On Estimation of the Spatial Clustering: Case Study of Epidemiological Data In Olomouc Region, Czech Republic

Lukáš Marek, Vít Pászto, Jiří Dvorský, Pavel Tuček

Department of Geoinformatics, Faculty of Science, Palacky University in Olomouc
17. listopadu 50, Olomouc, 771 46, Czech Republic
{lukas.marek, pavel.tucek}@upol.cz, vit.paszto@gmail.com,
jiri.dvorsky@vsb.cz

Abstract. An evaluation of spatial patterns and a clustering play an important role among methods of spatial statistics. However, traditional clustering techniques are seldom suitable for analyses of spatial data and patterns because they usually do not count on spatial relations and qualities of objects. This paper aims to introduce usage of methods of spatial clustering estimation, which are based mainly on the position of events and not only on the events attribute space. Firstly, the methods of the spatial clustering or randomness estimation are introduced and applied on a real dataset, then spatial clusters are identified and the intensity of processes is quantified. Non-spatial properties and a time are considered together with the location data. Also methods of the multivariate statistics are used for the purpose of the classification of regions with similar properties. Particularly, occurrence data of selected infectious diseases in Olomouc Region in period 2004 – 2010 provided by Regional Public Health Service in Olomouc are used for the case study.

1 Introduction

An application of the geographical information system, as well as a spatial statistics, for the exploration of the spatial pattern of health data has been highly discussed in the literature [2, 14, 18] and it is one of top topics in geosciences nowadays [5]. The literature also often uses phrases geographical epidemiology, spatial epidemiology, medical geography or even geomedicine that describe dynamic body of the theory and analytic methods concerned with the study of spatial patterns of the disease incidence and mortality [22]. Since John Snow's famous geographical study of the cholera in London in 1854 century, the interconnection of health data and analyses of spatial patterns became a standard procedure. Possibilities of current geographical information systems together with properties of databases, where health and epidemiological data are stored, make spatial evaluation easier than anytime before [13].

This contribution presents the usage of methods of spatial statistics applied on the epidemiological data from Olomouc region – one of administrative units of Czech

V. Snášel, K. Richta, J. Pokorný (Eds.): Dateso 2013, pp. 1–12, ISBN 978-80-248-2968-5.

Republic corresponding with NUTS3. Cases of one particular infectious disease – parotitis - are assessed within the period 2004-2010. Records about infection come from the database of infectious diseases, which is called EPIDAT. This database contains information about every single case of the infection which is reported by local doctors. EPIDAT also contains data about infected patients as the place of residence, the place of infecting, the time of treatment, the way of isolation etc. Although exact addresses of patients are reported, due to the necessity of preserving anonymity only approximate addresses are provided. That is why the exact geocoding is unfeasible and all records are aggregated into the regular fishnet or randomized within it.

2 Case study and Data

This section intents to describe data and their adjustment for the usage within techniques of Exploratory Spatial Data Analysis (ESDA) and Local Indicators of Spatial Association. Epidemiological data comes from the EPIDAT (EPIdemiological DATabase), which stores mandatory records about all infectious diseases in the Czechia. The case study is dealing mainly with spatial attributes of one selected disease – parotitis (i.e. mumps). Firstly, geocoded data are visualized in the form of (false) choropleth maps. Choropleth maps [12] allows to the researcher first, so called, "visual" analysis of spatial structure. Then quantitative analyses of possibilities of spatial clustering are proceeding – a kernel density estimation and G-function [3]. At last, data are aggregated and randomized into the regular hexagonal fishnet and spatial autocorrelation in local scale is explored using Moran's I, Getis Ord General G and LISA, which are comprehensively described in [1, 10]. Software tools used for the realization of case study was ArcGIS (choropleth maps, LISA) and R-project with packages for the spatial statistics, manipulating with spatial data and the visualization (spatstat, maptools, spdep, etc.).

The analyzed disease is a parotitis. Parotitis is an inflammation of one or both parotid glands, the major salivary glands located on either side of the face, in humans. The parotid gland is the salivary gland most commonly affected by an inflammation. The Mumps is an airborne virus and can be spread by an infected person coughing or sneezing and releasing tiny droplets of contaminated saliva, an infected person touching their nose or mouth, then transferring the virus onto an object, or sharing utensils [17]. Routine vaccinations have dropped the incidence of mumps to a very low level [15].

2.1 Data

The EPIDAT database is used to ensure the mandatory reporting, recording and analysis of infectious diseases in the Czech Republic. The database is used nation-widely by Public Health Service of the Czech Republic from 1st January 1993. The reporting of infectious diseases is the legal basis for local, regional, national and international control of infectious diseases (EU, WHO). The data storage is used to

secure exchange of actual data sets on the prevalence of infections among the departments of Public Health Service of CR, Ministry of Health of CR and Public Health Institute in Prague.

Total of 53 diagnoses of infectious diseases are monitored into the EPIDAT database. Each record contains 50 attributes. In terms of spatial analyses, the most important properties are information about the patient's residence as well as the place of infection and the place of sicken (the place where the patient became ill, often place of clinic or doctor's office).

The data set for this study was provided by the Regional Public Health Service in Olomouc. The original provided dataset contains 32 698 records of 11 selected infections from 53 diagnoses and covers the period 2004-2010, but only 958 records of one selected disease – parotitis. Because it is treated with sensitive personal data, the name, surname, identity number and full address is not included. Furthermore, geocoded data were randomized and anonymized using aggregation from the street network and municipality membership into the regular hexagonal fishnet, which is usual procedure in spatial epidemiology and econometrics [10]. The problem of aggregating the point based spatial phenomena into the district is well known as MAUP – Modifiable Area Unit Problem.

EPIDAT database is filled with data manually and it is a transcription of medical records. That situation guarantees the occurrence of errors, mistyped characters and different kinds of used abbreviations. A manual control and subsequent correction of mistakes could be time consuming and almost impossible because of the amount of records in database. That is why a tool for semi-automatic control, repairing and geocoding was developed as the result of collaboration between Department of Geoinformatics and Regional Public Health Service in Olomouc. Owing to this tool, whole process of repairing wrong records is time acceptable. Besides using our developed tool, the Google Geocoding API is used for geocoding addresses. Google Geocoding API substitutes the physical ownership of complete street data and the database of addresses and allows to geolocate with a suitable precision [23]. The process of geocoding of 32 698 addresses took 49 469 seconds (13.75 hours).

Several requisite steps have been done before any global or local analyses were executed. Firstly the data of selected infection diseases, as well as complete data, were randomized and aggregated into the regular hexagonal fishnet. Each hexagon has the area of 6.25 km^2, which is similar to the area of the cell established by Morishita index [16] and coincidentally, it is corresponding to the area of the average cadastral unit in the region. A number of inhabitants in hexagons was estimated from cadastral units and municipalities with usage of the areal weighting. Both, absolute and relative (prevalence on 1000 population) aggregation units were created but only the absolute occurrence entered the analysis. Secondly, spatial weight matrices were generated for each input of the disease. K-nearest neighbors method, with K = 12, was selected as the way of the spatial conceptualization, the other way is an assessment of the maximum threshold (distance) of possible connections among cases. The example of the spatial weight matrix is shown in the Figure 1.

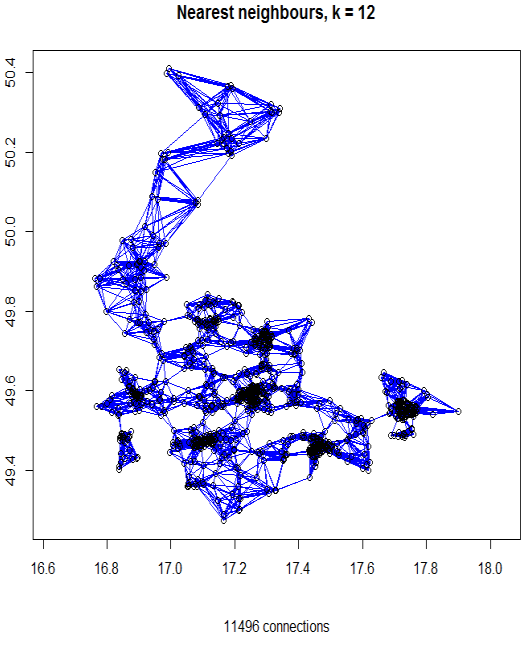

Fig. 1. The example of the generated spatial weight matrix. Points represent pseudo-randomized cases of the disease. Lines symbolize links among points, each case is connected to 12 other points / neighbors without taking into consideration of the distance

3 Methods

Analyses of spatial pattern of diseases occurrence, as well as their relations to potentially risk factors of the environment, are important parts of health studies. According to [7] three main broad areas of spatial epidemiology can be identified:

- disease mapping,
- geographic correlation studies,
- clustering, disease clusters, and surveillance.

The presented study is mainly focused on methods of the estimation and assessment of a spatial clustering as well as a multivariate clustering. Firstly, choropleth maps were constructed. Subsequently, the methods of the spatial clustering or randomness estimation were applied on the dataset and then spatial clusters are identified and the intensity of processes is quantified. Finally, the multivariate statistical clustering is executed, which extends the previously proceeded analyses.

Methods, which are introduced and described later in this chapter, are often covered by a common name Exploratory Data Analysis (EDA). This label

incorporates a wide group of statistical techniques that are very useful for both, an initial and deeper exploration of patterns and relations within the given data structure. In case of involving spatial metrics, the group of methods is called Exploratory Spatial Data Analysis (ESDA). The list of methods presented below in the text is not a complete enumeration, but only a brief overview of elementary techniques applicable to the problems and questions connected to the geographical space. It is worth to note, that most of E(S)DA methods are traditional techniques with basis in last century, but the progress in the (geographical) information science allows their wider usage.

3.1 Choropleth maps

Choropleth maps are probably the most common type of map for the display of areal data. These maps use different color and pattern combinations to depict different values of the attribute variable associated with each area, which is colored according to the category to which its corresponding attribute value belongs [22]. Although choropleth maps do not show continuously distributed values, they often portray densities [19]. Viewed in this way, one can consider them as a primitive visual tool for the analysis of spatial distribution of phenomenon.

3.2 Identification of Spatial Processes

A huge amount of methods for the estimation of the prevalent type of processes in the area are based on the testing of Complete Spatial Randomness (CSR) [6], the visual comparison of the plotted function with CSR or quadrat counts. E.g. quadrat test, G-function, K-function or Morishita index and Fry plot belong among these methods. Other suitable method is then a density kernel. It is appropriate to mention that the identification of spatial processes with the usage of previously presented methods is based mainly on the location of disease cases.

The evaluation of the spatial pattern of parotitis in this paper is realized by the plot of Morishita index and assesment of G-function. Morishita defined an index of spatial aggregation for a spatial point pattern based on quadrat counts. The spatial domain of the point pattern is first divided into Q subsets (quadrats) of equal size and shape [16]. The number of points falling in each quadrat is counted. If the pattern is completely random, index should be approximately equal to 1; values greater than 1 suggest clustering. Morishita plot is also helpful with an assessment of distances within clusters and also with the estimation of the pixel size or aggregating units in case of anonymization, for which the method of searching for the break point of the biggest change, so-called “elbow” method, is used.

G-function is the nearest neighbour distance distribution function (i.e. empirical cumulative distribution function of nearest neighbours). The shape of plotted G function is usually compared with the simulated envelope of random processes. This comparison allows unhiding the type of the possible spatial pattern. If the curve of G function takes place above the CSR envelope, then clustering is assumed. Its position

below the envelope means regular patterns and the position within the envelope refers to the random pattern [3].

3.3 Global and Local Spatial Clustering

Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locations on a two-dimensional (2-D) surface, introducing a deviation from the independent observations assumption of classical statistics [9]. Tobler's first law of geography encapsulates this situation, *"everything is related to everything else, but near things are more related than distant things".* The positive spatial autocorrelation refers to patterns where nearby or neighboring values are more alike; while the negative spatial autocorrelation refers to the patterns where nearby or neighboring values are dissimilar. One can distinguish two main types of spatial autocorrelation, which are global and local autocorrelation.

These techniques are collectively denoted as Exploratory Spatial Data Analysis (ESDA) and Local Indicators of Spatial Association (LISA), which are widely spread in geosciences and GIS software. Comprehensive description of theory, as well as detail examples of usage, can provide e.g. [1] or [10].

3.4 Multivariate Clustering

While the spatial clustering creates groups, which are based mainly on the similar location or the location and one common characteristic, methods of the multivariate statistics deals with an inverse situation. Thus, an aim of multivariate clustering is to categorize set of object with the emphasis on their quantitative and/or qualitative characteristics but mostly without implementing spatial dependencies [20], albeit several attempts for the combination of both approaches have appeared in recent years [4, 11].

For the purpose of this study, several methods of multivariate clustering, which were evaluated as the most suitable by simulations, were performed. Firstly, the similarity among all cases of parotitis through time is evaluated using the hierarchical clustering method with average linkage based on the Sokal-Michener dissimilarity distance measure for nominal variables [21]. Then similarity among hexagonal areas with aggregated values is calculated using Partitioning Around Medoids (pam) clustering algorithm based on the squared Euclidean dissimilarity distance measure. Moreover, DBSCAN algorithm - A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [8] – enables a different approach to the estimation of clusters, which is based on the internal density of clusters and is highly suitable for the usage with spatial database. The algorithm provides results, which are visually similar to the density kernel in case of incorporating simple locations. But it is more efficient in minimal requirements of domain knowledge to determine the input parameters, discovery of clusters with arbitrary shape and good efficiency on large databases [8].

4 Results

Two choropleth maps are constructed for the purpose of case study (Fig. 2). The first map (*left part*) expresses occurrences (i.e. absolute number) of parotitis in the municipalities of Olomouc region between January 2004 and December 2010. Darker areas mean that higher absolute number of cases was reported from the area. The second map expresses relative measure – prevalence, i.e. the number of cases of parotitis in the population of municipality. Both maps show that the more populated southern part of region is more affected by the parotitis than the northern part because the darkest areas in the first map match the biggest towns in the region. But second map is more particular and allows specifying of several centres.

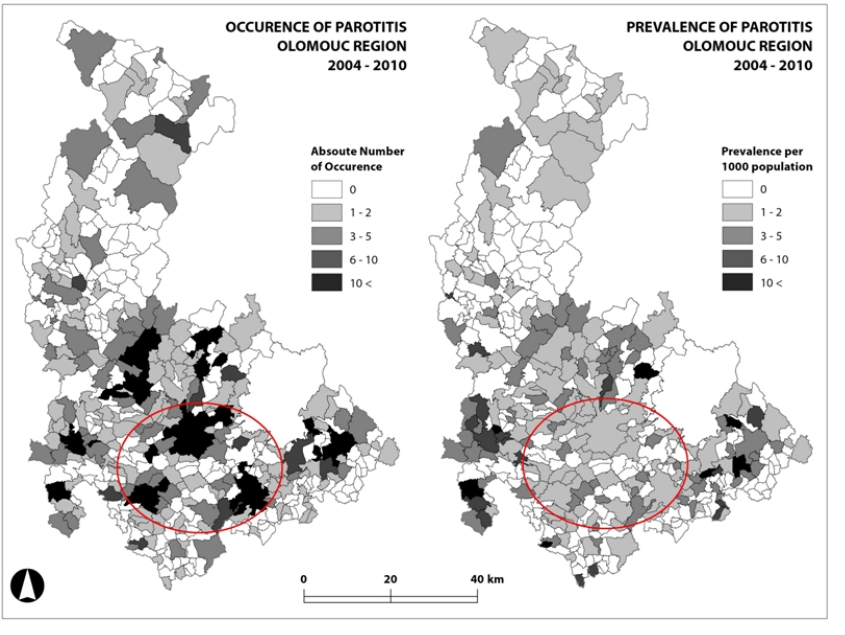

Fig. 2. Visualization of absolute (occurence - *left*) and relative (prevalence - *right*) values. Ellipses indicate places with high probability of misinterpretation

Results of Morishita index plot and the comparison between G-function and a simulated CSR process (Fig. 3) prove previously predicted fact that a possibility of clustering exists in several scales in the area. With the knowledge introduced before, it is evident that according to MI plot (Fig. 3 *left part*) clustering processes dominate in the area because the progress of function descends rapidly in the first part and after the "elbow point" it starts to converge to 1. G function (Fig. 3 *right part*) is compared with the envelope of CSR after 1000 simulations. The full line expresses observed value of function G for data pattern, dotted line stands for simulated CSR and light grey regions is 96% envelope of CSR. The curve of G function appears above the CSR line up to the value 0.05, which points out to clustering processes again.

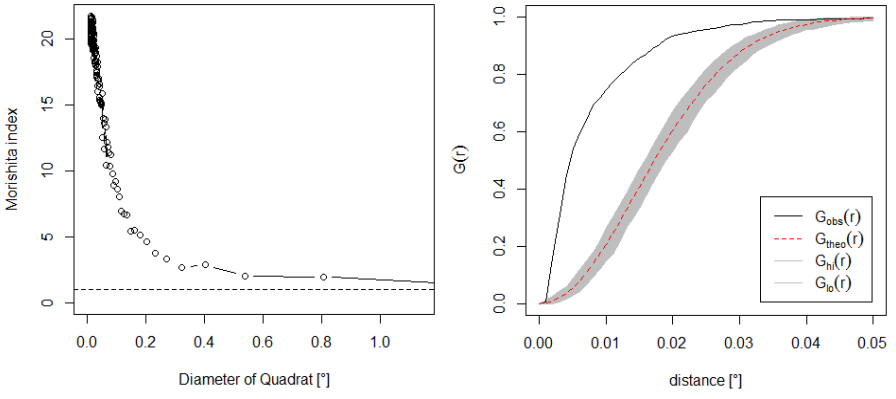

Fig. 3. Morishita index plot (*left*) and the curve G-function (*right*)

Density map is the method for quick visualization of significant clusters similar to choropleth maps or quadrat maps. The most important aspect in the estimation of kernel is not the type of kernel but its size, called bandwidth. In this study the quartic kernel with fixed bandwidth of size 0.01° (≈ 1 km) is used. This distance comes from the cross validated bandwidth selection for kernel density of point processes. Density map (Fig. 4 – *left*) then reveals primary spatial clusters, which are similar to those previously mentioned.

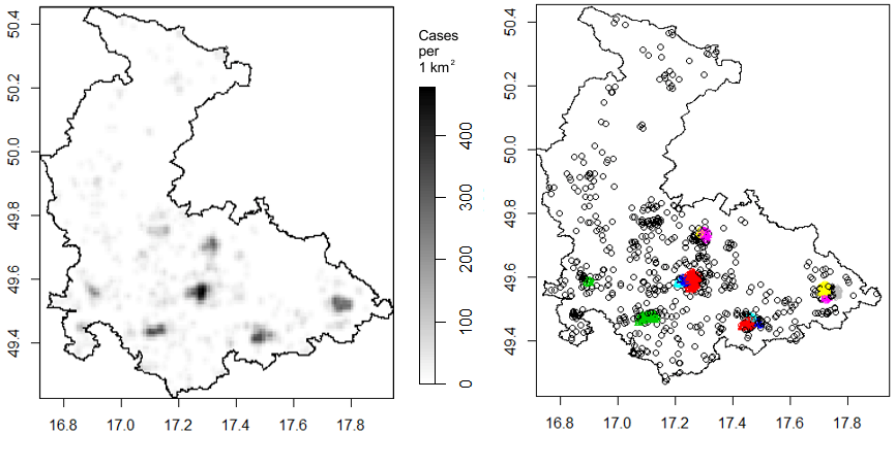

Fig. 4. Estimation of the number of cases by Kernel density (*left*) enables visual estimation of areas with dense occurrence of infection. The result of clustering by DBSCAN algorithm (*right*) –several clusters are identifiable in the southern part of study area or even in body of individual towns

Results of Moran's I, as well as Getis Ord G, prove to the presence of clustering processes. A neighbourhood for the analysis of local clustering and the size of aggregated units is based on the previous calculation of Morishita index. The global existence of clusters in the study area is proven by several methods, but their location is still not known. That is why it is proceeded to LISA.
Particular localizations of significant clusters, as well as their type, are then shown in the Figure 5. Main intensive clusters (grey areas) of high values (netted areas) with similar size of the area are identified near biggest towns in the southern and central part of the study area, while municipalities in the northern part of the region do not show any significant clusters or outliers. Clusters of high values are dominant in the area but also one outlier with low values appears.

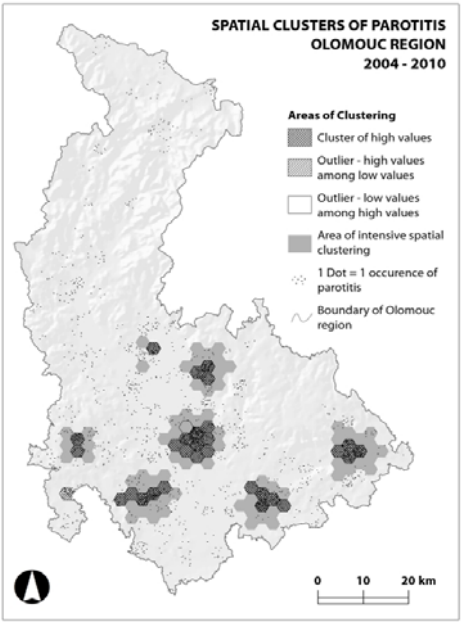

Fig. 5. Localization of disease clusters based on places of occurence of parotitis with usage of Local Indicators of Spatial Association (LISA)

For the purpose of this study, three methods of multivariate clustering, which were evaluated as the most suitable, were performed. Firstly, the similarity among all cases of parotitis through time is evaluated using the hierarchical clustering method with average linkage based on the Sokal-Michener dissimilarity distance measure for nominal variables (Fig. 6 *left part*). Then similarity among hexagonal areas with aggregated values is calculated using Partitioning Around Medoids (pam) clustering algorithm based on the squared Euclidean dissimilarity distance measure (Fig. 6 *right part*). Especially second case is useful, because it found 3 categories with a similar attribute space (categories 2-4), which corresponds with main towns in the study area and furthermore it divides these towns in separate classes.

Result of the third method is shown in the Figure 4 (*right part*). DBSCAN algorithm is highly effective for the estimation of spatial clusters even in the noise data. Fourteen clusters are found in the health data in the case study. These clusters not only correspond with settlements in the area of interest but also express inner differences in clusters and divide town into separate zones.

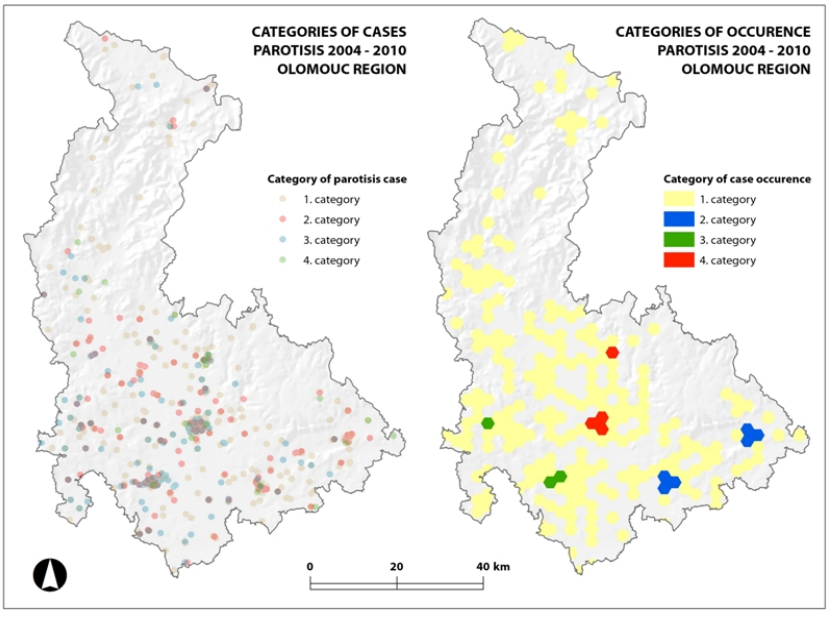

Fig. 6. Spatial visualization of multivariate clustering - similar cases (*left*), similar locations (*right*)

5 Discussion and Conclusion

One can easily explore spatial pattern and spatial relations with usage of spatial statistics and especially methods of spatial clustering estimation. But results of spatial statistics and mainly their interpretation are usually experience-dependent, i.e. subjective. Plenty of spatial techniques are also scale-dependent, thus their results are very sensitive on the precise adjustment of parameters and one small change can cause important differences in results. At least one example is given in this paper, which is an estimation of the kernel for density maps. Both, global and local indices of spatial autocorrelation are sensitive on the selection of parameters as well. The evaluation of spatial patterns is influenced also by others environmental factors, which some of them may be primarily hidden. Demographic factors are mostly the most influential element in case of health datasets.

Earlier presented procedures and techniques are only sample of suitable tools for spatial statistics, their further development and implementation lead not only to

spatial but to space-time and space-time-attribute analyses. Which are very complex and I believe they will be the predominate type of analyses in near future.

This paper presented techniques of the exploration of the spatial pattern. The usability of methods has been proved on the case study. An exploration of spatial patterns of the occurrence of parotitis (mumps) in the Olomouc Region was chosen as the model situation. Methods of EDA and ESDA confirmed the hypothesis that clustering processes exist in the area. Firstly the randomness of occurrence was tested and then the predominant type of the process was searched with a help of visualizing methods (choropleth maps and density maps) and tested (Morishita index plot). The spatial autocorrelation was explored on the aggregated data in the form of regular hexagons. Several clusters have been identified using methods of LISA. This clusters and their description are depicted on the map (Fig. 4). Clusters of high values with intensive processes are prevailing in around the towns Olomouc, Hranice, Přerov, Prostějov, Šternberk a Konice. Spatial statistics allows outstanding possibilities of exploration of spatial and space-time patterns, although some of methods have their strict limits and their interpretation can be subjective and experience dependent.

At last, data were evaluated by methods of multivariate statistics, which served to the searching of similar cases and regions with similar characteristics of disease occurrence.

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. Anselin, L.: Local indicators of spatial association—LISA. Geographical analysis. 27, 2, (1995).
2. Bergquist, R.: New tools for epidemiology: a space odyssey. Memórias do Instituto Oswaldo Cruz. 106, 7, 892–900 (2011).
3. Bivand, R.S. et al.: Applied Spatial Data Analysis with R. Springer New York, New York, NY (2008).
4. Carvalho, A. et al.: Spatial Hierarchical clustering. Rev. Bras. Biom. 27, 3, 411–442 (2009).
5. Davenhall, B.: Geomedicine: Geography and Personal Health. Esri, Redlands (2012).
6. Dixon, P.M.: Ripley 's K function. In: El-Shaarawi, A.H. and Piegorsch, W.W. (eds.) Encyclopedia of Environmetrics. pp. 1796–1803 (2002).
7. Elliott, P., Wartenberg, D.: Spatial Epidemiology: Current Approaches and Future Challenges. Environmental Health Perspectives. 112, 9, 998–1006 (2004).
8. Ester, M. et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International …. (1996).

9. Griffith, D., Arbia, G.: Detecting negative spatial autocorrelation in georeferenced random variables. International Journal of Geographical Information Science. 24, 3, 417–437 (2010).
10. Haining, R.: Spatial Data Analysis: Theory and Practice. Cambridge University Press (2004).
11. Horák, J. et al.: Methods of Spatial Clustering in a City. Geografie a Geoinformatika - Výzva pro praxi a vzdělávání. pp. 1–11 (2011).
12. Koch, T.: Cartographies of Disease: Maps, Mapping and Medicine. ESRI Press, Redlands, CA (2005).
13. Marek, L. et al.: Spatial Analyses of Epidemiological Data: Case Study In Olomouc Region. 12th International Multidisciplinary Scientific GeoConference SGEM: SGEM 2012, Proceedings Volume II. pp. 1155 – 1162 STEF92 Technology Ltd, Sofia, Bulgaria (2012).
14. Meade, M.S., Emch, M.: Medical geography. The Guilford Press, New York, NY (2010).
15. Medscape-Reference: Parotitis, http://emedicine.medscape.com/article/882461-overview.
16. Morishita, M.: Measuring of the dispersion of individuals and analysis of the distributional patterns. Memoir of the Faculty of Science. pp. 215 – 235 Kyushu University (1959).
17. NHS, Choices: Mumps - Causes, http://www.nhs.uk/Conditions/Mumps/Pages/Causes.aspx.
18. Ricketts, T.C.: Geographic information systems and public health. Annual review of public health. 24, 1–6 (2003).
19. Rushton, G.: Public health, GIS, and spatial analytic tools. Annual review of public health. 24, 43–56 (2003).
20. Tabachnick, B., Fidell, L.: Using multivariate statistics. Pearson (2007).
21. Walesiak, M., Dudek, A.: Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu. Zeszyty Naukowe Uniwersytetu Szczecinskiego. 450, 635–646 (2007).
22. Waller, L.A., Gotway, C.A.: Applied Spatial Statistics for Public Health Data. John Wiley & Sons (2004).
23. Batch geocoder, http://mapsapi.googlepages.com/batchgeo.htm.