

# APPLICATION OF MULTIVARIATE STATISTICAL METHODS IN THE ANALYSIS OF CZECH POPULATION LIFE QUALITY WITH ATTENTION TO REGIONAL DIFFERENTIATION

Andrea Jindrová<sup>1</sup>

<sup>1</sup>Department of Statistics, Faculty of Economics and Management, Czech University of Life Science Prague, Kamýcká 129, 165 21 Praha 6-Suchbát, Czech Republic

## Abstract

JINDROVÁ ANDREA. 2015. Application of Multivariate Statistical Methods in the Analysis of Czech Population Life Quality with Attention to Regional Differentiation. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 63(5): 1671–1678.

Quality of life in the regions is affected by many mutually interlinked factors. The paper is aimed at the research of regional disparities in CR population life quality as assessed from the viewpoint of economic efficiency of the region and the social and environmental conditions. The interregional disparities research started from statistical modeling based on identification of key indicators affecting life quality in the CR districts and the outcomes reached have been exploited further for multidimensional classification of districts as to the indicators analyzed. Attention has been paid also to the ways of application of cartographic map facilitating a clear visualization of regional disparities.

Keywords: quality of life, multidimensional statistical methods, disparity, economical indicators, social and cultural indicators, environmental indicators

## INTRODUCTION

The beginnings of research and life quality concept defining come to attention by the second half of the last century. By that time, studies were aimed at the research of state of the society as based on an objective look into the conditions of living (Andrews, Whitney, 1976).

As Andráško (2013) mentions the complexity and complicatedness of a human life is caused by overlapping dimensions and therefore it is not possible to focus on one particular part only when analysing it.

Quality of life became the object of systematic research not earlier than the last decades of 20<sup>th</sup> century. J. K. Galbraith and D. Riesman started writing in the Fifties on life quality as a new topic of sociology. In the Sixties, the first research application of the „quality of life“ concept (QOL) appears. The research was linked to the „Social Indicators“ movement, where quality of life was seen as not affected by economic indicators only, but by the environment, too, where people live, i.e.,

village or town. By that time, the quality of life title started being employed in politics, too (Vašurová, Mühlbacher, 2005).

A general target of the paper is statistical comparative analysis of regional disparities in the CR population quality of life, as seen from the viewpoint of a region's economic efficiency, social and environmental conditions. Attention is also paid to the chances of application of cluster analysis in the assessment of regional differentiation of separate CR districts.

Increased interest in quality of life is seen in the tendency to find separate aspects and factors affecting it. In spite of existence of many designs of definitions, ways of computation and selection of indicators included into the assessment, no generally accepted approach has been found so far. The statistical aspects of quality of life measurement are very heterogeneous and they are mainly aimed at a comparison of the positions of the regions studied. Based on such measurements, differences can be identified in the levels of economic, social

and environmental development of the regions. Such differences are called regional disparities. A disparity is each difference or inequality, the identification and comparison of which has some sense (social, economic, political, etc.).

A regional disparity can be understood „the difference or disproportion of various phenomena or processes having a unique regional location and being found within at least two entities of the regional unit“ (Hučka, Kutscherauer, 2008).

The CR Ministry for Regional Development defines regional disparities as „groundless regional differences in the levels of economic, social and ecologic development of the regions“. The disparities to be solved, are „the differences caused by subjective human activities, not the differences arising from objective cause, from natural conditions, for example“. A disparity is often seen a consequence of an undesired cause, i.e., a problem. On the other hand, positive disparities can be defined, too, meaning those caused by the strong lines of the region. They are the comparative advantages offering basis for development of the region in question.

Research of disparities should offer a basis for application of the regional policy instruments. According to the authors mentioned, a great importance is to be approved to the differentiation of partial look at regional differences from the application of synthetical indicators, when defining a „problem region“ in the process of imposing central regional policy (Binek, Galvasová, 2009).

Efforts in importing macro indicators on the regional level are typical in the research of life quality. The problem stands in the question of indicators selection, where it is not easy to come to agreement. The questions to be often answered when selecting indicators at a regional level are, e.g., accessibility of data in time and space, or level of homogeneity of the state in question. Selection of indicators can be affected, too, by the actual topical attention of research and by the research team structure. An exception can be expected in case of national income, being undiscutably linked to quality of life (Charvát, Petr, 2009).

When defining an indicator having a sufficient information power for informing about the development of a region, it is needed to be based on qualified statistical analyses and it is useful to include in these the indicators describing the region's development on the one hand, and indicators specific for the region's development on the other hand (Svatošová, Boháčková, Hrabánková, 2005).

## MATERIALS AND METHODS

Assessment of the degree of disparities in quality of life has been based on an analysis of the year 2012 indicators. Starting level for the regional differences assessment was an arrangement by

districts (NUTS 4), less the Capital Prague. City of Prague was excluded from the assessment due to its specific position as compared with other districts. The original data matrix contained 71 indicators subdivided into three topical domains: Economic (25 indicators), Social (32 indicators) and Environmental (14 indicators). Arrangement of partial survey domains was based on the CR Strategy of Regional Development descriptors selection, besides the development documents of separate CR Regions used for the design of regional disparity assessment methodology.

Selection of indicators and consequent reduction of these was based on correlation analysis and principal component analysis. Correlation analysis used the Spearman rank correlation coefficient. In case, a strong correlation ( $|r| > 0.8$ ) had appeared between some indicators, these were subsequently diagnosed by means of VIF (Variance Inflation Factor) due to the risk of undesirable multicollinearity. The VIF measure makes it possible, to establish the so-called variance expansion factors. A heuristic rule is usually employed in practice, according to which, VIF values  $> 10$  signal undesired multicollinearity between the variables (Kába, Svatošová, 2012).

Also the principal component analysis operation started from correlation matrix and it was applied for the assessment of inner links in the relationships between partial indicators and the visualization of these. Principal component analysis was applied in looking for correct dimension of the data set (those indicators have been chosen as sufficiently significant, that showed correlation between the indicator and the principal component above 0.7), and in defining new variables (transformation of the original variables,  $x_i$ ,  $i = 1, 2, \dots, m$  into a smaller number of mutually uncorrelated latent variables  $y_i$ ). (Meloun, Militký, 2006).

The new uncorrelated variables established made it possible to correctly explore the data using cluster analysis with the aim of classification of regions into groups showing similar ways of life quality assessment. A starting point when clustering data is the decision how to express similarity (distance) between separate regions. In the paper presented, Euclidean distance has been selected as the measure of distance. It is a classical measure of distance used in geometry, generalized for multivariate data. For clustering proper, the Ward method has been used, aimed at forming clusters with maximum inner homogeneity. The method belongs among hierarchic clustering methods. Hierarchic clustering starts at  $n$  clusters, where each observation makes an individual cluster and it ends at one cluster, gathering all observations. During every step, two closest observations or observation clusters are united into one new cluster. Progress of clustering is demonstrated by means of a special tree-like graph, called the dendrogram, demonstrating separate steps of hierarchic clustering, inclusive of the distances, at which separate clusters (or observations) have been

united. The dendrogram can also be applied when presenting the results but it is applicable in the case of a smaller number of objects studied only (30 units maximum). More info on this can be found e.g., in Hebák *et al.* (2007) or Řezanková *et al.* (2007).

Cartographic visualization has been chosen as the basic analytical instrument for representation of spatial relationships. Presenting indicator values in the shape of maps opens a way to discover the structure of indicators studied easily, and to obtain a basis for description of disparities and for discovering the causes of disparities.

IBM SPSS version 19 statistical software has been employed for data processing and multivariate statistical methods application.

## RESULTS AND DISCUSSION

Main target of the paper is the comparative analysis of regional disparities in the CR population quality of life as seen from the viewpoint of a region's economic efficiency, social and environmental conditions, applying cluster analysis. In order to reach this target, selection of key indicators suitable for assessment of regional disparities in quality of life between the regions chosen was the first step. As mentioned above, the primary data matrix contained 71 indicators. Primary selection of variables was based on correlation analysis. Based on this analysis, the original number of variables was reduced in the Economic Domain by 4 indicators, and in the Social and Environmental Domains by 5 indicators each.

A further reduction of dimensionality was based on principal component analysis. For the detection of key indicators, values of correlation coefficients were decisive, expressing the correlation between the indicator given and the component. Based on the values of these (correlation coefficient  $> 0.7$ ), all in all 29 indicators were excluded from the original matrix. The Economic Domain was reduced by 11 indicators, the Social Domain by 15, the Environmental Domain by 3 indicators.

Based on correlation and principal component analyses, the number of indicators over all three domains of study was reduced to the total number of 28 indicators and these were chosen as the key ones for the regions grouping. This way, data matrix has been reduced by 60% of its dimension.

**Economic Domain** has been represented by indicators:

- Railway network density (RND);
- Other transport areas density (OTAD);
- Number of persons travelling to place of job over 60 minutes/day (N60);
- Unemployment rate – females (URF);
- Average length of registration at Employment Office – males (AEOM);
- Number of job candidates per one vacancy (NJCV);
- Number of vacancies for graduates and youngsters (NVGY);

- Number of subjects without employees (NSWE);
- Number of subjects with 1–9 employees – microforms (MIKRO);
- Number of registered entrepreneur units – persons (NP FO).

**Social Domain** has been represented by indicators:

- Population density (Popu);
- Natural increase (NI);
- Index of age (IA);
- Number of medical staff (Medic);
- Number of dentists (Dent);
- Number of medical specialists (Special);
- Average pension (AP);
- Quotient of University graduates (UG);
- Quotient of journeymen (UJ);
- Quotient of elementary school graduates (ESG);
- Quotient of households with one automobile (HA);
- Quotient of households with telephone (HT).

**Environmental Domain** has been represented by indicators:

- Arable land per inhabitant (Arable);
- Quotient of gardens and orchards (QGO);
- Quotient of forest land (Forest);
- Quotient of dumping areas (QDA);
- Ecologic stability coefficient (ESC);
- Quotient of permanently occupied family houses with gas (Gas).

Considering size of the paper, it is not possible to explain whole the method of the variables number reduction. A detailed dimensionality reduction procedure is explained in the article (Jindrová, Poláčková, 2012).

Principal component analysis was not only aimed at the indicators number reduction, but at the transformation, too, of original variables into a smaller number of latent variables, used for data exploration by means of cluster analysis.

For grouping of objects, five principal components were used in the Economic Domain, explaining 74.66% of the total original data variance (selection had been done by means of Kaiser criterion, according to which only those principal components having the characteristic root above 1 are to be included in further analysis). Based on the outcomes of clustering, the best number of relatively homogeneous clusters was 7. After this finding had been accepted, co-ordination of separate regions into clusters was done. In order to make the results intelligible it was decided that the first cluster will represent the greatest group of regions, and the frequency of districts representation in the groups will gradually decrease with the number of the cluster rising.

Presentation of the outcomes started from a comparison of the selected indicators cluster average quotients for Economic Domain on all-Republic averages (minus Prague).

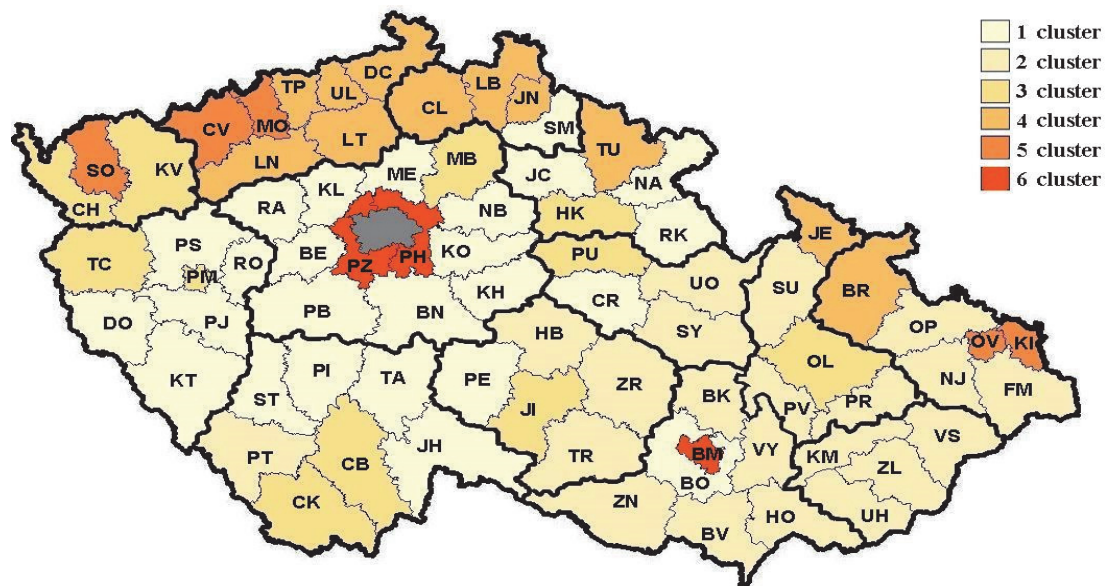




II: Quotients of selected indicator cluster averages on the Czech Republic averages

Cluster	Popu	NI	IA	Medic	Dent	Special	AP	UG	UJ	ESG	HA	HT
1	0.55	0.62	1.04	0.85	0.95	0.91	1.00	0.90	1.01	1.00	1.09	1.02
2	0.72	0.52	1.03	0.93	1.03	0.97	0.99	0.98	1.04	1.03	0.99	0.88
3	0.99	1.65	1.01	1.41	1.15	1.30	1.01	1.22	0.94	0.91	1.02	1.14
4	0.91	1.04	0.91	1.03	0.95	0.97	0.99	0.89	1.01	1.11	0.88	0.95
5	2.76	0.34	0.93	1.07	0.97	0.93	1.02	0.98	1.06	1.09	0.79	0.96
6	4.08	6.14	0.88	1.05	0.90	1.13	1.04	1.60	0.74	0.68	1.01	1.39

Source: own computation



2: Czech Republic districts as related to clusters Social Domain

Source: own computation

with the low value of the N60 (see above) and with those indicators that represent the numbers of economic subjects (NSWE, MIKRO, NP), showing values below the CR average. Based on the interplay of the values mentioned, can this cluster be presented as a group of districts offering few job chances but the inhabitants do not accept travelling for jobs to other more distant territories (more than one hour travel time). Low workforce migration is typical for this cluster. All the border districts of Ústecký Region belong here, and also the district Jablonec nad Nisou, and Moravian districts Opava, Přerov, Prostějov, Uherské Hradiště and Hodonín.

The fifth cluster contains 9 districts (11.8%) and it reports above-average values in all the selected labour market indicators. This cluster has the highest value (1.66 times higher than the CR average) of the NJCV indicator. The cluster has considerably below-average values of all transport infrastructure indicators. RND (Railway network density) is at the lowest level here, of all the clusters, being only 0.65 multiple of the CR average. The districts in this cluster can be considered the territories with low economic efficiency. These are the districts

Louny, Příbram, Chrudim, Jeseník, Bruntál and five mutually neighbouring Moravian districts.

The sixth cluster covers 2 districts only (2.6%), see Plzeň-město and Brno-město. High values of transport infrastructure indicators are typical for these, as well as values of the indicators representing numbers of economic subjects, while values of unemployment indicators are low, here. The two districts can be considered areas with the best conditions for living as assessed from the viewpoint of economic indicators.

The seventh cluster includes 2 only districts, too, see Ostrava and Karviná. Compared with CR averages, these two have high values of unemployment indicators. These are the districts, where a considerable decline of mining industry was recorded in the years past. The AEOM (Average length of registration at Employment Office – males) indicator value represents 2.15 multiple of the CR average. These are the districts with low numbers of small economic subjects and the cluster is an antithesis of the sixth cluster, which means these are the districts where quality of life is affected by low level of economic development.

Grouping of districts for the Social Domain was based on six principal components, explaining 76.5% of total variance of the original starting variables. Based on the assessment of separate clustering stages and on the assessment of distances applied for the aggregation, we can state that, 76 CR districts are to be subdivided into six subgroups. Specificities of the separate district groups were, same as in the case of the Economic Domain, assessed based on the value quotients of selected indicators, on the overall CR averages (Tab. II)

The first cluster was composed of 25 districts (32.9%). These were districts with typically a low value of Popu (Population density) and NI (Natural increase) indicators as compared with the CR averages. All indicators representing medical care show below-average values in this cluster of districts. Considering territorial distribution of regions grouped in this cluster, we can assume that, medical care is concentrated in the neighbouring large cities, having a status of independent districts. As Fig. 2 shows it, with the exception of one district (Mladá Boleslav), most districts of Středočeský Region and of Plzeňský and Jihočeský Regions form the first cluster. It can be stated in general, that this group of districts rather belongs to the inferior ones as assessed from the viewpoint of social indicators of the quality of life and that there is a considerably strong decline in the population numbers.

The second cluster, representing 27.6% districts (21 districts in absolute terms), shows a slightly higher population density than the first one, and higher values concerning medical care, too. Comparing the remaining indicators, we can state that the districts grouped in this cluster show a higher quotient of the IA (Index of age) indicator and the cluster can be characterized as one with a low increase of population and low education level. Most of Moravian and Silesian districts have been included in this cluster (Fig. 2).

The third cluster (14.5%) shows above-average values (1.65 times the CR average) of the NI (Natural increase) indicator. From the viewpoint of a complex assessment of all the quality of life indicators selected, the districts in this cluster come to the forefront of a hypothetical ladder. These are districts with a high level of medical care and a high percentage of University graduates live here.

The fourth cluster represents 14.5% districts and as it concerns the averages of indicators studied in

comparison with CR average we can say, the cluster covers districts at lower levels of education and at an average level of natural increase. The cluster does not present any extremal value in the indicators of Social Domain and is mostly set up of border districts. The area is rather specific as to social situation, partly because of rather distant location from the centres of political, social and cultural life.

The fifth cluster contains five districts (6.6%): Most, Chomutov, Sokolov, Karviná and Ostrava; these can be placed among the backward and problem regions, negatively affected by the process of restructuring the industry, ie., mainly of mining industry. As to quality of life from the social viewpoint, these are standing close to the bottom level of the ladder. An extremely low level of natural increase is typical for them.

The last, sixth, cluster covers three districts (3.9%): Praha-východ, Praha-západ a Brno-město. An extremely high increase of population numbers is typical for these (6.14 times the CR average) as well as high population density. In case of Praha-východ and Praha-západ districts we cannot forget that the quality of life within these is connected with the links to Capital Prague, affecting the situation of whole the region. Young people who work in Prague arrive in these districts, aiming at an improvement of their accommodation conditions in the places close to the Capital. These people mostly have reached a higher education, too, what connects to the Quotient of University graduates variable value, reaching here very above-average values as compared with other clusters and the CR average (1.6 times).

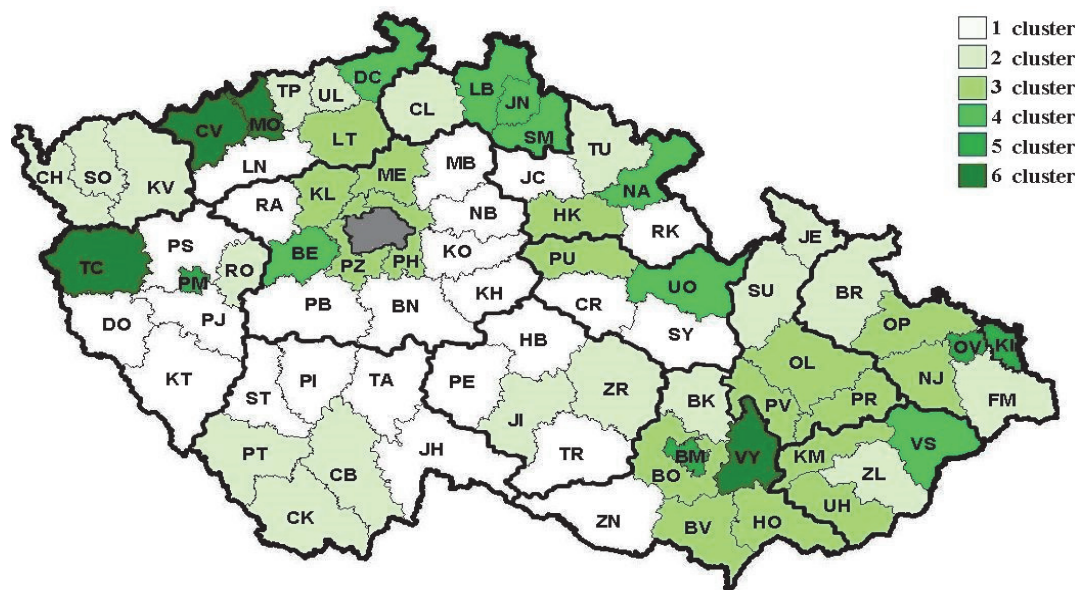
Clusters with similar values of life quality indicators for the Environmental Domain were found based on three principal components explaining 68.5% of the starting indicators variation. After assessment of outcomes of hierarchic grouping it has been stated that the family of CR districts rather clearly decomposes into six clusters of unequal sizes.

It is obvious from the Tab. III results that the first cluster, covering 24 districts (31.6% of the total number), shows an above-average value of the Arable land per inhabitant indicator. The cluster average prevailed 1.65 times the CR average, here. Low value of Ecologic stability coefficient, which is a quotient expressing the ratio of sizes of stable and unstable countryside-forming elements within

III: Quotients of selected indicator cluster averages on the Czech Republic averages

Cluster	Arable	QGO	Forest	QDA	ESC	Gas
1	1.65	0.79	0.90	0.83	0.71	0.71
2	0.68	0.66	1.33	1.22	1.60	0.98
3	0.89	1.33	0.71	0.98	0.47	1.30
4	0.50	1.07	1.27	0.40	1.88	0.84
5	0.09	2.66	0.63	0.24	0.50	1.61
6	1.03	0.63	1.14	3.01	0.88	1.26

Source: own computation



3: Czech Republic districts as related to clusters Environmental Domain  
Source: own computation

the area given, is also typical for this cluster of districts. Low values of this measure have been recorded in the areas with a considerable violation of natural cultures. It also means that, the districts where intensive farm production performs, have been included in the first cluster. The districts have suitable natural conditions and they are located in river-basins of big rivers and in the lowlands. Such a statement can be documented by the map showing adherence of districts to the clusters (Fig. 3).

The second cluster is represented by 19 districts (25%). High values of Ecologic stability coefficient (ESC) and of Quotient of forest land (Forest) are typical for this cluster and they are the positive indicators as concerns quality of life. On the other hand, the high value of the Quotient of dumping areas (QDA), standing here above the CR average, can be considered a negative factor for quality of life within this cluster.

A high percentage (22.4%) of districts has been included in the third cluster giving an above-average value of Quotient of gardens and orchards (QGO) that stands at the level 1.33 times of the CR average. This cluster also has high values of the indicator of level of gas supply to family houses and apartment houses, Quotient of permanently occupied family houses with gas (Gas) which is very favourable a factor as concerns cleanliness of atmosphere and also quality of life. However, if attention is paid to other indicators, it becomes obvious that districts in this cluster show very low values of the Ecologic

stability coefficient (ESC – expresses the ratio of the ecologically friendly areas to those areas that burden the environment) meaning that the quotient of unstable countryside-forming elements, such as built up areas, stands at a higher level, here.

The fourth cluster includes eight districts (10.5%) that stand above the CR average in the indicators Quotient of forest land (Forest) and Ecologic stability coefficient (ESC). Boundary districts form the greater part of this cluster. The Beroun district is the only exception, where 40% of the district area is covered by the protected areas CHKO Křivoklátsko and CHKO Český kras. The remaining districts in this cluster are situated in mountain areas.

The fifth and sixth clusters include by four districts each. The fifth cluster (5.3%) reports a minimum value of Arable land per inhabitant (Arable) and an above-average value of Quotient of gardens and orchards (QGO), reaching 2.66 times the level of CR average. Above the average stands in these districts also the value of (Gas) indicator. The districts Plzeň-město, Brno-město, Ostrava and Karviná belong in this cluster. Mostly they are the districts favourable as it concerns quality of life.

The sixth and last cluster (5.3%) is composed of Tachov, Chomutov, Most and Vyškov districts, connected as it concerns the quality of life assessment, by one of the negative indicators, Quotient of dumping areas (QDA). Its value gives a 3.01 times multiple of CR average, being at an extreme level within the Republic.

## CONCLUSION

Assessment of life quality and its development at the level of smaller areas is important for the establishment of disparity levels between regions. Practical importance of disparity assessment involves discovery of the negative factors and consequently it facilitates supplying support to the regions from the coffers of State, Region or Community administration bodies.

It is obvious from our results that cluster analysis offers a very suitable apparatus for multivariate extrapolation data analysis and its application facilitates assessment of regional disparities in all the domains of research. Based on the analysis it was possible to identify differences between regions and aggregate these into clusters reporting relatively homogeneous subclusters. (Nosek, Netrdová, 2010)

The optimal number of clusters was always chosen based on assessment of separate clustering stages represented in a dendrogram and on assessment of distances used at the stage of aggregation. The Economic and Social Domains were represented by seven clusters of districts each, the Environmental Domain had six clusters. Districts whose areas connect to regional capitals, such as the districts Plzeň-město, Brno-město as well as the districts Ostrava and Karviná that again belong to the structurally damaged ones, are often found within one common cluster. One disadvantage of cluster analysis when defining regional differentiation, appears in such a case, when districts are aggregated in a cluster and the chance to quantify accurately differences between separate districts is thereby lost. It does not facilitate establishing concrete numerical measures that could enable the researcher to discover differences between separate districts as concerns study of regional disparities in the life quality indicators.

## Acknowledgement

The author gratefully acknowledges the support from the Faculty of Economic and Management, Czech University of Life Sciences, via IGA grant, No. 11170/1312/3139.

## REFERENCES

- ANDRÁŠKO, I. 2013. *Quality of life: an introduction to the concept*. Brno: Masarykova univerzita.
- ANDREWS, F. M. and WHITHEY, S. B. 1976. *Social indicators of well-being: American's perceptions of life quality*. New York: Plenum.
- BINEK, J. and GALVASOVÁ, I. 2009. Regionální politika v ČR: Efekty a nové výzvy. In: *Efekty a nové výzvy: sborník z konference Regionální politika v ČR*. 5.–6. května. Jihlava: GaREP, 14–27.
- HEBÁK, P., HUSTOPECKÝ, J., PECÁKOVÁ, I. et al. 2007. *Vícerozměrné statistické metody [3]*. Praha: Informatorium.
- HUČKA, M. a KUTSCHERAUER, A. 2008. Metodologická východiska zkoumání regionálních disparit. In: *Mezinárodní kolokvium o regionálních vědách: sborník z kolokvia*. Pavlov 18.–20. června. [CD-ROM]. Brno: Masarykova univerzita, 8–14.
- CHARVÁT, O. a PETR, O. 2009. Kvalita života v příhraničních oblastech ČR. In: *Mezinárodní kolokvium o regionálních vědách: sborník z kolokvia*. Bořetice 17.–19. června. Brno: Masarykova univerzita, 214–226.
- JINDROVÁ, A. and POLÁČKOVÁ, J. 2012. Dimensionality reduction of quality of life indicators. *Acta Univ. Agric. Silvic. Mendelianae Brun.*, 60(7): 147–154.
- KÁBA, B. a SVATOŠOVÁ, L. 2012. *Statistické nástroje ekonomického výzkumu*. Plzeň: Aleš Čeněk, s. r. o.
- MELOUN, M. a MILITKÝ, J. 2006. *Kompendium statistického zpracování dat*. Praha: Academia.
- NOSEK, V. and NETRDOVÁ, P. 2010. Regional and Spatial Concentration of Socio-economic Phenomena: Empirical Evidence from the Czech Republic. *Ekonomický časopis*, 58(4): 344–359.
- ŘEZANKOVÁ, H., HÚSEK, D. a SNÁŠEL, V. 2007. *Shluková analýza dat*. Příbram: Professional Publishing.
- SVATOŠOVÁ, L., BOHÁČKOVÁ, I. a HRABÁNKOVÁ, M. 2005. *Regionální rozvoj z pozice strukturální politiky*. České Budějovice: JČU.
- VAĐUROVÁ, H. a MÜHLPACHER, P. 2005. *Kvalita života: Teoretická a metodologická východiska*. Brno: MSD Brno.

## Contact information

Andrea Jindrová: jindrova@pef.czu.cz