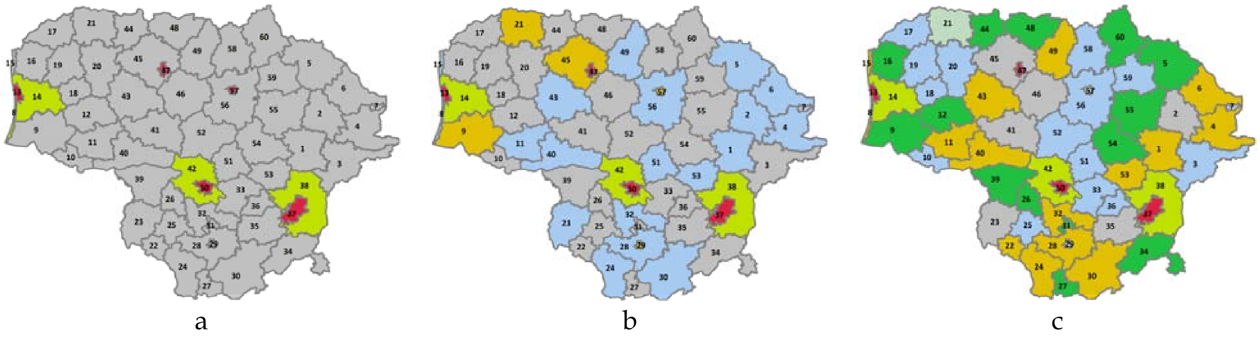


Fig. 2. Dependency of total error on number of clusters. *Source*: own study.

The clusterization results using 3, 5 and 7 clusters are shown in Fig. 3. In all three cases, the four largest cities of Lithuania (Klaipeda c. (13); Vilnius c. (37); Siauliai c. (47); Kaunas c. (50)) are assigned to one cluster. If municipalities are grouped to three clusters, the fifth largest city (Panevezys c. (57)) is also assigned to the same cluster. Similarly, the districts of the three largest cities are assigned to another cluster as they demonstrate similar development of macro-economic indicators.



a b c

Fig. 3. Clusterization result after 100 runs for 3 clusters (a), 5 clusters (b) and 7 clusters (c). *Source*: own study.

The lists of municipalities assigned to clusters are presented in Table 1 if the number of clusters is equal to 5. The numbers in the brackets next to the title refer to the municipality identity number in Fig. 3.

Table 1

Municipality groups after clusterization to 5 clusters

Cluster 1 (red)	Klaipeda c. (13); Vilnius c. (37); Siauliai c. (47); Kaunas c. (50)
Cluster 2 (green)	Klaipeda d. (14); Vilnius d. (38); Kaunas d. (42)
Cluster 3 (blue)	Neringa c. (8); Jurbarkas d. (40); Ignalina d. (4); Taurage d. (11); Prienai d. (32); Kelme d. (43); Pakruojis d. (49); Palanga c. (15); Lazdijai d. (24); Varena d. (30); Sirvintos d. (53); Moletai d. (1); Zarasai d. (6); Alytus d. (28); Rokiskis d. (5); Vilkaviskis d. (23); Jonava d. (51); Utena d. (2); Panevezys d. (56)
Cluster 4 (grey)	Svencionys d. (3); Rietavas c. (18); Kalvarija d. (22); Kaisiadorys d. (33); Pagegiai d. (10); Silale d. (12); Kretinga d. (16); Skuodas d. (17); Plunge d. (19); Telsiai d. (20); Marijampole d. (25); Kazlu Ruda d. (26); Birstonas d. (31); Salcininkai d. (34); Sakiai d. (39); Akmene d. (44); Joniskis d. (48); Kedainiai d. (52); Ukmerge d. (54); Anyksciai d. (55); Pasvalys d. (58); Kupiskis d. (59); Birzai d. (60); Visaginas c. (7); Druskininkai c. (27); Elektrenai d. (36); Raseiniai d. (41); Radviliskis d. (46); Trakai d. (35)
Cluster 5 (orange)	Mazeikiai d. (21); Panevezys c. (57); Alytus c. (29); Siauliai d. (45); Silute d. (9)

**d – district, c – city*

Source: own study.

The distances between the centroids of 5 clusters are given as the distance matrix in Fig. 4.The centroids of the first two clusters (cluster of largest cities and cluster of their districts) significantly differ from the remaining three clusters which represent the remaining municipalities. It is also worth noticing that the in-between distances between clusters 3, 4, 5 are smaller than the in-between distances between clusters 1 and 2.

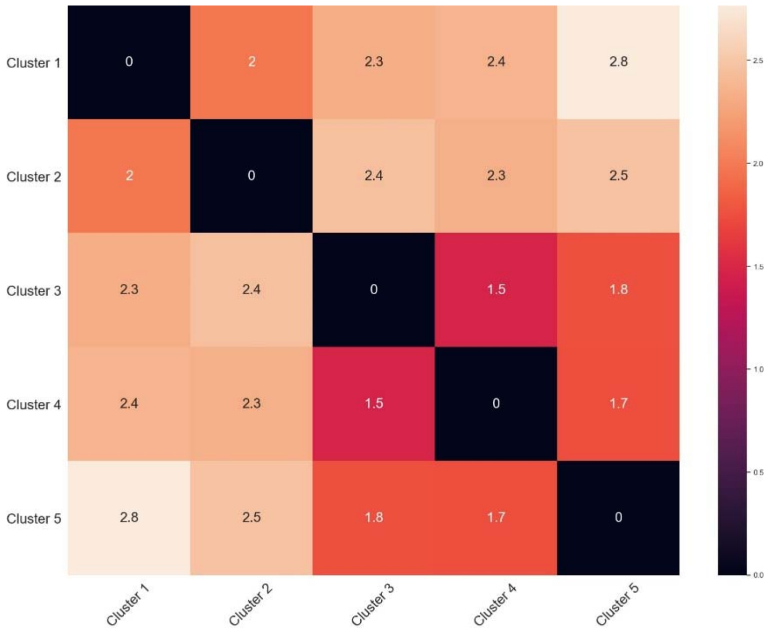


Fig. 4. Distance matrix which represents the distance between the centroids of the clusters. *Source*: own study.

In order to analyze influence of the different groups of macro-indicators, clusterization to 5 clusters using separate groups of economic, housing and demographic indicators was performed. As in the previous clusterization, 3 principal components were used in the analysis. The results of the clusterization have been presented in Fig. 5. All results show that the two largest cities (Vilnius c. (37)

and Kaunas c. (50)) are grouped to the same cluster. For clusterization based on economic indicators, the third largest city (Klaipėda c. (13)) is also included in this cluster. Clustering based on the demographic results groups the four largest cities into the same cluster. It is worth noticing that clusterization based on the demographic results groups the municipalities into 4 clusters, although 5 clusters were used as the initial number of clusters. Besides the cluster of the four largest cities, there is also a cluster of two smaller cities (Panevezys c. (57) and Alytus c. (29)). The remaining two clusters represent rural areas and smaller cities.

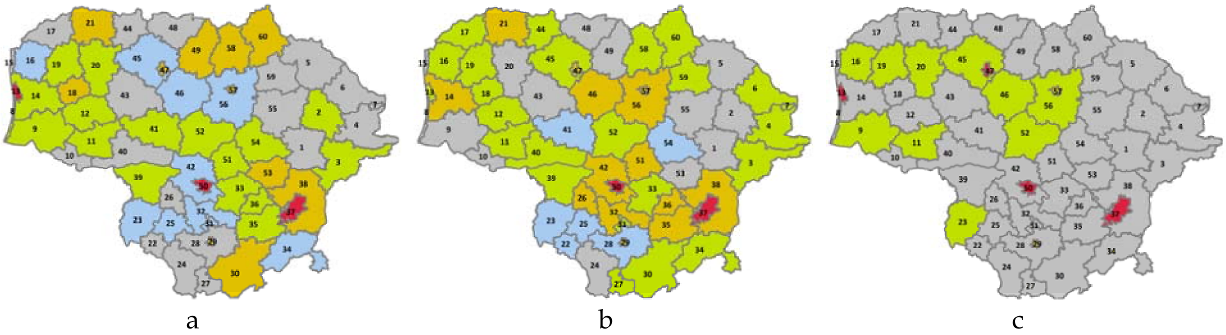


Fig. 5. Clusterization result after 100 runs for 5 if clusterization is performed using only economic indicators (a), housing indicators (b) and demographic indicators (c). *Source*: own study.

Similar tendencies of differentiating between cities and rural areas were observed for all groups of indicators. As an example to represent the dynamics of indicators, mean, minimum and maximum values of monthly wages in clusters obtained for economic indicators have been presented in Table 2. The colors correspond to the cluster colors in Fig. 5 b.

Table 2

Mean, min. and max. values of the monthly wages for the clusters grouped by economic indicators

		Cluster [1] [red]	Cluster [2] [orange]	Cluster [3] [green]	Cluster [4] [blue]	Cluster [5] [grey]
2006	Mean	473.8	362.8	354.1	321.9	348.1
	Min	428.1	310.8	305.3	300.9	301.5
	Max	525.1	484.5	462.8	361.7	579.5
2007	Mean	568.2	441.5	431.4	390.2	413.2
	Min	521.9	379.7	369.0	366.9	364.9
	Max	626.4	567.9	554.6	434.4	636.0
2008	Mean	672.5	534.2	534.4	476.3	507.3
	Min	618.6	465.4	457.6	442.5	455.9
	Max	735.6	633.7	688.1	522.5	677.4
2009	Mean	644.9	511.8	508.7	469.6	496.4
	Min	591.7	465.7	449.5	443.4	447.2
	Max	698.9	603.9	587.6	510.6	649.9
2010	Mean	626.1	511.8	491.1	448.1	475.7
	Min	568.5	442.5	437.9	411.3	422.0
	Max	684.7	573.4	576.6	501.6	607.9
2011	Mean	643.8	506.2	506.6	454.3	486.1
	Min	588.2	457.9	450.4	409.8	415.9
	Max	704.1	584.7	605.6	506.5	624.7
2012	Mean	666.0	521.4	522.6	471.2	499.7
	Min	607.3	474.4	462.5	427.8	442.2
	Max	731.6	609.4	626.2	541.6	657.4
2013	Mean	696.7	549.5	551.2	499.6	527.4
	Min	645.3	500.2	498.1	458.2	463.1
	Max	759.7	642.1	669.3	573.2	691.6
2014	Mean	733.3	572.3	574.0	520.8	548.2

	Min	679.8	519.9	523.3	470.4	487.3
	Max	797.6	645.6	684.2	583.4	716.3
2015	Mean	768.5	609.0	610.5	555.1	578.2
	Min	713.9	537.5	552.4	498.5	516.4
	Max	832.4	689.5	726.7	638.5	761.8
2016	Mean	829.6	662.0	669.2	617.6	624.2
	Min	780.1	598.0	610.7	560.7	566.7
	Max	890.4	737.8	777.2	725.5	792.9
2017	Mean	900.5	723.5	723.8	669.4	671.0
	Min	854.5	650.6	648.2	609.7	594.9
	Max	968.7	809.6	823.4	785.8	826.8
2018	Mean	993.8	780.2	791.7	730.4	728.6
	Min	941.5	677.6	705.2	663.6	641.7
	Max	1069.5	863.2	910.1	861.2	885.3

Source: own study.

In order to represent the dynamics of monthly wages in different clusters, the dynamics of mean values are provided in Fig. 6. Although all curves show similar tendencies (growth until 2008, plateau or small decrement in the period of 2008-2011 and growth again in the later years), this also demonstrates the significant gap between the first cluster of large cities and other municipalities.

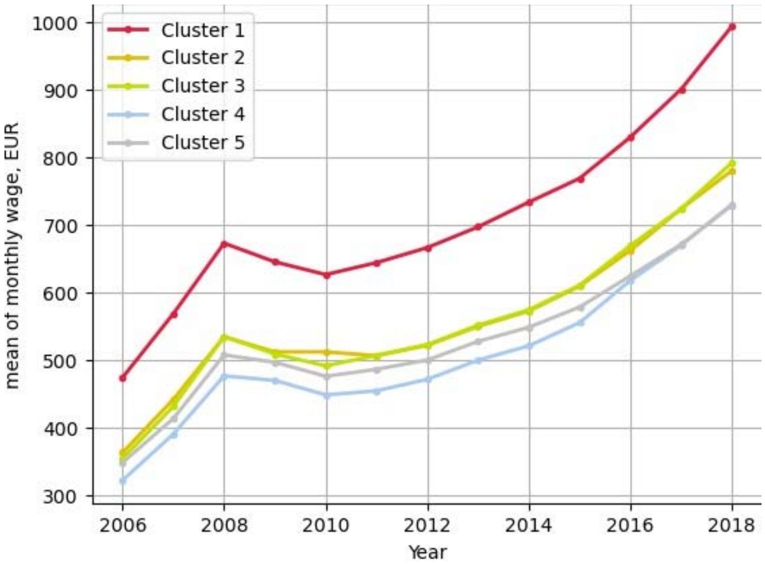


Fig. 6. Dynamics of mean monthly wage in different clusters. *Source*: own study.

Similar research based on the capability to attract foreign investment in municipalities of Lithuania for the period from 1996 to 2001 was provided in (Burinskienė & Rudzkiene, 2004). The results showed that the development of municipalities depends on the location, thus the distance to the largest cities. Moreover, the socio-economic situation is similar in most municipalities of Lithuania. Although the periods and indicators differed from the ones used in this research, the results are found to be corresponding due to rather slow economic processes in the region and relationships between the indicators which define the economic situation. The results obtained in both studies show similar trends to those of the welfare index of municipalities for 2019 published in (Vilnius Institute of Policy Analysis, 2019). However, the approach proposed in this article enables to group municipalities based not only by the indicators of one year, but also by the change in indicators.

4. Discussion and conclusions

The time series clustering analysis requires specific trends to be analyzed in a more comprehensive manner. It can, however, be used to quickly identify similarities and differences between the regions without specific expertise. Afterwards, the insights can be used to deepen the analysis for policy makers or new investors to quickly receive basic knowledge of a new market. Additionally, the obtained distance matrix between the regions can be used as an adjustment coefficient when applying the comparative method for real estate valuation.

The proposed methodological approach could be applied in various fields to improve the decision making process. For instance, Asamoah et al. (2019) conducted an extensive literature analysis of the economic indicators influencing construction industry. The research identified 59 indicators based on literature analysis. The provided methodological approach in our publication could be applied to the dataset to help identify dependency between the time series data. Nugroho et al. (2020) analyzed housing prices of Indonesia and focused on multiple regions and the growth rate of housing prices. Their research focused mainly on the analysis of 4 growth types of housing price; however, based on the presented methodology, it seems that the types were determined based on the analytical and/or expert approach, and not statistical analysis. The presented approach of time-series clustering may be used to provide a more robust analysis approach, which would help to quantify the housing price growth more precisely. Kokot (2020) analyzed the influence of socio-economic factors on the housing market in Poland. Firstly, a correlation analysis was carried out in the publication; afterwards, the variables were normalized and the k-mean clustering method used. The applied clustering analysis focused on correlation indicators and not the actual indicators as in a time series. Thus, the application of the proposed methodological approach in our publication could provide a more in-depth analysis of socio-economic factors.

In the case of our publication, the obtained clusterization results showed that there is a distinct cluster of largest cities of Lithuania. This cluster is clearly defined in all results obtained for economic, housing and demographic indicators. Another cluster is extracted for regions which are close to the largest cities. It should be noted that differences between other clusters are not significant. However, some tendencies can also be identified; for example, whether the economy of municipalities in the same cluster is based on agriculture, resort activities, etc. The main limitation of the proposed method is sensitivity to the number of clusters and number of principal components used in the analysis. As these parameters must be selected in advance, careful analysis must be performed on what values are appropriate for the analysis. Several future research areas can be considered in order to improve the practical application of the proposed algorithm. The calculation of the common space for each cluster enables the most important information for the objects of the cluster to be obtained and the respective variables combined. For different clusters, the most significant variables can differ. For future research, the number of principal components can be chosen for each cluster individually as the number of eigenvalues which were calculated for the covariation matrix of the cluster objects and exceed a determined threshold value.

In the future, additional types of indicators such parameters for agriculture, sustainable energy consumption and other areas can be integrated in the time series clustering process. The sustainability aspects of the region could better explain the performance of social-economic change in the regions. The publication proposed a way to cluster municipalities and to derive initial information about them without the involvement of experts. The proposed approach can be used by different institutes, which can integrate other types of variables and derive a precise indicator to quantify the performance level of the regions. Afterwards, a classification algorithm may be used to estimate the tendency of the region based on the new time series data. The proposed methodology can be used by policy makers to improve the decision-making process and to improve the performance measurement process of the recommended policies. Alternately, the proposed approach can be used to quantify differences between regions when applying the comparative method for real estate valuation. In this case, the data set consisted of socio-economic indicators, though it can be supplemented with more precise risk indicators related to the business or real estate sector.

In conclusion, the approach proposed in this article may be an example of multivariate time series analysis in economics. This makes it possible to obtain clusters which are based not only on the current situation but on the changes of the socio-economic indicators in the analyzed period to be constructed. The approach of clustering municipalities with respect to the series of socio-economic indicators can be applied in various areas, including but not limited to developing pilot projects in order to stimulate social welfare and sharing good practices between the municipalities of the same cluster and expanding the dataset which is used to evaluate real estate or insurance prices.

Funding

This research was supported by the Research, Development and Innovation Fund of Kaunas University of Technology (PP-91L/19).

Conflicts of interest

The authors declare that they have no conflict of interest.

Availability of data and material

All data used in this research was published as open data in Official Statistics Portal (https://osp.stat.gov.lt/), Statistics Lithuania.

5. References

Asamoah, R. O., Baiden, B. K., Nani, G., & Kissi, E. (2019). Review of exogenous economic indicators influencing construction industry. *Advances in Civil Engineering, 2019*, 1–8. Advance online publication. https://doi.org/10.1155/2019/6073289

Athey, S. (2019). The Impact of Machine Learning on Economics. *The Economics of Artificial Intelligence: An Agenda*,. 548–551.

Athey, S., & Luca, M. (2019). Economists (and economics) in tech companies. *The Journal of Economic Perspectives, 33*(1), 209–230. https://doi.org/10.1257/jep.33.1.209

Augustyński, I., & Laskoś-Grabowski, P. (2018). Clustering macroeconomic time series. econometrics, 22(2), 74–88. https://doi.org/10.15611/eada.2018.2.06

Blien, U., Hirschenauer, F., & Thi Hong Van, P. (2010). Classification of regional labour markets for purposes of labour market policy. *Papers in Regional Science, 89*(4), 859–880. https://doi.org/10.1111/j.1435-5957.2010.00331.x

Brauksa, I. (2013). Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia. *Journal of Economics, Business and Management, 1*(1), 42–45. https://doi.org/10.7763/JOEBM.2013.V1.10

Burinskienė, M., & Rudzkiene, V. (2004). Comparison of spatial-temporal regional development and sustainable development strategy in Lithuania. *International Journal of Strategic Property Management, 8*(3), 163–176. https://doi.org/10.3846/1648715X.2004.9637515

Einav, L., & Levin, J. (2013). The Data Revolution and Economic Analysis. In *NBER Working Paper 53*. https://doi.org/10.3386/w19035

Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. *Social Indicators Research, 141*(1), 61–94. https://doi.org/10.1007/s11205-017-1832-9

Gružauskas, V., Kriščiūnas, A., Čalnerytė, D., & Navickas, V. (2020). Analytical Method for Correction Coefficient Determination for Applying Comparative Method for Real Estate Valuation. *Real Estate Management and Valuation, 28*(2), 52–62. https://doi.org/10.1515/remav-2020-0015

Kazak, J., van Hoof, J., Świąder, M., & Szewrański, S. (2017). Real estate for the ageing society–the perspective of a new market. *Real Estate Management and Valuation, 25*(4), 13–24. https://doi.org/10.1515/remav-2017-0026

Kleinert, C., Vosseler, A., & Blien, U. (2018). Classifying vocational training markets. *The Annals of Regional Science, 61*(1), 31–48. https://doi.org/10.1007/s00168-017-0856-z

Kokot, S. (2020). Socio-Economic Factors as a Criterion for the Classification of Housing Markets in Selected Cities in Poland. *Real Estate Management and Valuation, 28*(3), 77–90. https://doi.org/10.1515/remav-2020-0025

Li, H. (2019). Multivariate time series clustering based on common principal component analysis. *Neurocomputing, 349*, 239–247. https://doi.org/10.1016/j.neucom.2019.03.060

Majerova, I., & Nevima, J. (2017). The measurement of human development using the ward method of cluster analysis. *Journal of International Students, 10*(2), 239–257. https://doi.org/10.14254/2071-8330.2017/10-2/17

Manzhynski, S., Siniak, N., Źróbek-Różańska, A., & Źróbek, S. (2016). Sustainability performance in the Baltic Sea Region. *Land Use Policy, 57*, 489–498. https://doi.org/10.1016/j.landusepol.2016.06.003

Mattes, M. D., & Sloane, M. A. (2015). Reflections on Hope and Its Implications for End-of-Life Care. *Journal of the American Geriatrics Society, 63*(5), 993–996. https://doi.org/10.1111/jgs.13392 PMID:25940710

Nugroho, A. A., Purnama, M. Y. I., & Fauzia, L. R. (2020). Clustering and regional growth in the housing market: Evidence from Indonesia. *Jurnal Keuangan Dan Perbankan, 24*(1), 83–94. https://doi.org/10.26905/jkdp.v24i1.3565

Řezanková, H. (2014). Cluster analysis of economic data. *Statistika, 94*(1), 73–86.

Rovan, J., & Sambt, J. (2003). Socio-economic Differences Among Slovenian Municipalities : A Cluster Analysis Approach. Developments in Applied Statistics.

Salvati, L., & Carlucci, M. (2014). A composite index of sustainable development at the local scale: Italy as a case study. *Ecological Indicators, 43*, 162–171. https://doi.org/10.1016/j.ecolind.2014.02.021

Seidel, C., Heckelei, T., & Lakner, S. (2019). Conventionalization of Organic Farms in Germany: An Empirical Investigation Based on a Composite Indicator Approach. *Sustainability (Basel), 11*(10), 2934. https://doi.org/10.3390/su11102934

de Senna, L. D., Maia, A. G., & de Medeiros, J. D. F. (2019). The use of principal component analysis for the construction of the Water Poverty Index. *RBRH (Brazilian Journal of Water Resources), 24*, *e19*, 1–14. https://doi.org/10.1590/2318-0331.241920180084

Serra, P., Vera, A., & Tulla, A. F. (2014). Spatial and Socio-environmental Dynamics of Catalan Regional Planning from a Multivariate Statistical Analysis Using 1980s and 2000s Data. *European Planning Studies, 22*(6), 1280–1300. https://doi.org/10.1080/09654313.2013.782388

Usman, H., Lizam, M., & Adekunle, M. U. (2020). Property price modelling, market segmentation and submarket classifications: A review. *Real Estate Management and Valuation, 28*(3), 24–35. https://doi.org/10.1515/remav-2020-0021

Vilnius Institute of Policy Analysis. (2019). Municipality welfare index.