USING CLUSTERING IN GEOSCIENCES: EXAMPLES AND CASE STUDIES

M.Sc. Lukáš Marek*

M.Sc. Vít Pászto

M.Sc. Pavel Tuček, Ph.D.

Department of Geoinformatics, Faculty of Science, Palacky University in Olomouc; 17. listopadu 50, Olomouc, 771 46, **Czech Republic**

*Corresponding author: lukas.marek@upol.cz

ABSTRACT

Clustering methods are used in many scientific disciplines for various purposes. Geosciences also adopted principles of clustering and added a spatial component into clustering procedures. The aim of the clustering is to explore the spatial and/or attribute space of the data and then to find groups of similar objects. The similarity of objects, then, should be high within groups (clusters) while clusters themselves should be as different as possible. In this paper, two case studies using clustering methods on geospatial data are presented. In the first one, spatial clustering is applied on mortality rates in the Czech Republic. In the second case study, attribute clustering (based on fractal dimension) is performed in order to find clusters of river drainage systems. The paper offers examples of possible use of clustering methods in geosciences.

Keywords: geoinformatics, multivariate statistics, geocomputation, clustering, geodata

INTRODUCTION

Multivariate statistical methods, especially the clustering, are common in the variety of scientific fields that utilize their advances and complexity in the abstraction and modelling of real-world problems. Geosciences are not the exception since the amount of spatial and non-spatial data usually processed during geoanalyses seems to be feasible for the application of multivariate statistics. The combination of spatial information and analyses together with multivariate methods provides a strong tool for the complex evaluation of the geographical problems allowing the reduction of the dimension, spatial classification or the pattern analysis. One of the most used groups of multivariate statistics is the clustering. The aim of the clustering is to explore the spatial and/or attribute space of the data and then to find groups of similar objects. The similarity of objects, then, should be high within groups (clusters) while clusters themselves should be as different as possible.

In this paper, we present original examples and case studies where multivariate statistics is used with the emphasis placed on the clustering of geospatial data. We also applied geospatial tools in order to visualize the results. In the case studies, we tried to cover three main fields of the geography - human geography, physical geography and urban geography. The first case study deals with the mortality rates in the Czech Republic. Clustering methods helps to identify regions with similar mortality rate and its structure. The second case study is focused on the application of the fractal dimension calculation of river drainage systems and consequent clustering in order to quantify the similarity of individual river drainage systems. The clustering is based strictly on geometrical properties of the geomorphology rather than other characteristics. In both case studies, geographical interpretation of the results is discussed.

CLUSTER ANALYSIS

Cluster analysis is a group of methods whose purpose is to identify subsets of similar objects in disordered data. Therefore, it is a method of classification, which leads to the formation of classes. Methods usually provide suitable results especially when there are assumptions that the set of objects tend to be grouped into natural clusters [5]. At the end of the whole process, it is usually advisable to provide the characterization of each class and the interpretation of results.

There are more than one type of clustering that differ in processes of the class creating. The main clustering types are hierarchical (e.g. agglomerative or divisive) and non-hierarchical clustering [6, 12]. The type of the most suitable method can be chosen based on two main situation. The first situation is when the number of desired data classes is known, then e.g. k-means algorithm can be utilized. The second case is the unknown number of data classes, when the optimal number of classes is stated during the clustering process and analysis (e.g. hierarchical methods). The determination of similarities and differences within clusters can be realized on the basis of a large number of rules, such as the degree of similarity, degree of geographic distance or median. The results of hierarchical clustering can be visualized in the form of a dendrogram, where different groups are clearly expressed in the all levels of aggregation.

At the beginning of the clustering process, each multivariate record is considered a separate cluster (class). Then, the most similar records/clusters are merged together. On one hand objects belonging to different clusters should be as different as possible. On the other, objects that were once connected to the cluster, remain together. Cluster analysis and its individual variants are very important in the process of analysis of geodata. Besides the factor analysis and principal component analysis, they are the most common methods of multivariate statistics.

Examining the similarities of individual objects is based on the similarity of their properties as a whole and also on their proximity in space. The objective of cluster analysis is that elements within the cluster should as similar as possible, but simultaneously the individual clusters should differ. When creating a matrix of similarity among objects, the importance of individual factors can be taken into account using weighting (e.g. frequency of causes, geographical distance, etc.). Parameters for the conducting of cluster analysis were selected based on automated simulations.

GEODEMOGRAPHIC CLASSIFICATION OF CZECH MUNICIPALITIES REGARDING THE MORTALITY STRUCTURE (CASE STUDY I)

The main aim of the case study is to show the possible classification of municipalities in the Czech Republic according to the structure of mortality and causes of death. Furthermore, the classification was made in three years (1994, 2003, 2012) so the changes in the mortality structures can be easily followed and described. The case study combines methods of multivariate statistics and spatial analyses in order to provide the concept of geo-demographic evaluation of the mortality and its changes from near past to recent time.

The main data source in the case study were provided by the Czech Statistical Office, which records all information about the mortality of population including its structure, causes, etc. However, the detailed data are not provided because of its confidentiality and possible reidentification of individuals. Particularly, the data from 1994, 2003 and 2012 aggregated according to 20 main causes of death were analysed in the case study. The absolute or relative frequency of deaths were substituted by the estimate of the relative risk of death on the individual cause - standardized mortality rate. The relative risk was gained as the ratio between the absolute observed frequency of particular cause of death and the expected frequency estimated on the basis of standard/overall population of the Czech Republic and its mortality and the population in the municipality. Relative risk may result in non-negative values between zero and the infinity with some restrictions related to input values [2]. The estimation of relative risk supplied the standardization of mortality rates that were hardly possible to compute considering input data that did not contain the information about the age structure. According to some authors, the standardization of the data is not necessary in the case of the comparison of relative risk that is based on the overall characteristics of population [1]. The relative risk estimation of individual death causes was computed on the basis of Poisson probability distribution using R project and its package SpatialEpi [7].

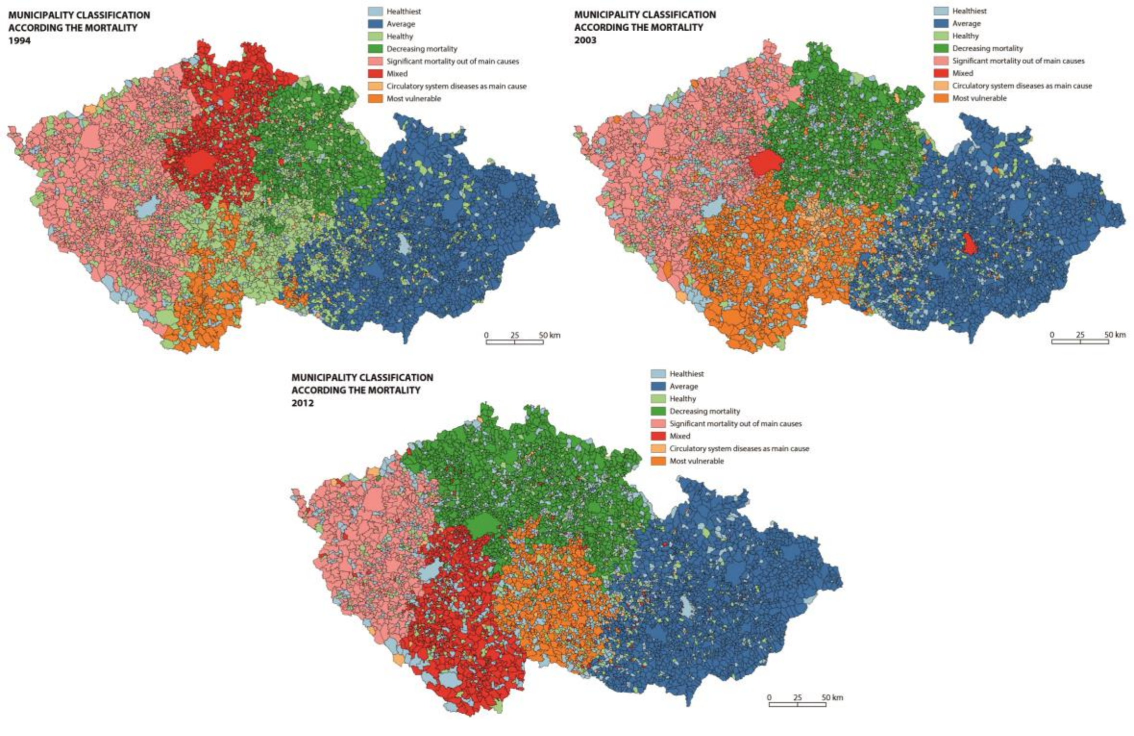
MUNICIPALITY CLASSIFICATION ACCORDING THE MORTALITY 1994
MUNICIPALITY CLASSIFICATION ACCORDING THE MORTALITY 2003
MUNICIPALITY CLASSIFICATION ACCORDING THE MORTALITY 2012
Healthiest
Average
Healthy
Decreasing mortality
Significant mortality out of main causes
Mixed
Circulatory system diseases as main cause
Most vulnerable
0 25 50 km

Fig. 1 Classification of Czech municipalities according to the mortality structure

The clustering in the case study was based on the similarity matrix. The squared Euclidean distance was used to calculate the similarity matrix among municipalities. Moreover, the similarity matrix was weighted by spatial distance matrix in order to get spatially contiguous regions. Specifically, the hierarchical clustering into eight groups using Ward's method was selected. The aim of Ward's method is to minimize heterogeneity in clusters according to criteria that minimize intra-sum-square. Results are depicted on the Figure 1.

Municipalities in the Czech Republic were categorized into 8 groups during analysed years. Groups characteristics are briefly described in Table 1 and groups are described verbally in the following section:

Tab. 1 Cluster characteristics (relative risk in %).

				1994				
Cause of death / Category	1	2	3	4	5	6	7	8
Circulatory system diseases	0	123	115	149	140	150	338	174
Neoplasms	28	108	108	124	148	146	50	180
External causes of morbidity and mortality	49	117	48	150	164	166	12	164
Respiratory diseases	0	147	40	202	228	138	20	343
Digestive diseases	0	155	34	152	166	158	6	206
Endocrine, nutritional and metabolic diseases	0	112	19	211	243	183	5	348
Nervous system diseases	0	117	15	160	199	237	24	384
Other causes	27	136	25	140	230	196	5	198
*Total relative risk**	*11*	*121*	*97*	*146*	*153*	*153*	*202*	*188*
Number of municipalities	**588**	**1458**	**1587**	**596**	**926**	**597**	**302**	**197**
				2003				
Cause of death / Category	1	2	3	4	5	6	7	8
Circulatory system diseases	31	120	153	129	117	5915	353	110
Neoplasms	53	107	78	118	132	0	55	187
External causes of morbidity and mortality	54	152	0	152	149	0	93	117
Respiratory diseases	64	107	0	222	147	0	92	176
Digestive diseases	5	132	3	132	164	0	57	204
Endocrine, nutritional and metabolic diseases	0	118	0	157	204	0	48	133
Nervous system diseases	9	112	0	144	188	0	74	168
Other causes	1	136	0	180	169	0	45	186
*Total relative risk**	*37*	*119*	*100*	*134*	*130*	*3083*	*214*	*141*
Number of municipalities	**1244**	**1491**	**705**	**856**	**921**	**2**	**271**	**761**
				2012				
Cause of death / Category	1	2	3	4	5	6	7	8
Circulatory system diseases	35	117	180	106	123	97	384	154
Neoplasms	55	104	95	105	131	167	5	140
External causes of morbidity and mortality	30	133	5	134	171	156	0	182
Respiratory diseases	67	123	19	137	135	140	0	195
Digestive diseases	9	142	10	146	120	113	12	159
Endocrine, nutritional and metabolic diseases	34	120	0	161	232	153	0	209
Nervous system diseases	9	137	0	127	153	215	0	192
Other causes	34	121	2	133	146	132	0	214
*Total relative risk**	*40*	*117*	*114*	*114*	*133*	*127*	*190*	*160*
Number of municipalities	**1351**	**1477**	**548**	**1211**	**491**	**598**	**135**	**440**

The healthiest municipalities - areas with the lowest relative risk in the most of death causes and low mean overall relative risk.

Average municipalities - municipalities with an average or slightly above-average relative risks in all death causes.

Healthy municipalities - municipalities with an average relative risk of death from neoplasms and slightly above-average relative risk the circulatory system diseases, while the mortality on the remaining causes is very low and average mortality corresponds to assumptions.

Municipalities with the decreasing mortality - the municipalities with gradually decreasing relative risk of mortality, which approaches the average value in the most causes of death.

Municipalities with a significant mortality out of main causes of death - the municipalities with the above-average relative risk of death especially among the causes of deaths that are different from diseases of the circulatory system and neoplasms.

Mixed group - a group of municipalities that varies in time that cannot be clearly defined.

Municipalities with diseases of the circulatory system as the predominant cause of death - the relative risk of death from diseases of the circulatory system significantly exceeds the average value, but relative risks of other causes of death are very low.

The most vulnerable municipalities - municipalities with high relative risks across the whole spectrum of causes of death.

CLUSTERING OF RIVER DRAINAGE SYSTEMS ACCORDING TO THEIR FRACTAL DIMENSION (CASE STUDY II)

The main aim of the second case study is to perform a cluster analysis of river drainage network systems. The idea was to cluster individual river drainage networks in order to create groups which would be corresponding with drainage systems types (such as dendritic, parallel, radial etc.). The clustering is based solely on fractal dimension values of the selected river networks. Fractal dimension is a description tool for measuring a complexity of a shape. It gained large popularity in many fields of natural sciences (e.g. [8, 14]), including e.g. ecology, geography, GIScience, where measures of object's shape are essential. There are several techniques how to measure fractal dimension and for this case study a box-counting method was used (for more details see e.g. [4]).

River drainage networks were obtained from free Internet source. Ten examples from each selected drainage system were selected and examined. Data are available from DIVA-GIS website (*http://www.diva-gis.org/Data*) and were used to select appropriate areas with typical drainage systems. There are several types of drainage systems, each of them typical for a certain type of relief [9]. In this case study, selected types of drainage systems were – dendritic, parallel, trellis, rectangular, radial and deranged (Fig. 2).

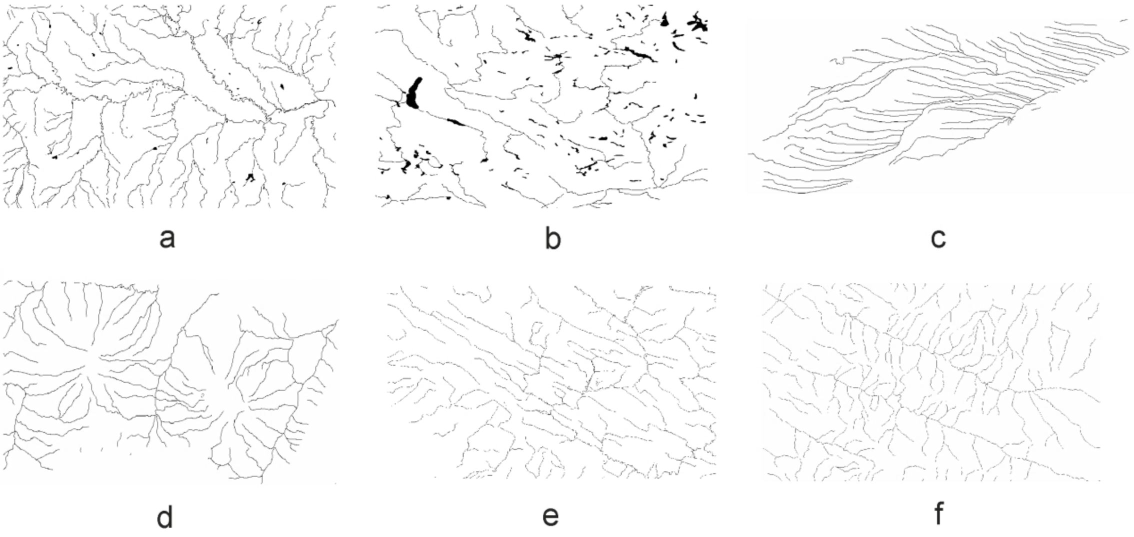

Fig. 2 Examples of selected river drainage systems, a – dendritic (Thailand), b – deranged (Poland), c – parallel (Canada), d – radial (Indonesia), e – rectangular (Iran), f – trellis (Brazil).

Fractal dimension calculation results are depicted in Table 2. Standard deviation of fractal dimension and ratio of black (representing rivers) and white pixels were also calculated.

Tab. 2 Fractal dimension calculation results (including its standard deviation and ration of black and white pixels) according to river drainage systems.

River drainage system	Average fractal dimension	Standard deviation	Black/white pixels ratio
Dendritic – A	1,4262	0,039286	0,052163
Deranged – B	1,5127	0,075419	0,117499
Parallel – C	1,3093	0,070454	0,021445
Radial – D	1,3905	0,091077	0,049157
Rectangular – E	1,2988	0,076465	0,021179
Trellis – F	1,2280	0,049390	0,014330

Then, the cluster analysis was performed using PAM (Partitioning Around Medoids) method, which is non-hierarchical and partitioning - meaning that dataset is broken up into desired number of groups using medoids (representative objects of a dataset, whose average dissimilarity to all surrounding objects is minimized). This method is similar to the K-means clustering, but K-means uses means or centroids to cluster a dataset. The PAM is treated to be more robust than K-means because of minimizing dissimilarity instead of Euclidean distances [10, 13]. Resulting clusters according to the two main components are depicted via 2D graph (Fig. 3).

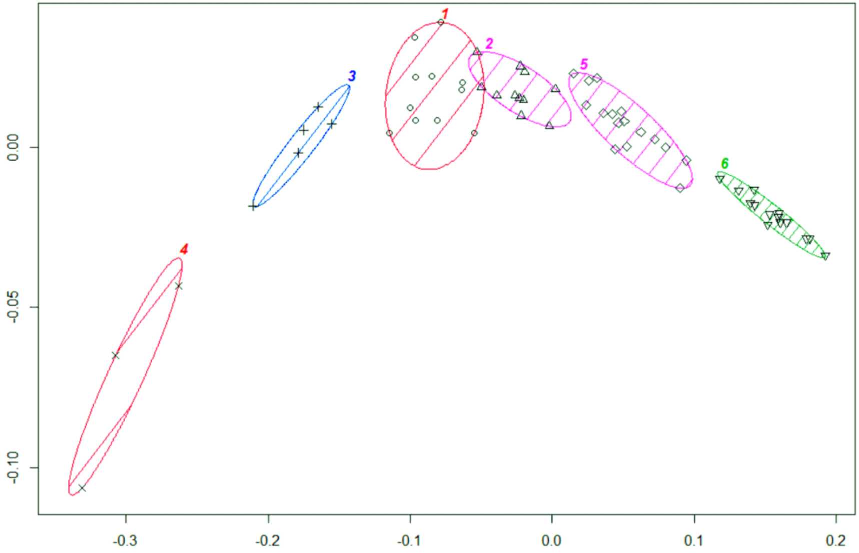

Fig. 3 Cluster analysis of river drainage networks using PAM method.

Cluster analysis was set to create six clusters, which should have been ideally formed by river networks from the same type (e.g. dendritic cluster, parallel cluster, trellis cluster etc.). Nevertheless, composition of clusters depicted in Figure 3 was rather heterogeneous than homogeneous. Particular composition of clusters is in Table 3.

Tab. 3 Cluster composition of river networks.

Cluster number	Cluster composition from particular river networks
1	A1, A3, A4, A7, B5, B8, B9, D1, D9, D10, E4
2	A2, A5, A8, A9, A10, B3, C3, C9, D4, D7, E8
3	A6, B4, B6, B7, B10
4	B1, B2, D2
5	C1, C2, C6, C7, C8, D3, D5, D6, D8, E6, E7, E9, E10, F5, F6, F10
6	C4, C5, C10, E1, E2, E3, E5, F1, F2, F3, F4, F7, F8, F9

Next step was to perform a cluster analysis only according to the average fractal dimension using hierarchical method (method which creates tree structure –

dendrogram) called Single. This algorithm constructs a hierarchy of clusters, starting with one large cluster containing all objects and then the cluster is divided until each cluster contains only a single object (more details in e.g. [3]). Cluster analysis was conducted for each drainage network type in order to examine the similarity of drainage systems. Results are depicted in Figure 4.

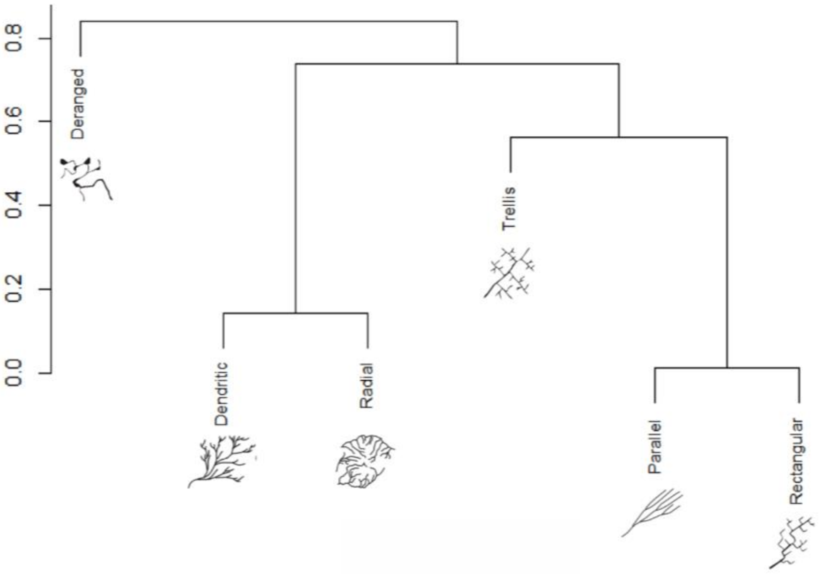

Fig. 4 Cluster analysis of river drainage types using Single method.

The most alike are parallel and rectangular drainage systems, which both embody quite similar patterns. Both of them (and also trellis drainage system) show regular patterns and their fractal dimension is the lowest. Radial and dendritic drainage systems embody more irregular shapes and therefore their fractal dimension is higher. The most irregular patterns are observed in deranged drainage system. Meanders, lakes and flood plain lobes are typical for deranged drainage systems and it is difficult to observe any distinguished structure in the shape of river network. The complexity of this type of drainage system is the greatest and analysis shows that the fractal dimension is also the highest. Cluster analysis shows that this type of drainage network is also the most dissimilar from any other drainage system types.

CONCLUSION AND DISCUSSION

The multivariate clustering is not a spatial method. Therefore, its results do not have to create spatially contiguous clusters. In fact, the situation is usually completely opposite and identified clusters (groups) of object are usually randomly placed because of the lack of spatial information [11]. The solution to this might be the weighting of the similarity matrix by the matrix of spatial relations (distance, contiguity), which helps to incorporate the First law of geography that say: "*Everything is related to everything else, but near things are more related than distant things.*" [15]. There are the space for objectivity and the analyst's experiences in the most of steps that are necessary in the process of clustering. Firstly, the suitable clustering procedure is needed to be chosen together with the selection of the type of the dissimilarity matrix that comes directly from the data character. Then, the number of clusters is opted. Lastly, the resulting clusters are described and interpreted.

In the first case study, classification of municipalities according to the mortality rate was performed. The result of the case study is the classification of the Czech Republic into eight categories based on their common characteristics of the mortality as well as their common location. However, there are no other socioeconomic characteristics and relations involved in the estimation of clusters. The second case study dealt with clustering of the most common types of river drainage systems. Clustering method PAM was used to classify individual river drainage networks according to their fractal dimension (and its standard deviation) and ratio of black/white pixels. This clustering brought rather unsatisfactory results. Then, clustering method Single was applied on river drainage types and was based only on fractal dimension calculation. Final clusters of river drainage systems were distinctive.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

REFERENCES

[1] Beale, L. et al.: Methodologic issues and approaches to spatial epidemiology. Environ. Health Perspect. 116, 8, 1105–10 (2008).

[2] Bencko, V. et al.: Statistické metody v epidemiologii. Nakladatelství Karolinum, Praha (2003).

[3] Florek, K. et al.: Sur la Liaison et la Division des Points d'un Ensemble Fini. Colloq. Math. 2, 3-4, 282–285 (1951).

[4] Hastings, H.M., Sugihara, G.: Fractals: A User's Guide for the Natural Sciences. Oxford University Press, Oxford (1994).

[5] Hebák, P. et al.: Vícerozměrné statistické metody 2. Informatorium, Praha (2005).

[6] Horák, J.: Prostorové analýzy dat. VŠB-TU Ostrava, HGF, Institut geoinformatiky, Ostrava (2011).

[7] Chen, C. et al.: SpatialEpi: Methods and Data for Spatial Epidemiology, http://cran.r-project.org/package=SpatialEpi, (2014).

[8] Kitchin, R., Thrift, N. eds: International encyclopedia of human geography. Elsevier Science (2009).

[9] Knighton, D.: Fluvial Forms and Processes: A New Perspective. Oxford University Press, New York (1998).

[10] Kumar, P., Wasan, S.K.: Comparative study of K-means, pam and rough K-means algorithms using cancer datasets. Proceedings of CSIT: 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009). pp. 136–140 (2009).

[11] Marek, L. et al.: On Estimation of the Spatial Clustering: Case Study of Epidemiological Data In Olomouc Region, Czech Republic. VŠB – Technická univerzita Ostrava, Ostrava (2013).

[12] Marek, L. et al.: Space-time evaluation of health data: Case of Olomouc area, Czech Republic. SGEM2013 Conference Proceedings. pp. Vol. 1, 911–918 STEF92 Technology Ltd., Sofia, Bulgaria (2013).

[13] Park, H.S., Jun, C.H.: A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl. 36, 2, 3336–3341 (2009).

[14] Peitgen, H.-O. et al.: Chaos and Fractals Mew Frontiers of Science. Springer, New York (2004).

[15] Tobler, W.R.: A Computer Movie Simulation Urban Growth in the Detroit Region. Econ. Geogr. 46, 332, 234–240 (1970).