Spatial Clustering of Disease Events Using Bayesian Methods

Lukáš Marek, Vít Pászto, Pavel Tuček, and Jiří Dvorský

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
http://www.geoinformatics.upol.cz/

Abstract. One of main aims of the spatial analysis of health and medical datasets is to provide additional information to the specialized medical research. These analyses can be used for disease mapping; searching for places with a higher intensity and probability of the disease event; or the influence assessment of selected natural or artificial phenomena. Suitably selected methods allow a proper analysis of these data and identification of irregularities and deviations of the phenomena in the area of interest. The structure of medical data usually needs to be standardized (over age structure of the population) before the comparison of different regions. Bayesian statistics derives the posterior probability as a consequence of a prior probability and a probability model for the data observed. Geosciences and geomedicine usually use the Bayesian theory for smoothing of data - to help depict the real spatial pattern and its changeability. The Bayesian principles, together with the spatial neighbourhood and statistical models, are successfully used also for the identification of spatial and space-time clusters with significantly higher/lower risk of incidence of the disease. These procedures are denoted as methods of spatial clustering and can be used with or without utilization of properties of certain phenomena. Particularly, occurrence data of campylobacteriosis infection in four Moravian regions in period 2008 – 2012, which were provided by The National Institute of Public Health, were used for the case study.

1 Introduction

The disease mapping, visualization of disease rates and the clustering of disease data are still one of the most interesting topics in geosciences. It is because of the nature of the data which are often pure spatial with rich descriptive part and it is easy to combine them with other data (demographic, economic, etc.) [14]. This contribution aims to present the usage of empirical Bayesian methods in the disease mapping and subsequent creating of disease maps. Bayesian methods incorporate the prior knowledge about the phenomenon (or underlying processes) to provide more accurate and easily understandable description of the situation. Empirical Bayesian procedures are used for disease rates smoothing in the case of choropleth map. They also help to identify local clusters of more/less affected areas. The main topic of the case study in this paper is the analysis of the spatial distribution of disease called campylobacteriosis in

J. Pokorný, K. Richta, V. Snášel (Eds.): Dateso 2014, pp. 25–34, ISBN 978-80-01-05482-6.

Moravian regions between years 2008 and 2012 with usage of Bayesian estimates based on Poisson distribution.

2 Case study and Data

The case study, where further described methods are applied, is dealing with the spatial distribution of the campylobacteriosis in four Moravian regions (Moraskoslezsky, Zlinsky, Olomoucky and Jihomoravsky) between years 2008 and 2012. There were almost 49 thousand of cases of the disease during that period, while only 34 thousand were expected according to previous records. Using disease counts and disease rates calculated for the municipalities in the area of interest, we tried to identify areas that are possibly more vulnerable to the disease than their neighbourhood. The 5-year observed number of cases, expected number of cases and relative risk (SIR) were used as main disease characteristics for this study.

Campylobacteriosis is caused by bacteria called Campylobacter jejuni, which is found worldwide in the intestinal tracts of animals. The bacteria are spiral shaped and can cause disease in animals and humans. Most cases of campylobacteriosis are associated with handling or eating raw or undercooked poultry meat or fresh milk. Campylobacteriosis causes gastrointestinal symptoms, such as diarrhoea, cramping, abdominal pain, and fever in domestic animals and humans. Young animals and humans are the most severely affected [23].

2.1 Data

The data set for this study was provided by The National Institute of Public Health of the Czech Republic. The database contains almost 50 thousands records of the campylobacteriosis occurrence in the period 2008 – 2012. Names, surnames, identity numbers and sometimes also the full addresses are not included because it is treated with sensitive personal data. The data were firstly cleansed of inconsistencies and then the geocoding process was run. Furthermore, the individual records were aggregated to the municipalities - administrative units - due to the clarity of the visualization and analyses [15]. The problem of the conversion of spatial phenomena between different areal or administrative units is well known as MAUP – Modifiable Area Unit Problem [18]. During the calculation of disease rates and expected number of cases, the population data from the Population and Housing Census of the Czech Republic were used as the main basis for the data standardization.

Figure 1 shows the probability density function of disease events counts, total population and standardized incidence ratio in Moravian municipalities visualized in the logarithmic scale (upper graph) and in the logarithmic scale and centred (lower graph) in order to simplify visual analysis. The probability function of population and diseases events counts are fairly similar, which indicates the need for standardization and also analysis that considers this close relation. Figure 2 then depicts the spatial distribution of standardized incidence ratio (SIR). SIR is the ratio of the number of disease cases observed in the study group or population to the number that would be expected if the study population had the same specific rates as the standard population, multiplied by 100 and usually expressed as a percentage [10]. By this way, SIR expresses

relative risk (or vulnerability) of the municipality to certain disease. Municipalities traversing value of 1 are more vulnerable to disease, while municipalities with SIR lower than 1 are healthier.

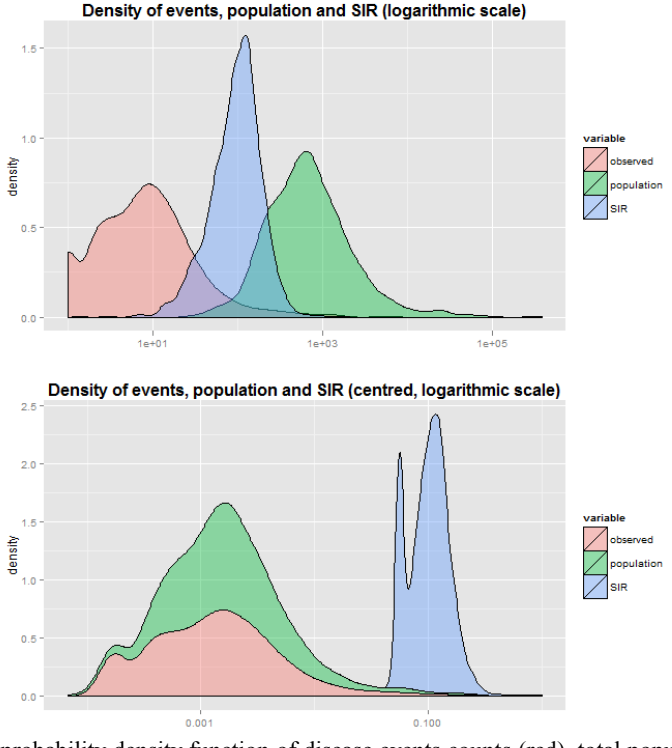

Fig. 1. The probability density function of disease events counts (red), total population (green) and standardized incidence ratio (blue) in Moravian municipalities visualized in the logarithmic scale (upper graph) and in the logarithmic scale and centred (lower graph) in order to simplify visual analysis.

3 Methods

During the study of disease spatial distribution, mainly in the case of aggregated data, it is often suitable to focus on the local variability of the disease occurrence or relative risk rather than examine the study area as a whole. This procedure is usually denoted the disease cluster detection. The general review of methodology as well as usage of spatial clustering methods and its Bayesian enhancements in the literature, e.g [6, 11, 21] etc.

In geosciences the spatial clustering is often encapsulated as the analysis of the spatial autocorrelation. The spatial autocorrelation is the correlation among values of a single variable, which is 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 [7]. Positive spatial autocorrelation refers to the patterns where nearby or neighbouring values are more alike; while negative spatial

autocorrelation refers to the patterns where nearby or neighbouring values are dissimilar. One can distinguish two main types of spatial autocorrelation, which are global and local. The null hypothesis for global clustering is simply that no clustering exists (i.e. random spatial dispersion ≈ CSR). Probably the most used method for both global and local analyses of spatial autocorrelation is Moran's I statistics (together with e.g. Getis-Ord G and Geary's C statistics). Moran's I coefficient of autocorrelation is similar to Pearson's correlation coefficient, and quantifies the similarity of an outcome variable among areas that are defined as spatially related [16]. The problem with variance instability for rates or proportions, which served as the motivation for applying smoothing techniques to maps may also affect the inference for Moran's I test for spatial autocorrelation [1]. The implementation of the adjustment procedure of Assuncao and Reis (1999), which uses a variable transformation based on the Empirical Bayes principle may be one of solutions. This yields a new variable that has been adjusted for the potentially biasing effects of variance instability due to differences in the size of the underlying population at risk [1].

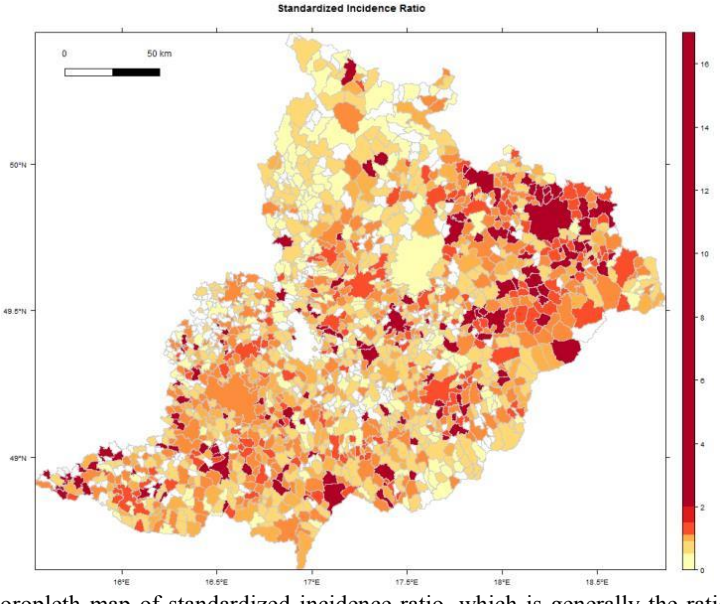

Fig. 2. Choropleth map of standardized incidence ratio, which is generally the ratio between observed disease cases and its potential (expected) amount, which is based on the population and its age structure in individual municipalities.

3.1 Spatial clustering of case events data

If a cluster is described as an uncommon collection of events, then it is needed to detect these collections observed within the data set. Such methods define a set of potential clusters, collections of events each of which we might define as a cluster if the collection appears unusual enough (discrepant from the null model of interest), then identifies the most unusual of these [21]. This general idea motivated the "geo-

graphical analysis machine" (GAM) of Openshaw where potential clusters were defined as collections of events falling within circular buffers of varying radii [17]. The buffers were centred at each point in a fine grid covering the study area and the GAM approach mapped any circle whose collection of events were detected as unusual, e.g., those circles where the number of events exceeded the 99.8th percentile of a Poisson distribution with mean defined by the population size within the buffer multiplied by the overall disease risk [21]. GAM is very useful for descriptive purposes, but should not be used for hypothesis testing.

Scan statistics provide another approach that is similar to the local case/control ratios. A scan statistic involves definition of a moving window and a statistical comparison of a measurement (e.g., a count or a rate) within the window to the same sort of measurement outside the window. Kulldorff [8] defines a spatial scan statistic very similar to the GAM and other methods, but with a slightly different inferential framework. The primary goal of a scan statistic is to find the collection(s) of cases least consistent with the null hypothesis, i.e. the most likely cluster(s) but Kulldorff goes a bit further and seeks to provide a significance value representing the detected cluster's unusualness, with an adjustment for multiple testing [22]. Kulldorff [8] considers circular windows with variable radii ranging from the smallest observed distance between a pair of cases to a user-defined upper bound. He builds an inferential structure based on earlier works where authors note that variable-width one-dimensional scan statistics represent collections of local likelihood ratio tests comparing a null hypothesis of the constant risk hypothesis compared to alternatives where the disease rate within the scanning window is greater than that outside the window. The maximum observed likelihood ratio statistic provides a test of overall general clustering and an indication of the most likely cluster(s), with significance determined by Monte Carlo testing of the constant risk hypothesis [22].

The outstanding description of methods including their mathematical apparatus or their possible implementations and applications provide mainly [8, 9, 17].

3.2 Bayesian mapping and spatial clustering of case events data

Presentation of disease rates in area units as choropleth maps can inadvertently provide misleading information. This fact is well known mainly in the case of small-area studies that introduces an extra source of variability into the map because of random variation. Typically, sparsely populated areas with few (or zero) cases can generate extreme values of the SMR (and also prevalence), as the variance of the SMR is inversely related to expected number of cases and small populations have large variability in the estimated rates [5] and that is why risk estimates and other rates are rather unstable.

Bayesian methods provide a solution for this kind of bias. They use probability models to obtain smoothed estimates consisting of a compromise between the observed rate for each region and an estimate from a larger collection of cases and persons at risk (e.g., the rate observed over the entire study area or over a collection of neighbouring regions) [22]. The basic principle of Bayesian methods is that uncertain data can be strengthened by combining them with prior information [19]. In the case of

empirical Bayes estimation of spatially-varying disease risk, posterior risk can be estimated from a weighted combination of the local risk (also called the likelihood) and the risk in surrounding areas, the latter representing the prior information [4].

The set of areal units on which data are recorded can form a regular lattice or differ largely in both shape and size, so data typically exhibit spatial autocorrelation, with observations from areal units close together tending to have similar values. A proportion of this spatial autocorrelation may be modelled by including known covariate risk factors in a regression model, the residual spatial autocorrelation can be induced by a number of factors, and violates the assumption of independence that is common in many regression models [12]. The most common remedy for this residual autocorrelation is to augment the linear predictor with a set of spatially correlated random effects, as part of a Bayesian hierarchical model. These random effects are typically represented with a conditional autoregressive (CAR) model, which induces spatial autocorrelation through the adjacency structure of the areal units. However, the CAR priors force the random effects to exhibit a single global level of spatial autocorrelation, ranging from independence through to strong spatial smoothing. Such a uniform level of spatial smoothness for the entire region is unrealistic for real data, which are instead likely to exhibit sub-areas of spatial autocorrelation separated by discontinuities. Such localized spatial smoothing may occur where rich and poor communities live side-by-side, and in this context the response variable is likely to evolve smoothly within each community with a sudden change in its value at the border where the two communities meet [12].

To be more particular, the analysis provided in the case study is based on the function that fits a Poisson log-normal random effects models to spatial count data, where the random effects are modelled by the localised conditional autoregressive (CAR) model proposed by [13]. The random effects in neighbouring areas (e.g. those that share a common border) are modelled as correlated or conditionally independent, depending on whether the populations living in the two areas are similar (correlated random effects) or very different (conditionally independent). The model represents the natural log of the mean function for the set of Poisson responses by a combination of covariates and a set of random effects. Inference is based on Markov Chain Monte Carlo (MCMC) simulation, using a combination of Gibbs sampling and Metropolis steps [12]. The outstanding overview of Bayesian techniques are provided in [11, 12] and others.

4 Results

Firstly, the original data of disease events needed to be aggregated to the municipality level, filtered to selected area of the Czech Republic. Subsequently, aggregated counts that represented actually observed cases served as the bases for the calculation of expected number of cases in the area that were found out using internally indirect standardization. SIR, which is the ratio between observed and expected number of cases and expresses the relative risk of the area can be seen in the Fig. 2.

Both values served as inputs for the Openshaw's Geographical Analysis Machine that allowed the identification of possible disease clusters in the area. Radius for the analysis was chosen as 7 km, the alpha value for the cluster identification was 99.8

quantile of the Poisson probability distribution. Several significant clusters can be identified throughout the study area (Fig. 3), but 2 most visible can be seen – the first is located in the southern part of the area near the Brno municipality, while the second cluster is placed in densely populated surroundings of Ostrava (but surprisingly except the city itself). As it was mentioned before, GAM is very useful for descriptive purposes, but should not be used for hypothesis testing because of the overestimation of clusters. That is why other methods - scan statistics and Bayesian identification of inference in the area, were performed.

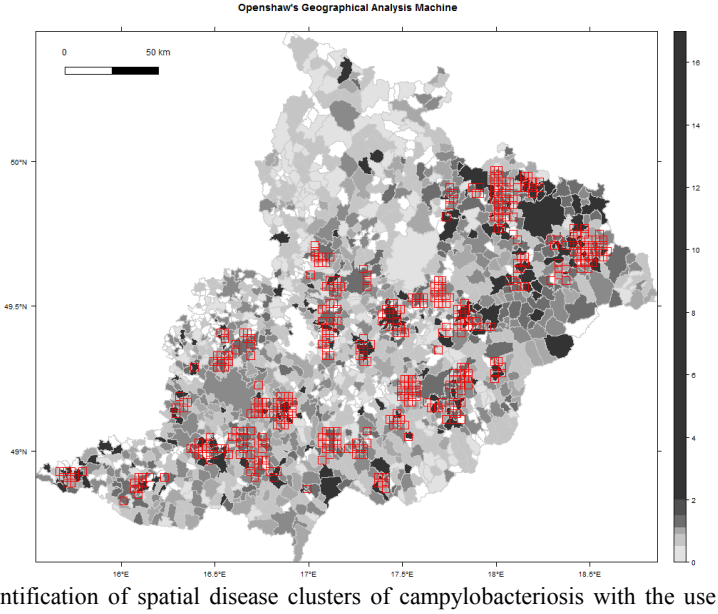

Fig. 3. Identification of spatial disease clusters of campylobacteriosis with the use of Openshaw's Geographical Analysis Machine. Colours and legend depicts standardized incidence ration, while red squares identify locations that are involved in probable disease spatial clusters.

Results of scan statistics used for the identification of spatial disease clusters of campylobacteriosis with the use of the clustering function for Kulldorff and Nagarwalla's statistic are shown on the Fig. 4. The scan statistics is based on the Poisson distribution of disease events, 15 % significance and 5 % fraction of total population. Unlike GAM results, only one significant cluster was identified in the northern part of the study area and it is located in the surrounding of the Ostrava with the core in the village Kateřinice (dark grey area on the Fig. 4).

The last analysis is based on the function that fits a Poisson log-normal random effects models to spatial count data, where the random effects are modelled by the localised conditional autoregressive (CAR) model. The model is based on the list of binary neighbourhood with the queen contiguity conceptualization of space. The observed amount of cases is modelled as the of logarithmical scale of amount of expected number of disease events (intercept) and the ratio between young people (under 15) and elderly people (64+), which is also the basis of the dissimilarity matrix. The analysis detected only two areas (Fig. 5 - left part) that might be the cores of possible

clusters. The first municipality is located in the south-western part of the study area (village Podhradí nad Dyjí). The second theoretical core area is placed in the village Nelepeč-Žernůvka in the west of the study area. That might indicate other necessary customization of the model with the use of other characteristics of the area. Similar analysis based on the distribution of the population was performed due to the comparison. It is depicted in the right part of Fig. 5. Unlike the previous analysis, the result showed significantly more borders between clustering areas and their neighbourhood. On the other hand most of them are densely populated, so the analyst should consider their importance carefully and focus on several individual locations.

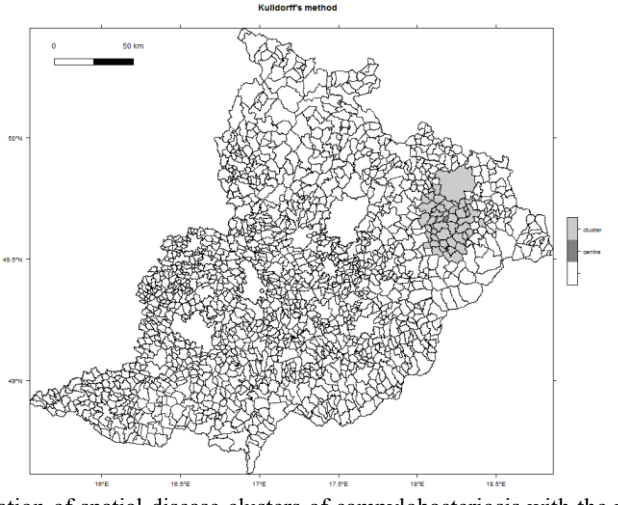

Fig. 4. Identification of spatial disease clusters of campylobacteriosis with the use of the clustering function for Kulldorff and Nagarwalla's statistic. Dark grey areas stand for central (core) area, light grey colour stand for other municipalities in the spatial cluster.

5 Discussion and Conclusion

The contribution aimed to introduce methods of spatial clustering and Bayesian spatial clustering that were based either on the location of disease events in the study area of four Moravian regions or their locations and demographical characteristics of municipalities. One has to realize that all presented methods are dependent on the scale and also on the prior information, which is entering the models mostly in the form of the probability distribution. Therefore, results and their evaluation have to be performed carefully in order to avoid misinterpretation. The aim of the contribution is therefore not only to use methods in real case study but also to show several different results that originally come from the same data.

Firstly Openshaw's GAM detected high number of possible diseases clusters, but due to its disadvantages, results were taken just as informative and an initial step for further analysis. Then, scan statistics based on Kulldorff and Nagarwalla's statistic was used for the identification of spatial disease clusters of campylobacteriosis. The scan statistics discovered one statistically significant cluster on the north of the study area. Lastly, the Poisson log-normal random effects models to spatial count data, where the

random effects are modelled by the localised conditional autoregressive (CAR) model, was used to proceed more detailed and complex analysis. This model incorporated the information about neighbourhood of individual municipalities and also the dissimilarity matrix based on the age structure of the population in the neighbouring villages or cities. The model was able to identify two core areas of possible clusters.

One has to realize that Bayesian techniques usually tend to shift values to the mean risk – global or local by incorporating information between areas. The risks in areas with more information (e.g., urban areas) are usually less smoothed than in areas that exhibit higher sampling variation (typically those with low number of cases), and thus produce more stable estimates of the pattern of underlying disease risk [20]. However, although raw risks can produce "noisy" maps that are difficult to interpret, over-smoothed maps may produce a homogeneous risk surface, masking the true risk distribution [3]. It is important to mention that all analyses presented in this paper are heavily dependent on the scale. We chose the scale of municipal districts but results on other scales could show differences. When someone chooses to broad scale for the analysis, results will probably reveal one (or several) large cluster so the local variance disappears. On the other hand, to local scale may not lead to identification of any clusters. The extension of Bayesian model using other characteristic of the population, spatial unit or disease is possible; however their dynamic properties are mainly shrunk to the sequential procession of time series or time slices.

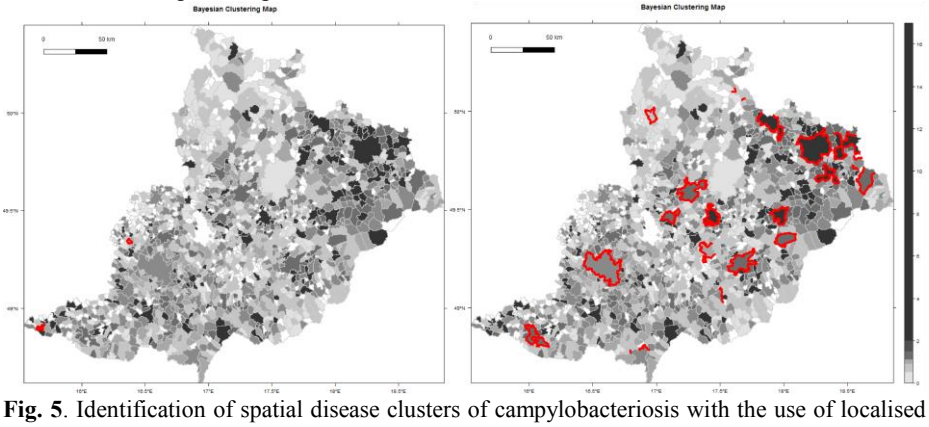

Fig. 5. Identification of spatial disease clusters of campylobacteriosis with the use of localised conditional autoregressive (CAR) model based on dissimilarity metrics with binary neighbourhood relations to spatial Poisson data. Colours and legend depicts standardized incidence ration, while red areas identify locations that are centres of probable disease spatial clusters. The left part describes the relation of likely clusters to the ratio of old people to children, while the right part is based on the population.

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.: GeoDa^{TM} 0.9 User's Guide. (2003).
2. Assuncao, R., Reis, E.: A new proposal to adjust Moran's I for population density. Stat. Med. 2162, November 1998, 2147–2162 (1999).
3. Beale, L. et al.: Methodologic issues and approaches to spatial epidemiology. Environ. Health Perspect. 116, 8, 1105–10 (2008).
4. Clayton, D., Bernardinelli, L.: Bayesian methods for mapping disease risk. In: Elliott, P. et al. (eds.) Geographical and Environmental Epidemiology: Methods for Small Area Studies. Oxford University Press, Oxford (1996).
5. Elliott, P., Wartenberg, D.: Spatial Epidemiology: Current Approaches and Future Challenges. Environ. Health Perspect. 112, 9, 998–1006 (2004).
6. Goodchild, M., Haining, R.: GIS and spatial data analysis: Converging perspectives. Pap. Reg. Sci. 44, 0, 1–26 (2004).
7. Griffith, D., Arbia, G.: Detecting negative spatial autocorrelation in georeferenced random variables. Int. J. Geogr. Inf. Sci. 24, 3, 417–437 (2010).
8. Kulldorff, M.: A spatial scan statistic. Commun. Stat. - Theory Methods. 26, 6, 1481–1496 (1997).
9. Kulldorff, M., Nagarwalla, N.: Spatial disease clusters: Detection and inference. Stat. Med. 14, 8, 799–810 (1995).
10. Last, J., Abramson, J.: A Dictionary of Epidemiology. Oxford University Press, USA (2001).
11. Lawson, A.B.: Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. CRC Press (2009).
12. Lee, D.: CARBayes: An R Package for Bayesian Spatial. J. Stat. Softw. 55, 13, 24 (2013).
13. Lee, D., Mitchell, R.: Boundary detection in disease mapping studies. Biostatistics. 13, 3, 415–26 (2012).
14. Marek, L. et al.: Bayesian mapping of medical data. (2014).
15. 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).
16. Moran, P.: Notes on continuous stochastic phenomena. Biometrika. 37, 1, 17–23 (1950).
17. Openshaw, S. et al.: A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. Int. J. Geogr. Inf. Syst. 1, 4, 335–358 (1987).
18. Openshaw, S.: The Modifiable Areal Unit Problem. , Norwhich (1984).
19. Pfeiffer, D. et al.: Spatial analysis in epidemiology. Oxford University Press (2008).
20. Richardson, S. et al.: Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies. Environ. Health Perspect. 112, 9, 1016–1025 (2004).
21. Waller, L.: Detection of clustering in spatial data. SAGE Handb. Spat. Anal. 34 (2009).
22. Waller, L.A., Gotway, C.A.: Applied Spatial Statistics for Public Health Data. John Wiley & Sons (2004).
23. The Center for food security & public health: Campylobacteriosis, available at: http://www.cfsph.iastate.edu/FastFacts/pdfs/campylobacterosis_F.pdf, (2013).