# Cluster analysis of the Liberec region municipalities

Miroslav Žižka[1]

**Abstract.** The article deals with the cluster analysis of municipalities in the Liberec Region. It builds on the results of factor analysis, which defined seven significant factors that characterize the socio-economic status of communities. The input for cluster analysis was the factor loadings, which allowed making multi-dimensional classification of the studied municipalities. To build the structure of clusters there were used both hierarchical clustering methods and non-hierarchical clustering methods with the help of k-means. In the first case the Ward's method was utilized to create clusters and the results were graphically displayed by tree diagram. With the help of the division procedure an initial cluster was divided into 2 to 4 smaller clusters. Two basic clusters can be classified as urban and rural. Urban cluster is characterized by better features of population settlement, age structure, civic amenities and the branch structure. On the contrary, cluster consisting of rural communities has an aging population, population migrating to cities, poor civic amenities and the employment structure is dominated by manufacturing and agriculture. When lowering the clustering level, it can be found that rural cluster breaks down to agricultural, submountaneous and cross-border areas. In the second stage, the method of non-hierarchic clustering using k-means was applied on the results of factor analysis. A set of municipalities of the Liberec Region was gradually divided into 2 to 10 clusters, which were analyzed in detail using the R-square index and Calinski-Habarasz F index. Taking into account the requirement that the clusters were not formed only by a few elements and outlying values, the region was finally divided into 6 clusters, which can be classified as urban, micro-regional, economically weak and rural focused on housing, services and agriculture.

**Keywords:** Cluster analysis, hierarchical cluster analysis, non-hierarchical cluster analysis, factor loadings, Ward's method, k-means clustering, measure of disagreement, recall coefficient, F-ratio, Calinski-Habarasz F index.

**JEL Classification:** C38, R11, R15
**AMS Classification:** 62H30, 91C20

## 1 Introduction

Cluster analysis is one of the frequently used multivariate statistical methods in social sciences. The examples of practical application of cluster analysis include classification of the EU regions in terms of business environment on the basis of factor loadings of the individual regions [6], the division of countries in Europe, the Middle East and Africa in similar groups in terms of the position of high tech companies, the economic environment of the country and its competitiveness index [4] or using a hierarchical clustering method to classify SMEs in terms of their ICT competencies [1]. Another example is the creation of a number of client profiles of bank customers with the help of the modified hierarchical method using a CF tree [2].

In the article [9], a methodology for evaluating the socio-economic status of municipalities in the Czech Republic was characterized. Based on this methodology, seven significant factors (F1 - unemployment, F2 - domicile attractiveness, F3 - population settlement, F4 - age structure, F5 - civic amenities, F6 - branch structure and F7 - economic activity; below are used only abbreviated designation of factors F1 to F7) have been identified, including the factor loading of the individual municipalities.

This article aims to build on the results of previous research and divide communities characterized by means of factor loadings into homogeneous clusters, if possible, in which the individual municipalities will be as similar as possible in terms of factor loadings. Individual clusters give an idea of socio-economic status of the cluster while allowing a classification of municipalities surveyed with regard to individual factors of the socio-economic environment. In the first stage, methods of hierarchical clustering were used to construct clusters, in the second stage; they were followed by methods of non-hierarchical clustering.

[1] Technical University of Liberec/Faculty of Economics, Department of Business Administration, Voroněžská 13, 461 17 Liberec, Czech Republic, miroslav.zizka@tul.cz.

## 2 Divisional hierarchical clustering

In this procedure, we assume that all municipalities form a cluster and by its gradual division we obtain a larger number of clusters so that we would end up in each municipality. In fact, we will attempt to divide a basic set of the Liberec Region municipalities into several large, internally homogeneous clusters that will vary in individual factors of the socio-economic environment.

Similarity of municipalities was evaluated first by tree diagram of objects; the distance of objects was measured by the squared Euclidean distance which forms the basis of Ward's clustering method (for example, see [5]). Ward's method was used to construct clusters because graphical analysis showed that the surveyed municipalities tend to be grouped into multi-dimensional ellipsoids. One problem was the fact that there were 215 municipalities and it did not allow a clear description of their names on the vertical axis of tree diagram (see Figure 1). For the assignment of individual municipalities in clusters was therefore used the table of amalgamation schedule in the STATISTICA CZ 10 programme. Clustering level h = 210 split the set of municipalities into 2 unequally large clusters. The first cluster includes 46 municipalities and is made up of larger regional settlement (the average number of inhabitants is 7,454); while the second cluster contains 169 mostly small municipalities (average number of inhabitants in this cluster is 569). We can say that on the significance level $\alpha = 5\%$ difference in the average number of inhabitants in the two clusters is statistically significant (p-value < 0.0001).

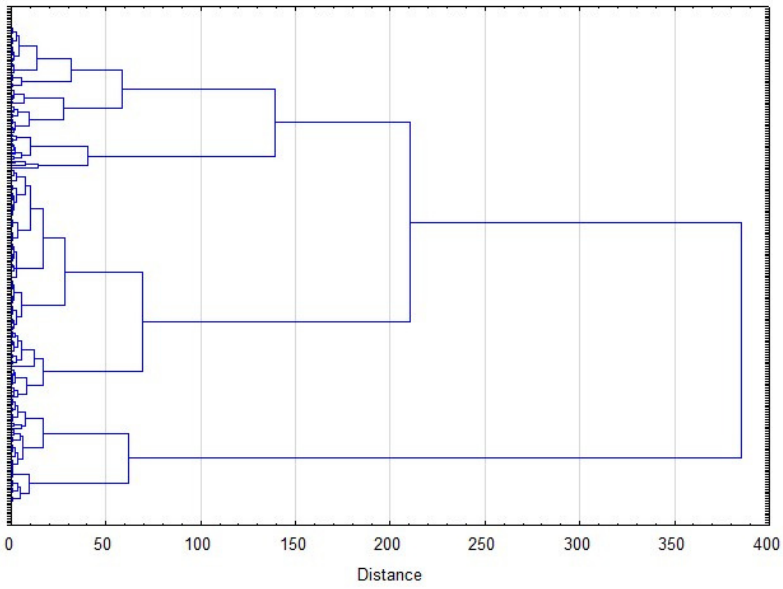


**Figure 1** Tree Diagram

Ward's Method, Squared Euclidean distance

When comparison was made of the averages of factor loadings for both clusters (see Figure 2) at significance level $\alpha = 5\%$ substantial differences were verified between clusters of factors F3 (p-value < 0.0001), F4 (p-value = 0.0453), F5 (p-value < 0.0001) and F6 (p-value < 0.0001). On the contrary, a statistically significant difference between the average of factor loadings were not detected in clusters of factors F1 (p-value = 0.2989), F2 (p-value = 0.6800) and F7 (p-value = 0.9015). Based on these findings, we can say that cluster 1 is characterized by better parameters of indicators relating to population settlement, age structure, civic amenities and the branch structure. Due to the fact that this cluster usually includes cities and larger municipalities, the results are logical. As shown in the atlas [8], larger communities are characterized by higher population density, usually a positive migration balance, are equipped with basic educational and medical facilities, and most of the population is economically active in the tertiary sector. On the contrary, cluster 2 is composed of mostly small, rural communities, often facing an aging population and outflow of inhabitants to the cities, there is a need to commute for civic amenities and manufacturing industry and agriculture play more important role in the employment structure.

With the decrease of clustering level at h = 140, we obtain three clusters of municipalities which include 46, 105 and 64 municipalities. The first cluster of municipalities yet remains the same (see bottom of Figure 1). The second cluster now includes a municipality with an average population of 474 and the third cluster has an average size of 724 residents, but in both cases with a high degree of variability (v = 0.67, 1.15 respectively). Cluster 2 is represented by rather agricultural areas of the region, while cluster 3 includes the submountaneous and cross-border areas of the region. If we reduce the clustering level arbitrarily at h = 70, we obtain 4 clusters of municipalities with 46, 105, 16 and 48 municipalities each. In an analogous way, we can continue to form other,

smaller clusters of municipalities. In principle, this hierarchical approach provided us with the basic information about the structure of clusters of municipalities in the Liberec Region and for further analysis we will use the method of non-hierarchic clustering.

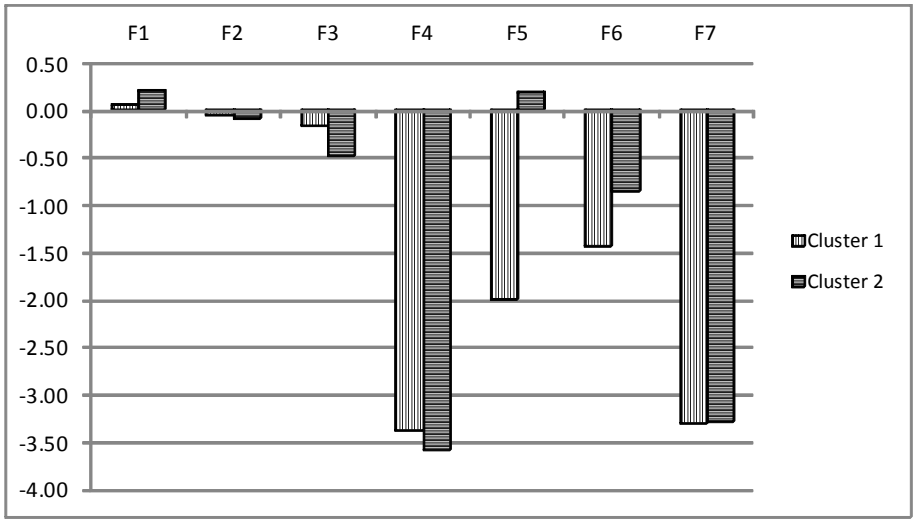


**Figure 2** Average values of factor loadings in clusters 1 and 2

# 3 Clustering by the k-means method

The aim of the application of this method is the division of individual municipalities of the Liberec Region in the predetermined number of clusters. The algorithm is described, for example, in the work [3]. The initial decomposition can be determined randomly, in our case, however, it can be based both on the previous result of hierarchical clustering, where "reasonable" number of clusters appears to be 2 to 4, and on the administrative structure of the Liberec Region that is divided in 4 districts and 10 administrative districts of municipalities with extended powers. For these reasons, the division of municipalities in the region was tested in 2 to 10 clusters. Calculations were again performed in the STATISTICA CZ 10 programme. Three options to determine the initial cluster centers are available: maximizing the initial distance between clusters, ordering observations according to distance and choice of objects in constant intervals or selection of the first N observations. Individual variants were compared using analysis of variance to determine whether the averages of the individual factors are significantly different between groups. For the option k = 2 clusters, the least suitable method appeared to be arrangement according to distance (at significance level $\alpha = 5\%$ the null hypothesis of conformity of variances was not rejected in the case of 4 factors, in other two methods, H0 was not consistently rejected in one variable). For the option k = 3 clusters, H0 was not rejected consistently in the case of one factor in the following methods - sort distances and choose the first N observations. In the case of k = 4 and a larger number of clusters, it can be already said that congruence of variance was not verified in any factor (see Table 1), and for all methods. Table 2 shows the number of municipalities across clusters.

The determination of "optimal" number of clusters is a highly discussed question of cluster analysis. As stated in [5], there is no objective way to determine such a termination criterion. In the literature, (see for example [5], [7]), there is therefore described a number of different criteria and methods for assessing the quality of clusters.

| Factor | between SS | Df | within SS | Df | F | p-value |
|---|---|---|---|---|---|---|
| F1 | 97.1268 | 3 | 66.97478 | 211 | 101.9974 | 0.000000 |
| F2 | 3.1079 | 3 | 57.59572 | 211 | 3.7952 | 0.011119 |
| F3 | 4.8871 | 3 | 46.76894 | 211 | 7.3494 | 0.000104 |
| F4 | 20.1172 | 3 | 51.18085 | 211 | 27.6453 | 0.000000 |
| F5 | 192.8578 | 3 | 33.22964 | 211 | 408.1998 | 0.000000 |
| F6 | 23.8639 | 3 | 30.18421 | 211 | 55.6060 | 0.000000 |
| F7 | 44.8476 | 3 | 27.92743 | 211 | 112.9457 | 0.000000 |

**Table 1** Analysis of variance

Notes: N = 4 clusters, method of maximizing of initial distances between clusters

One of the simpler methods for evaluating clusters is the **measure of disagreement *MD*** (1), based on confusion matrix [7]. Suppose that we know in advance the structure of *P* division of municipalities in clusters and further the structure *C* obtained by cluster analysis. Confusion matrix contains the number of objects occurring simultaneously in cluster structure *P* and cluster structure *C*. Number of common objects shall be denoted as $n_{hh}$. The number of clusters in *C* and *P* is the same. The structure of clusters known in advance was derived from the administrative structure of the Liberec Region, i.e. 4 districts and 10 administrative districts of municipalities with extended power (MEP).

$$MD = \frac{n - \sum_{h=1}^{k} n_{hh}}{n} \tag{1}$$

When comparing the composition of the municipalities in 4 districts with the structure of 4 clusters obtained by cluster analysis, we find that MD = 0.56. For 10 MEP and 10 clusters, the degree of disagreement is even higher, namely 0.72. Therefore, it can be said that there are clusters of municipalities across districts and MEP which do not copy the existing administrative structure of the region that much.

Other measures based on a comparison with the predetermined classification are the **accuracy coefficient** $P_{hh'}$, expressing a share of common objects of two clusters on the number of objects from the structure *C* or the **recall coefficient** $R_{hh'}$, indicating a share of common objects on the number of objects of the cluster structure *P*. Combining the two above-mentioned measures we receive the **F-ratio** (2), which is their harmonic average. The result is a value from 0 to 1.

$$F_{hh'} = 2\frac{P_{hh'}R_{hh'}}{P_{hh'} + R_{hh'}} \tag{2}$$

In the case of F-ratio and 4 clusters, cluster 4 (0.61) shows the highest value where there is a high number of common municipalities in the Semily district, while cluster 2 consists of municipalities from all districts (F = 0.17). If we compare the structure of the emerged 10 clusters with the structure of 10 MEP, then the highest congruence was observed in cluster 9 (F = 0.32), where there are most of the municipalities from Turnov MEP.

| Cluster No. | k = 4 | k = 5 | k = 6 | k = 7 | k = 8 | k = 9 | k = 10 |
|---|---|---|---|---|---|---|---|
| 1 | 54 | 70 | 68 | 1 | 1 | 15 | 2 |
| 2 | 13 | 46 | 15 | 15 | 39 | 4 | 22 |
| 3 | 39 | 35 | 39 | 36 | 4 | 40 | 16 |
| 4 | 109 | 12 | 35 | 79 | 18 | 26 | 3 |
| 5 | | 52 | 12 | 33 | 71 | 1 | 33 |
| 6 | | | 46 | 39 | 39 | 38 | 1 |
| 7 | | | | 12 | 15 | 52 | 15 |
| 8 | | | | | 28 | 2 | 38 |
| 9 | | | | | | 37 | 26 |
| 10 | | | | | | | 59 |

**Table 2** Number of municipalities in individual clusters

To assess the number of clusters, **R-square index** was used. It measures the share of between-groups variability $SS_B$ on total $SS_T$ variability, see equation (3). Generally, an increasing number of these clusters become more homogeneous, and therefore the value of the determination ratio increases. A similar index based on an analysis of variance is **Calinski Habarasz F index** (*CHF*) as given in equation (4). It holds true that high values of *CHF* show well-formed clusters. In the analysis, therefore, the maximum within a certain interval is searched. From Table 3 it is obvious that both indices reach the highest values in case of 9 clusters. On the other hand, it is evident that when there is a large number of clusters, low-element clusters are formed. They include outlying observations, and are difficult to interpret. Therefore, 6 clusters, which were subjected to a further analysis, are the most sensible number of clusters.

$$RSQ = \frac{SS_T - SS_W}{SS_T} \tag{3}$$

$$CHF = \frac{\frac{SS_B}{k-1}}{\frac{SS_W}{n-k}} \tag{4}$$

| Cluster No. k | Total squares B | Total squares W | RSQ | CHF |
|---|---|---|---|---|
| 4 | 5.68 | 39.75 | 0.13 | 10.06 |
| 5 | 9.11 | 39.45 | 0.19 | 12.12 |
| 6 | 15.65 | 34.79 | 0.31 | 18.80 |
| 7 | 24.57 | 33.83 | 0.42 | 25.17 |
| 8 | 38.72 | 31.12 | 0.55 | 36.80 |
| 9 | 49.93 | 14.51 | 0.77 | 88.58 |
| 10 | 65.37 | 27.70 | 0.70 | 53.74 |

**Table 3** RSQ and CHF indexes for different number of clusters

Municipalities in **cluster 1** include small to medium-sized municipalities (the largest municipalities is Cvikov with 4.5 thousand inhabitants, the smallest Janovice with 80 inhabitants). The cluster is characterized by favorable employment features, domicile attractiveness, age structure and branch structure. By contrast, the factor of population settlement and civic amenities develops negatively (see Figure 3). Cluster 1 can be classified as rural, fulfilling mainly the function of housing and recreation.

**Cluster 2** contains all four district towns, all 10 seats of MEP and 5 other larger communities. In this cluster, population settlement, age structure, civic amenities and the branch structure can be positively evaluated. On the contrary, factors of unemployment, the domicile attractiveness and economic activity appear to be problematic. Cluster 2 can be described as urban, providing catchment function for the surrounding communities and branch focus on service.

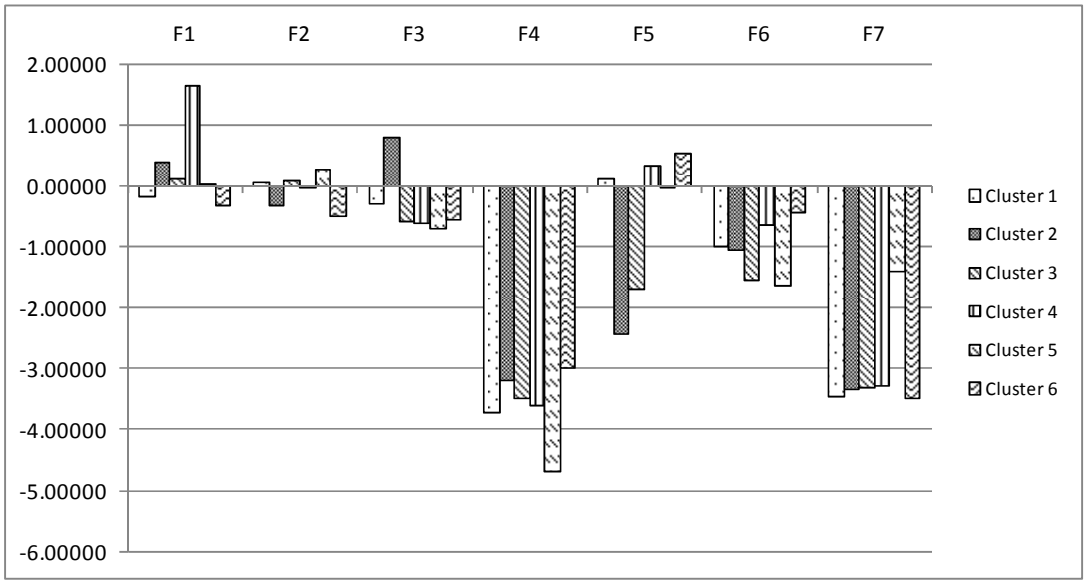


**Figure 3** Average values of factor scores in clusters 1 to 6

In **cluster 3**, there are medium-sized municipalities (the average population of around 2 thousand), out of which 7 of them are municipalities with authorized municipal office that perform the functions of government. Other communities are centres of subregional units. The cluster is characterized by good domicile attractiveness and convenient civic amenities. The weaknesses are unemployment and population settlement. The cluster can be described as micro-regional providing essential central functions.

**Cluster 4** includes rather medium-sized municipalities (average population of about 800), which are located along the borders of the Liberec Region. Of all the observed clusters, this one is characterized by high unemployment, negative domicile attraction, sparse population settlement and poor age structure. The cluster can, therefore, be classified as economically weak area of the region.

**Cluster 5** contains rather smaller municipalities (except Plavy) mostly in the northwestern part of the region. The cluster shows good characteristics in the domicile attractiveness, very good age structure and average ratings in civic facilities (mainly thanks to the presence of primary schools in these municipalities). The unemployment factor can be evaluated as neutral. The weakness of municipalities in the cluster is a low density of settlement. In the cluster, there are tourist attractive municipalities (Bezdez, Zahradky, Bedrichov). The cluster can be characterized as rural with a predominant focus on tourism services.

In the last **cluster 6**, there are mostly small communities in the southeast region (MEP Turnov, Semily and Zelezny Brod). The cluster is characterized by favorable employment parameters and age structure. On the contrary, more domicile attractiveness, settlement and civic facilities can be viewed as weakness. The branch structure shows a higher proportion of a primary sector. The cluster can be classified as rural, with a higher share of employment in agriculture.

# 4 Conclusion

Although the determination of "optimal" number of clusters is relatively complex and ambiguous question, the analysis provides quite logical results. If we split the region just into two clusters, in the first cluster, there are larger cities and in the latter, there are small, rural municipalities. In the case of three clusters, small municipalities can be further divided into municipalities focusing on agriculture and cross-border and submountaneous communities, where services play a greater role in tourism. When using k-means method the division of the region into 6 clusters was eventually chosen. They can be classified as urban, rural - with the function of housing, rural - focused on services, rural - agricultural, micro-regional with central functions and the economically weak area. This division can be used as appropriate in formulating suitable development strategies of cities, municipalities and microregions. The creation of more clusters does no longer seem reasonable, since further created clusters begin to contain fewer elements.

## Acknowledgements

The article was prepared with the support of the project Technology Agency of the Czech Republic ev. No. TD010029 "Definition of subregions for distinguishing and the solution of social and economic disparities".

## References

[1] Antlová, K., Popelínský, L., and Tandler, J.: Long Term Growth of SME from ICT Competencies and Web Presentation. *E&M Ekonomie a Management* **14** (2011), 125-139.
[2] Draessler, J., Soukal, I., and Hedvičáková, M.: Shluková analýza poptávkové strany trhu základních bankovních služeb. *E&M Ekonomie a Management* **14** (2011), 102-114.
[3] Hebák, P., et al.: *Vícerozměrné statistické metody [3].* INFORMATORIUM, Prague, 2007.
[4] Kraftová, I., and Kraft, J.: High tech firmy a tvorba bohatství v zemích EMEA. *E&M Ekonomie a Management* **11** (2008), 6-20.
[5] Meloun, M., and Militký, J.: *Kompendium statistického zpracování dat.* Academia, Prague, 2006.
[6] Odehnal, J., and Michálek, J.: Empirická analýza regionálního podnikatelského prostředí vybraných zemí EU. *Politická ekonomie* **59** (2011), 242-262.
[7] Řezanková, H.: Hodnocení kvality shluků. In: *Analýza dat 2008/II – Statistické metody pro technologii a výzkum* (Kupka, K., ed.). TriloByte Statistical Software, Pardubice, 2008, 19-40.
[8] Žižka, M., et al.: Atlas ekonomických ukazatelů. Professional Publishing, Prague, 2011.
[9] Žižka, M.: Model for Assessment of the Social Economic Level of Municipalities. In: *Proceedings of the 29th International Conference on Mathematical Methods in Economics 2011 – part II.* (Dlouhý, M., and Skočdopolová, V., eds.). Professional Publishing, Prague, 2011, 786-791.