INTEGRATED APPROACH TO ESTABLISH REGIONAL CLUSTERS IN LATVIA

Juris Hazners, Elita Benga
Institute of Agricultural Resources and Economics, Latvia
juris.hazners@arei.lv, elita.benga@arei.lv

Abstract. The EU Common Agricultural Policy supports the socio-economic viability of rural areas through funding and actions proposed in the second Pillar of Rural Development Programmes. One of the long-term objectives of the policy is achieving a balanced territorial development of rural economies and communities. A number of ex-ante evaluations of Latvian Rural Development programmes state that the issues related to regional aspects are addressed rather superficially. A comprehensive analysis to prioritize the areas with the highest need for support is lacking. The approach based on administrative regions would be misleading as there are marked differences between the districts in the same region. The objective of the study is the creation of regional clusters with the identification of possible outlier districts and overlap of the administrative regions to adjust the programme support promoting more balanced territorial development. An integrated methodological approach is used in a study triangulating the results obtained by K-means cluster analysis, Hierarchical Agglomerative clustering and Discriminant analysis. Five indicators are selected for the forming the clusters: Gross Value Added per capita, Territorial Development Index and share of the enterprises in Agriculture, Forestry and Food Processing sectors in total number of enterprises. The research results show that within the clusters there are outlier districts either with relatively poor performance in affluent clusters and vice versa. The regions overlap within the clusters with clusters stretching beyond the regional borders. There are marked differences in socio-economic indicators between the clusters. There is a strong negative correlation between the performance and the share of the agricultural enterprises pointing towards a necessity to strengthen the support for diversification outside the primary sector.

Keywords: rural development, K-means clustering, hierarchical agglomerative clustering, discriminant analysis.

Introduction

The EU Guidelines for the ex-ante evaluation of Rural Development Programmes (RDP) recommend the examination of the degree of budgetary consistency across territories; accordingly, those territories which have been prioritized should be weighted more highly in terms of resource allocation [1]. The assessment of regional specificities is a demanding task due to lack of regional data in many occasions. Moreover, there are marked differences in socio-economic indicators between the districts and/ or counties within the administrative regions. The first stage of the ex-ante evaluations includes the SWOT analysis. In the ex-ante evaluation of the Latvian RDP 2022-2027, about two thirds of elements of the SWOT analysis with the regional relevance are not regionally analysed [2]. Creation of district or county clusters would allow for a more targeted geographical prioritization of the programme eligibility in specific areas. The European Cluster Collaboration Platform defines clusters as regional ecosystems, as economic actors and institutions of related industries and competences with a sufficient scale to develop specialised expertise, services, resources, suppliers and skills [3]. Thus, the clusters are predefined and their denomination as business clusters would be more appropriate. Cluster analysis or clustering is a data analysis method that divides a set of objects into groups where objects in the same group are more similar to each other than to those in other groups. The number of clusters and their structure are not predefined. These two concepts are often misunderstood and improperly attributed. Often the analysis of the predefined clusters mistakenly is called cluster analysis. Research in Latvia so far mostly has been focused either on the identification of business clusters [4], or on the evaluation of business cluster policy [5]. Clustering of the administrative units is rarely applied.

Materials and methods

Clustering algorithms are a type of unsupervised machine learning used to group data based on similarities. Two methods, K-means clustering and hierarchical clustering can segment datasets into clusters. In K-means clustering clusters are distinct with their numbers assigned according to the user pre-selected total number of clusters. K-means groups data by minimizing intra-cluster variation to create compact, distinct clusters. It requires specifying number of k clusters upfront. K-means clustering is based on the iterative relocation of data points between clusters. It is used to divide the units of a dataset into non-overlapping groups, or clusters, based on the selected unit characteristics, usually a number of variables. The expected result is a pre-determined number of distinct groups of panel data

units with a high degree of similarity within each group and a low degree of similarity between groups. K-means clustering uses the Lloyd algorithm (1957) centroid model [6]. A centroid is expressed as a mean of each pre-selected variable that characterizes the unit. First, the k initial centroids are chosen based on an arbitrarily selected single variable, either randomly or by dividing the range of a specified variable into intervals. After that, iterations are done in two steps. In the first one, each case of the data set is assigned to a cluster based on its distance from the cluster's centroids, using one of the distance metrics. Usually Euclidian distance is used:

$$d(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}, \quad (1)$$

where $P = (p_1, p_2, \ldots, p_n)$ - vector P;
$Q = (q_1, q_2, \ldots, q_n)$ - vector Q;
n - number of data points;
$d(P,Q)$ - Euclidian distance between the vectors P and Q.

Among other metrics, Manhattan distance can be used:

$$d(P,Q) = \sum_{i=1}^{n} |p_i - q_i| \quad (2)$$

where $P = (p_1, p_2, \ldots, p_n)$ - vector P;
$Q = (q_1, q_2, \ldots, q_n)$ - vector Q;
n – number of data points;
$d(P,Q)$ – Manhattan distance between the vectors P and Q.

In the second step, the previous value of centroid (usually mean) is replaced by the mean of all cases assigned to the centroid in the first step. Then the sum of the squared error (SSE) is calculated as a sum of squared minimum distances between the cases. These two step iterations are repeated either until the SSE does not become lower or assigned cluster numbers for each unit do not change.

The Caliński-Harabasz index (CHI), introduced by Caliński and Harabasz in 1974 [7] is a metric for evaluating the clustering quality. The assessment is based solely on the dataset and the clustering results. CHI index is defined as the ratio of the between-cluster separation (BCSS) to the within-cluster dispersion (WCSS), normalized by their number of degrees of freedom:

$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)} \quad (3)$$

where $BCSS$ – weighted sum of squared Euclidean distances between each cluster centroid;
$WCSS$ – sum of squared Euclidean distances between the data points and cluster centroid;
k – number of clusters;
n – number of data points;
CH – Caliński-Harabasz index.

BCSS (Between-Cluster Sum of Squares) is the weighted sum of squared Euclidean distances between each cluster centroid (mean) and the overall data centroid (mean):

$$BCSS = \sum_{i=1}^{k} n_i \|c_i - c\|^2 \quad (4)$$

where n_i – number of points in cluster C_i;
c_i – centroid of C_i;
c – overall centroid of the data;
$BCSS$ – weighted sum of squared Euclidean distances between each cluster centroid.

BCSS measures how well the clusters are separated from each other (the larger the better). WCSS (Within-Cluster Sum of Squares) is the sum of squared Euclidean distances between the data points and their respective cluster centroids:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2 \quad (5)$$

where k – number of clusters;
c_i – centroid of C_i;
x – data point;
$WCSS$ – within cluster sum of squares.

WCSS measures the compactness of the clusters (the smaller the better). The objective of the centroid-based clustering is the simultaneous maximizing BCSS and minimizing WCSS.

The Elbow method, proposed by Thorndike in 1953 [8] is a heuristic graphical method for finding the optimal number of clusters in a k-means clustering algorithm. The WCSS values are plotted on the y-axis with the corresponding k values plotted on the x-axis. The optimal number of clusters is determined by forming of an elbow in the graph. The optimal number of clusters suggested by the elbow method could be retained only if the corresponding CHI index has the highest value.

In hierarchical clustering assignment of the units to clusters and number of clusters are determined by patterns detected from a Cluster Dendrogram. Hierarchical clustering creates a hierarchy of clusters organized in a tree structure based on similarity. The advantage of hierarchical clustering is the allowing for an unknown number of clusters to be identified. The hierarchical cluster analysis method that builds a hierarchy of clusters was introduced by Anderberg in 1973 [9]. There are two approaches for hierarchical clustering. Agglomerative clustering is a "bottom-up" process. Each observation starts in its own cluster, and pairs of clusters are merged moving up the hierarchy. Divisive clustering is a "top-down" approach. All observations start in one cluster, and clusters are split recursively moving down the hierarchy. Agglomerative cluster analysis of n units is defined by a stepwise algorithm which merges two data panel units at each step, the two which have the least dissimilarity. Dissimilarities between clusters of objects can be defined by the maximum dissimilarity (complete linkage), minimum dissimilarity (single linkage) or average dissimilarity (average linkage). First, the dissimilarity matrix is created. In each step, both rows and columns or columns of a matrix are clustered by merging the respective rows and columns. The process is repeated n-1 times. In every step, the size of the matrix diminishes by one. The results of a cluster analysis are presented by a binary tree, or dendrogram with n-1 nodes. Each pairing has the distance between its constituents attached.

Linear discriminant analysis is an easy to use statistical method for classifying data units into a number of groups or categories. It was first proposed by Fisher in 1936 [10]. The assignment of a unit to the specific group is done by user before the analysis. Usually, discriminant analysis is used for assessing the adequacy of a previously established group membership, as well as for predictions with respect to an expected group number of an additional data unit previously not included in a sample. Discriminant analysis is often performed after cluster analysis to confirm the correctness of estimated cluster membership. Independent variables are presented by a matrix X where k units are put in rows and their respective group variables in g columns. Dependent variable Y is presented by a column vector whose k cells equal to expected group number for each unit assigned after the calculations. Independent grouping variable G is presented by a column vector whose k cells equal to previously assigned group number for each unit. A mean corrected matrix X^o is calculated by dividing each cell in a matrix X to its respective column mean. Prior probabilities are calculated as number of units in a group divided by the total number of units:

$$p_i = \frac{n_i}{N} \quad (6)$$

where n_i – number of units in group i;
p_i – prior probability for group i;
N – total number of units.

For each group i, covariance matrix is calculated:

$$c_i = x_i^o (x_i^o)^T / n_i \quad (7)$$

where n_i – number of units in group i;
x_i^o – submatrix of matrix X^o containing units of group i;
$(x_i^o)^T$ – transposed matrix of matrix x_i^o;
c_i – covariance matrix for group i.

After that, pooled within group covariance matrix is calculated:

$$C(r,s) = \frac{1}{N}\sum_{i=1}^{s} n_i c_i(r,s) \quad (8)$$

where n_i – prior probability for group i;
x_i^o – submatrix of matrix X^o containing units of group i;
$(x_i^o)^T$ – transposed matrix of matrix x_i^o;
N – total number of units;
r – row number;
s – number of columns;
c_i – covariance matrix for group i.

Discriminant function for the group i takes the following matrix form:

$$f_i = \mu_i C^{-1} x_k^T - \frac{1}{2}\mu_i C^{-1}\mu_i^T + ln(p_i) \quad (9)$$

where p_i – prior probability for group i;
C^{-1} – inverse matrix of pooled within covariance matrix,
μ_i – vector of means in group i (average of x_i);
μ_i^T – transposed vector of means in group i (average of x_i);
C_i – pooled within covariance matrix for group i.
f_i – discriminant function.

The unit is assigned to the group that has the maximum discriminant function.

Data for the cluster parameters were retrieved from the National Statistics database [11].

Results and discussion

K-means clustering with Euclidean distance is performed with pre-selected user defined cluster numbers ranging from two to seven. The elbow plot and CHI indices are plotted in Fig. 1.

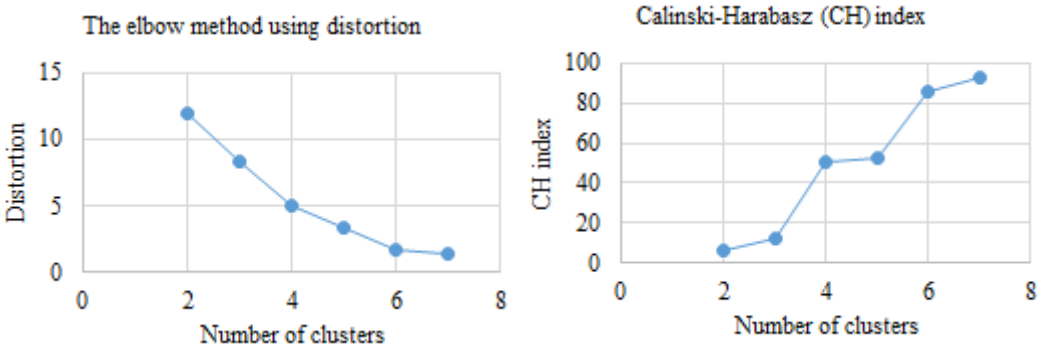
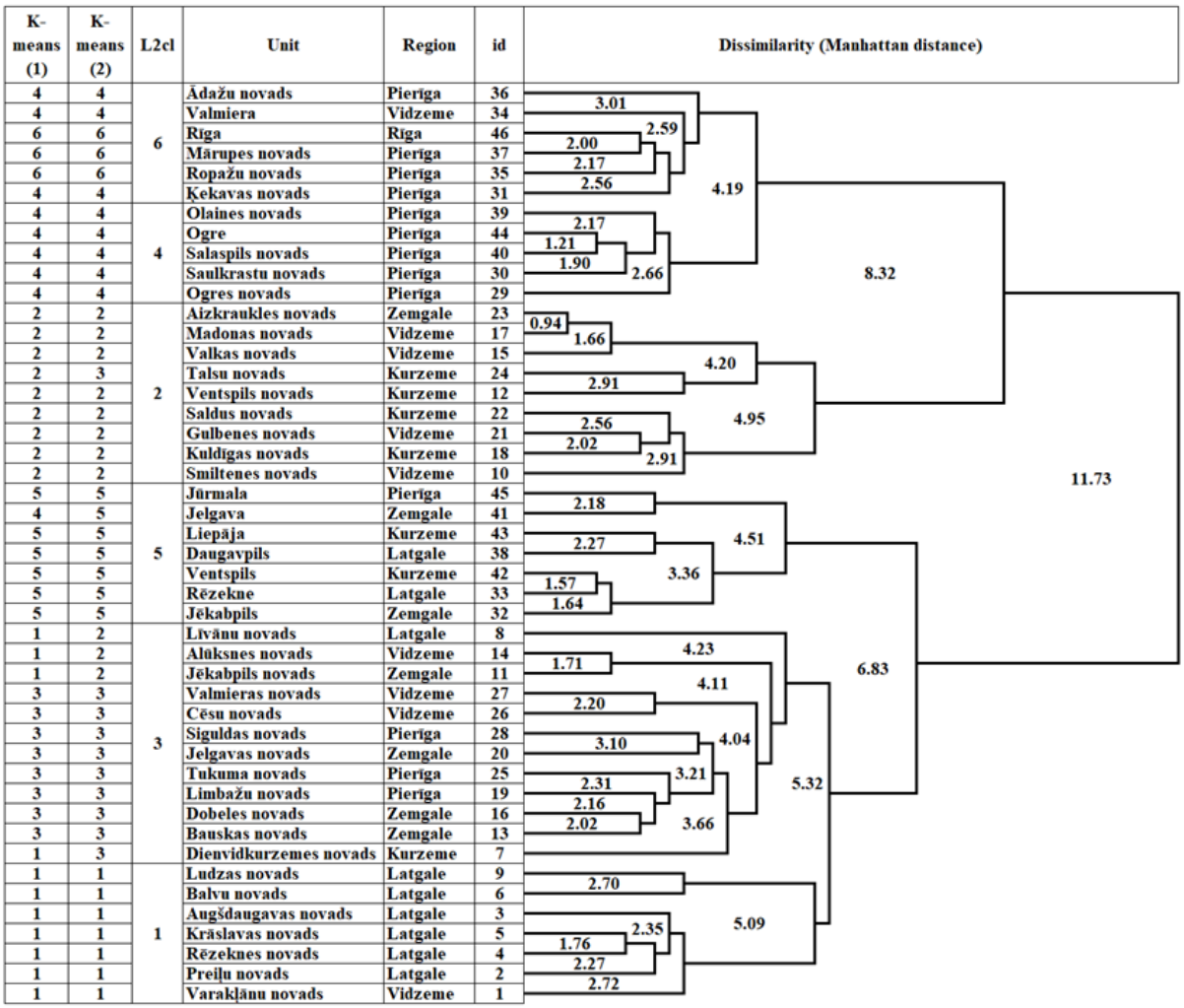


Fig. 1. **Elbow plot using distortion and Calinski-Harabasz index**

As suggested by the elbow plot, the optimal number of the clusters is six. The corresponding CH index value also is the highest. Hence, the proceeding with the six clusters is justified. To cross-validate the results obtained with the Euclidean distance, additional k-means clustering with Manhattan distance is performed for the six cluster option. For the agglomerative clustering, a Complete Linkage (denoted L2cl) technique is selected. The 46*46 dissimilarity matrix is constructed, and 45 steps of merging the pairings are performed. The resulting dendrogram with the distances (dissimilarities) along with the results of two previously performed k-means clusterings with the Euclidean distance and Manhattan distance is shown in Fig. 2.

First, the units from the dendrogram based on the results of the agglomerative clustering are assigned their respective cluster number yielded by K-means clustering with the Euclidean distance. Second, the cluster numbers from K-means clustering with the Euclidean distance and K-means clustering with the Manhattan distance are put into the first two columns in the picture.



K-means (1)	K-means (2)	L2cl	Unit	Region	id
4	4	6	Ādažu novads	Pierīga	36
4	4	6	Valmiera	Vidzeme	34
6	6	6	Rīga	Rīga	46
6	6	6	Mārupes novads	Pierīga	37
6	6	6	Ropažu novads	Pierīga	35
4	4	6	Ķekavas novads	Pierīga	31
4	4	4	Olaines novads	Pierīga	39
4	4	4	Ogre	Pierīga	44
4	4	4	Salaspils novads	Pierīga	40
4	4	4	Saulkrastu novads	Pierīga	30
4	4	4	Ogres novads	Pierīga	29
2	2	2	Aizkraukles novads	Zemgale	23
2	2	2	Madonas novads	Vidzeme	17
2	2	2	Valkas novads	Vidzeme	15
2	3	2	Talsu novads	Kurzeme	24
2	2	2	Ventspils novads	Kurzeme	12
2	2	2	Saldus novads	Kurzeme	22
2	2	2	Gulbenes novads	Vidzeme	21
2	2	2	Kuldīgas novads	Kurzeme	18
2	2	2	Smiltenes novads	Vidzeme	10
5	5	5	Jūrmala	Pierīga	45
4	5	5	Jelgava	Zemgale	41
5	5	5	Liepāja	Kurzeme	43
5	5	5	Daugavpils	Latgale	38
5	5	5	Ventspils	Kurzeme	42
5	5	5	Rēzekne	Latgale	33
5	5	5	Jēkabpils	Zemgale	32
1	2	3	Līvānu novads	Latgale	8
1	2	3	Alūksnes novads	Vidzeme	14
1	2	3	Jēkabpils novads	Zemgale	11
3	3	3	Valmieras novads	Vidzeme	27
3	3	3	Cēsu novads	Vidzeme	26
3	3	3	Siguldas novads	Pierīga	28
3	3	3	Jelgavas novads	Zemgale	20
3	3	3	Tukuma novads	Pierīga	25
3	3	3	Limbažu novads	Pierīga	19
3	3	3	Dobeles novads	Zemgale	16
3	3	3	Bauskas novads	Zemgale	13
1	3	3	Dienvidkurzemes novads	Kurzeme	7
1	1	1	Ludzas novads	Latgale	9
1	1	1	Balvu novads	Latgale	6
1	1	1	Augšdaugavas novads	Latgale	3
1	1	1	Krāslavas novads	Latgale	5
1	1	1	Rēzeknes novads	Latgale	4
1	1	1	Preiļu novads	Latgale	2
1	1	1	Varakļānu novads	Vidzeme	1

Fig. 2. **Cluster dendrogram and the cluster numbers from k-means clustering**

To finalize the choice of the clusters assigned, the results form all three methods are compared. There are certain discrepancies between the results obtained with these three approaches. Using the K-means clustering with the Euclidean distance, Jelgava State City belongs to the cluster #4, while the K-means clustering with the Manhattan distance yields the cluster #5. The latter seems more properly assigned as the cluster #5 contains predominantly large cities. Līvānu novads, Alūksnes novads and Jēkabpils novads belong to the cluster #1 and #2, respectively, while the dendrogram suggests these should belong to the cluster #3. Dienvidkurzemes novads belongs to the cluster #1 and #3. As cluster #3 is suggested also by a dendrogram, this district has to be assigned to the cluster #3. Clusters #1 and #2 have the same assignments from the results obtained with all three methods. Clusters #4 and #6 should be merged into one cluster as the cluster #6 from the dendrogram can be viewed as a part form the cluster #4.

The main conclusion from the adjusted results can be drawn with respect to the overreliance on the geographical administrative breakdown into regions. As seen from the dendrogram, a few units do not belong to clusters suggested by the administrative regions. Analysis suggests that forming the clusters can be useful to recognize patterns or detect anomalies between the regional units which would not be recognizable if an official administrative breakdown is used.

Based on similarities or differences between clustering variables, the following descriptive titles can be attributed to clusters as follows.

1. Cluster#1 – Poor "East" with the lowest per capita GVA, high share of agriculture in total number of companies, and the second lowest TDI index ranking;
2. Cluster#2 – Average performing "Northeast and West" with low per capita GDP, the second highest share of agriculture in total number of companies, and the medium TDI index ranking;
3. Cluster#3 – Average preforming "Northeast and South" with the second lowest per capita GDP, the third highest share of agriculture in total number of companies, and rather high TDI index ranking;
4. Cluster#4 – Rich "Capital City and Satellites" with the highest per capita GDP, the lowest share of agriculture in total number of companies, and the highest TDI index ranking;
5. Cluster#5 – "State Cities" with high per capita GDP, the lowest share of agriculture in total number of companies, and the lowest TDI index ranking.
6. Cluster#6 – "Broader Metropolitan area" with high per capita GDP, the lowest share of agriculture in total number of companies, and high TDI index ranking.

Note that "Northeast" districts belong to two distinct clusters.

Such a breakdown better reflects the district characteristics with respect to future targeted policy planning. For example, geographical or performance eligibility criteria for regionally targeted programmes can be modified by using more detailed district breakdown.

As there are units that do not have equal cluster assignments from three various cluster analyses - Talsu novads, Alūksnes novads, Jēkabpils novads, Līvānu novads, Dienvidkurzemes novads and Jelgava, a discriminant analysis is performed. Six outlier units are not included in the initial analysis, and these are kept for predictions.

The assigned cluster numbers from three types of cluster analysis and discriminant analysis with outliers in bold are shown in Table 1.

Table 1

Division of administrative units into clusters with various methods

Unit	Region	K-means (1)	K-means (2)	L2cl	Discriminant	Final choice
Ādažu novads	Pierīga	4	4	6	4	4
Valmiera	Vidzeme	4	4	6	4	4
Rīga	Rīga	6	6	6	6	6
Mārupes novads	Pierīga	6	6	6	6	6
Ropažu novads	Pierīga	6	6	6	6	6
Ķekavas novads	Pierīga	4	4	6	4	4
Olaines novads	Pierīga	4	4	4	4	4
Ogre	Pierīga	4	4	4	4	4
Salaspils novads	Pierīga	4	4	4	4	4
Saulkrastu novads	Pierīga	4	4	4	4	4
Ogres novads	Pierīga	4	4	4	4	4
Aizkraukles novads	Zemgale	2	2	2	2	2
Madonas novads	Vidzeme	2	2	2	2	2
Valkas novads	Vidzeme	2	2	2	2	2
Talsu novads	**Kurzeme**	**2**	**3**	**2**	**2**	**2**
Ventspils novads	Kurzeme	2	2	2	2	2
Saldus novads	Kurzeme	2	2	2	2	2
Gulbenes novads	Vidzeme	2	2	2	2	2
Kuldīgas novads	Kurzeme	2	2	2	2	2
Smiltenes novads	Vidzeme	2	2	2	2	2
Jūrmala	**Pierīga**	**5**	**5**	**5**	**4**	**5**
Jelgava	**Zemgale**	**4**	**5**	**5**	**4**	**5**
Liepāja	Kurzeme	5	5	5	5	5
Daugavpils	Latgale	5	5	5	5	5
Ventspils	Kurzeme	5	5	5	5	5
Rēzekne	Latgale	5	5	5	5	5
Jēkabpils	Zemgale	5	5	5	5	5
Līvānu novads	**Latgale**	**1**	**2**	**3**	**3**	**3**
Alūksnes novads	**Vidzeme**	**1**	**2**	**3**	**1**	**2**

Table 1 (continued)

Unit	Region	K-means (1)	K-means (2)	L2cl	Discriminant	Final choice
Jēkabpils novads	**Zemgale**	**1**	**2**	**3**	**1**	**2**
Valmieras novads	Vidzeme	3	3	3	3	3
Cēsu novads	Vidzeme	3	3	3	2	3
Siguldas novads	Pierīga	3	3	3	3	3
Jelgavas novads	Zemgale	3	3	3	3	3
Tukuma novads	Pierīga	3	3	3	3	3
Limbažu novads	Pierīga	3	3	3	3	3
Dobeles novads	Zemgale	3	3	3	3	3
Bauskas novads	Zemgale	3	3	3	3	3
Dienvidkurzemes novads	**Kurzeme**	**1**	**3**	**3**	**1**	**2**
Ludzas novads	Latgale	1	1	1	1	1
Balvu novads	Latgale	1	1	1	1	1
Augšdaugavas novads	Latgale	1	1	1	1	1
Krāslavas novads	Latgale	1	1	1	1	1
Rēzeknes novads	Latgale	1	1	1	1	1
Preiļu novads	Latgale	1	1	1	1	1
Varakļānu novads	Vidzeme	1	1	1	1	1
Ādažu novads	Pierīga	4	4	6	4	4
Valmiera	Vidzeme	4	4	6	4	4
Rīga	Rīga	6	6	6	6	6
Mārupes novads	Pierīga	6	6	6	6	6
Ropažu novads	Pierīga	6	6	6	6	6
Ķekavas novads	Pierīga	4	4	6	4	4
Olaines novads	Pierīga	4	4	4	4	4
Ogre	Pierīga	4	4	4	4	4
Salaspils novads	Pierīga	4	4	4	4	4
Saulkrastu novads	Pierīga	4	4	4	4	4
Ogres novads	Pierīga	4	4	4	4	4

The final choice is not entirely predetermined by the quantitative results. The finalized assignment of clusters is determined by the authors' choice. Note, that sequence of the cluster numbers assigned by the k-means method differs from a consequential assignment of cluster numbers with hierarchical agglomerative clustering. The layout of clusters on a Latvian map is shown in Fig. 3.

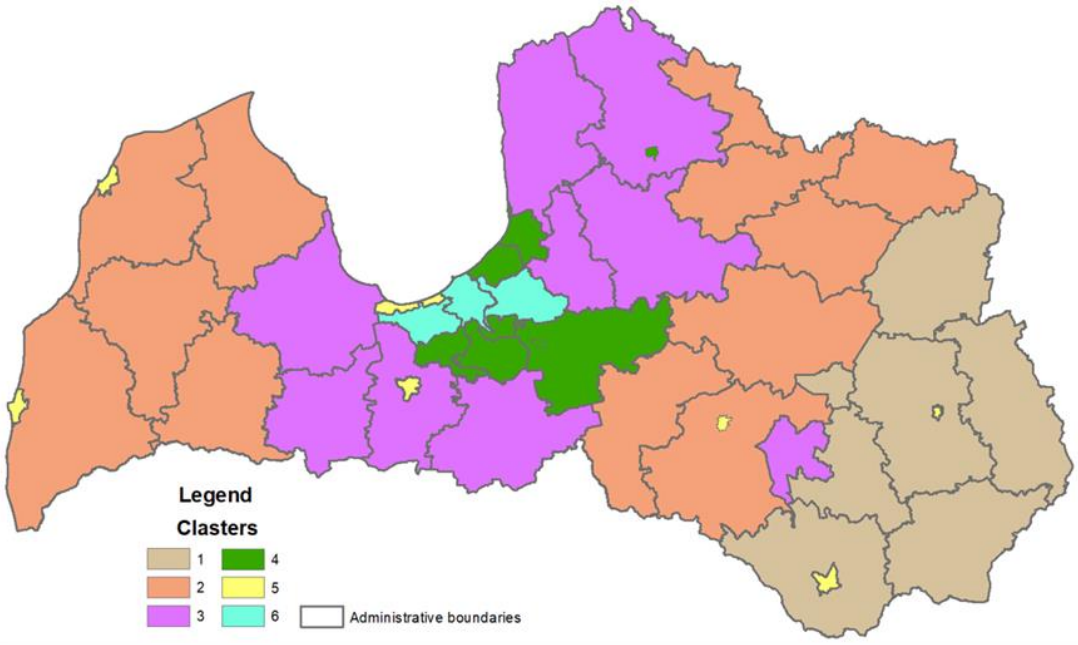


Fig. 3. **Clusters of Latvian administrative units**

Empirical results obtained show that combination of three selected methods allow for a more sensible assignment of territorial units to the clusters. In k-means clustering, cluster numbers are unambiguously assigned while the adjacency of clusters can not be established. In hierarchical agglomerative clustering, cluster numbers are assigned arbitrarily by researcher while the adjacency of clusters is clearly visible. Moreover, hierarchical agglomerative clustering enables to reveal possible overlapping of clusters which is not the case with k-means clustering. Predictive discriminant analysis helps finalize the choice of clusters for unambiguous units. Application of the clustering methods with respect to Local Administrative units is not very common. In their study on Polish provinces and Czech regions [12], the authors establish clusters based upon the indicators of human capital revealing marked differences in socio-economic variables between the clusters. In Turkey use of a clustering approach by K-means methodology allows for the delineation of five homogenous precipitation regions [13]. In Indonesia, three provincial clusters are identified with respect to the coronavirus disease risk levels [14]. In Germany ten clusters are identified showing lower economic capabilities of Eastern regions [15]. In Lithuania, authors experiment with three, five and seven clusters by combining k-means clustering with the Principal Components Analysis [16]. In Latvia, the only similar study [17] divides the country in five clusters that resembles the spatial distribution of the existing study in general. In the UK, the Office for National Statistics in the latest Census in 2021 includes experimental statistics with clustering local authorities against subnational indicators in England [18]. Clusters are defined as similar local authorities. Using the k-means clustering method, seven different models on different themes using data available are created allowing users to understand the similarities between local authorities, while providing local authorities with control groups for investigating the impact of policy interventions. The possible extensions of the clustering model to more detailed administrative breakdown with a larger number of parameter variable areas and variables of interest such as labour pendulum migration would ask for a substantial input of data engineering resources.

Conclusions

1. Based on five socio-economic parameter variables, 46 district and State City Local Administrative units in Latvia can be attributed to six clusters. The clusters are assigned distinct descriptive titles that characterize the relative situation in a particular cluster in general. Analysis suggests that forming the clusters can be useful to recognize patterns or detect anomalies between the regional units which would not be recognizable if an official administrative breakdown is used.
2. Exclusive use of k-means clustering alone does not provide the insights of cluster adjacency or remoteness. These characteristics can be obtained by simultaneous or subsequent hierarchical agglomerative clustering. While the k-means clustering yields six clusters, the hierarchical agglomerative clustering clearly shows that Cluster#4 is entirely encircled by the Cluster#6. Thus, the suggested number of clusters is five. Also, the largest dissimilarity between the poor Cluster#1 and affluent combination of Cluster#4 and Cluster#6 would not be detectable solely by k-means clustering.
3. Empirical results obtained by several methods strongly indicate towards a necessity to use a mixed approach with a combination of clustering techniques and methods applied that yield slightly varying results. Discriminant analysis should be considered a valuable tool for cross-validation of the results obtained with clustering methods.
4. Outlier units bring special interest as they provide valuable insights for policy makers. Thus, Līvānu novads by the official regional breakdown belongs to poor Latgale region albeit clustering suggests its adherence to average performing Cluster#3.
5. An extension of the clustering approach to more detailed regional breakdown at the parish level would demand for certain amount of data engineering resources due to numerous possible interactions between the large number of administrative units.

Author contributions

Conceptualization, J.H.; methodology, J.H. and E.B.; software, J.H.; validation, J.H.; formal analysis, J.H and E.B.; investigation, J.H.; data curation, J.H.; writing - original draft preparation, J.H.; writing - review and editing, J.H.; visualization, J.H. All authors have read and agreed to the published version of the manuscript.

References

[1] European Commission. European Evaluation Network for Rural Development. Getting the most from your RDP. Guidelines for the ex-ante evaluation of 2014-2020 RDPs. June 2014. [online] [16.12.2024]. Available at: https://ec.europa.eu/enrd/sites/default/files/guidelines_for_the_ex_ante_evaluation_2014_-_2020_rdps.pdf
[2] AREI. Atskaites. Ex-ante novērtējums Lauku attīstības programmai 2021-2027. 2. posma atskaite. [online] [16.12.2024]. Available at: https://www.arei.lv/sites/arei/files/files/lapas/KLP%202023-2027%20EX-ANTE%20zinojums_1.posms_AREI_2020.pdf
[3] European Cluster Collaboration Platform. [online] [17.02.2025]. Available at: https://www.clustercollaboration.eu/cluster-definitions
[4] Kulakova N. Business clusters formation in Latvia. Procedia-Social and Behavioral Sciences, 110, 2014, pp.174-181.
[5] Boronenko V., Zeibote Z. The potential of cluster development and the role of cluster support policies in Latvia. Economic Annals, 56(191), 2011, pp. 35-67.
[6] Lloyd S.P. Binary block coding. Bell System Technical Journal, vol. 36(2), 1957, pp. 517-535.
[7] Caliński T., Harabasz J. A dendrite method for cluster analysis. Communications in Statistics, vol. 3(1), 1974, pp. 1-27.
[8] Thorndike R. L. Who belongs in the family? Psychometrika, 18(4), 1953, pp. 267-276.
[9] Anderberg M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.
[10] Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol. 7, 1936, pp. 179-188.
[11] Centrālā Statistikas pārvalde. Oficiālās Statistikas portāls. Datubāze. [online] [16.12.2024]. Available at: https https://data.stat.gov.lv/pxweb/lv/OSP_PUB/START__POP__IR__IRS
[12] Wielechowski M., Cherevyk D., Czech K., Kotyza P., Grzęda Ł., Smutka L. Interdependence between human capital determinants and economic development: K-means regional clustering approach for Czechia and Poland. Entrepreneurial Business and Economics Review, vol. 9(4), 2021, pp. 173-194.
[13] Türkeş M., Tatlı H. Use of the spectral clustering to determine coherent precipitation regions in Turkey for the period 1929-2007. International Journal of Climatology, vol. 31(14), 2011, pp. 2055-2067.
[14] Abdullah D., Susilo S., Ahmar A. S., Rusli R., Hidayat R. The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Quality & quantity, vol. 56(3), 2022, pp. 1283-1291.
[15] Sternberg R., Litzenberger T. Regional clusters in Germany - their geography and their relevance for entrepreneurial activities. European Planning Studies, vol. 12(6), 2004, pp. 767-791.
[16] Gružauskas V., Čalnerytė D., Fyleris T., Kriščiūnas A. Application of multivariate time series cluster analysis to regional socioeconomic indicators of municipalities. Real estate management and valuation, vol. 29(3), 2021, pp. 39-51.
[17] Dahs A., Krumins J., Berzins A., Lece K. Vanguard and laggard rural population clusters in Latvia. Economic Science for Rural Development, 2024, pp. 279-289.
[18] Office for National Statistics. Data and analysis from Census 2021. [online] [09.02.2025]. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/clusteringlocalauthoritiesagainstsubnationalindicatorsengland/2023-02-24