Hot spot analysis: Improving a local indicator of spatial association for application in traffic safety

Hot Spot Analysis: Improving a Local Indicator of Spatial Association for Application in Traffic Safety

Elke Moons, Tom Brijs and Geert Wets

Transportation Research Institute, Hasselt University,
Science Park 1/15, 3590 Diepenbeek, BELGIUM
elke.moons@uhasselt.be

Abstract. As in most European countries, traffic safety has become top priority in the National Safety Plan in Belgium. The first phase in every safety analysis concerns the identification of the hazardous locations. In this respect, a local indicator of spatial association (Moran's I) is improved and applied to determine hot spots locations on highways in Limburg, a province in Belgium. However, the analysis is complicated by the fact that accident data have a very specific nature: they form a Poisson random process rather than a Gaussian random process and they are prone to sparseness. Therefore, the well-established indicator needs some adaptations and simulations are required to determine the underlying distribution of Moran's I. This paper emphasizes the importance of using a correct distribution and indicates what could go wrong otherwise e.g. at policy level.

Keywords: spatial statistics, local indicator of spatial association (LISA), Moran's I, hot spot accident analysis, traffic safety

1 Introduction

In the last years, accident figures have become a 'hot' topic in the media, for policy makers, for academics and for the broad audience. In comparison to most other European countries, Belgium performs below par concerning traffic safety. The risk of having a fatal accident (per vehicle kilometers driven) is 33% higher in Belgium when compared to the European average [3]. Therefore, it seems more than natural that traffic safety takes a top priority in the National Safety Plan. Moreover, the States General of traffic safety set the amibitious goal to reduce the number of fatal accidents per year by 2015 up to 500 [14].

An important topic of interest in safety analysis is investigating the location of and the reason for the dangerous sites, referred to as *hot spot analysis*. Hot spot analysis in traffic safety can in general be split up into four phases. Firstly, the hazardous locations (hot spots) need to be identified. Next, the locations will be rank ordered. If required, the severity of the accident can be accounted for (e.g. [6], [7], [25] and [31]). Consequently, one tries to find an explanation why some locations are hot spots and others not (i.e. profiling of hot spots). Several techniques can be applied for this purpose. Examples are the analysis of manoeuvre diagrams, conflict observation studies and methods based on information from the traffic accident records [28], [12], [11], [15], [16]. Finally, a selection needs to be made about the sites to treat [26]. This often proves to be a policy decision, and the choice can depend on many factors: e.g. the financial supplies, whether or not it is preferred to look at a group of locations or at every site separately, on a cost-benefit analysis [4], [18]. In this paper, the focus lies on the first of the four phases, i.e. on the identification of hot spots.

Hot spot safety work can be described as the task to enhance traffic safety by adapting the geometrical characteristics and the environmental features of problematic locations in the existing road network. More in specific, this task entails localizing and treating intersections and road segments with an unusual high number of accidents, the so-called hot spots. To reduce the number of accidents, one needs to know where concentrations of accidents occur. As a consequence, the geographical aspect is of major importance to describe and tackle the problem of traffic accidents in order to indicate the most critical hazardous locations in a scientifically sound and practical, workable way. Although, one acknowledges the importance of the geographical aspect, very often statistical - non-spatial - regression models are used to model the number of accidents with respect to some explanatory variables, hereby ignoring the existing geographical relationship between the different locations [13], [10], [24]. Hierarchical Poisson-gamma models are well-known in traffic accident modelling, next to zero-inflated Poisson and negative-binomial models and fairly recently Poisson-lognormal models were used to consider the small accident counts [21], [19], [20], [22], [29].

In this paper, the geographical aspect is directly taken into account by applying a local indicator of spatial association to identify hazardous locations on highways and by the consideration of distances along the road network. Because of the nature of traffic accidents, the normal use of the indicator has serious weaknesses and some adaptations were necessary to serve the purpose. This will be explained in Sect. 2. The application on highways in a province of Belgium is the content of Sect. 3, while Sect. 4 ends with conclusions and some ideas for future research.

2 Methods

Recently, in traffic safety literature, there is a tendency to use spatial data analysis techniques next to statistical (Poisson) regression models [13], [10] to determine locations with a high number of accidents. This enables to account for the spatial character of a location. In this paper, a spatial autocorrelation index is used. The application of spatial autocorrelation indexes allows to search for relationships between locations that are close to another in space within a study area and moreover they are similar or opposite to each other regarding a certain variable under study.

One distinguishes between global and local spatial autocorrelation. The global measure investigates globally if locations that belong to the study area are spatially correlated. Next to the global measure, which gives an idea about the study area as a whole, it might be interesting to limit the analysis to a part of it. Indeed, it might happen that parts of the study area show a spatial autocorrelation, which was not noticed in the global measure. On the other hand, if global spatial autocorrelation is present, the local indexes can be useful to point at the contribution of smaller parts of the area under investigation. These local indexes are considered to be *Local Indicators of Spatial Association (LISA)* [1], if they meet two conditions:

- it needs to measure the extent of spatial autocorrelation around a particular observation, and this for each observation in the data set;
- the sum of the local indexes needs to be proportional to the global measure of spatial association.

The global version of Moran's I was first discussed in [27], though in this paper its local version will be applied. The local version of Moran's I at location i that satisfies the above two requirements can be written down as follows:

$$I_i = \frac{n}{(n-1)S^2}(x_i - \bar{x}) \sum_j w_{ij}(x_j - \bar{x}) \tag{1}$$

$$\text{with} \begin{cases} \text{x}_i \text{ the value of the variable under investigation, } X\text{, at location } i, \\ \bar{x} \text{ the average value of } X, \\ \text{w}_{ij} \text{ a weight that denotes the proximity between location } i \text{ and } j, \\ \text{S}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{, the variance of the observed values and,} \\ \text{n the total number of locations.} \end{cases}$$

A nice property of Moran's I is the fact that it looks relative with respect to an average value. Because of computational issues, it is often impossible to compute the index for the study area as a whole in one time, and one needs to split up the study area in smaller parts (as will be the case in Sect. 3). By plugging in the average of the entire study area as $\bar{x}$ (instead of just the average of the smaller part), all results can easily be combined. So, $\bar{x}$ might serve as a reference value for the study area under investigation.

Anselin [1] derives the mean and variance of I_i under the randomization assumption for a continuous X-variable. The expected value of I_i is, for example [30]:

$$E[I_i] = \frac{-1}{n-1} \sum_{j=1}^{n} w_{ij} \quad .$$

The exact distributional properties of the autocorrelation statistics are elusive, even in the case of a Gaussian random field. The Gaussian approximation tends to work well, but the same cannot necessarily be said for the local statistics [30]. Anselin [1] recommends randomization inference, e.g. by using a permutation approach. However, Besag and Newell [5] and Waller and Gotway [32] note that when the data have heterogeneous means or variances, a common occurrence with count data, such as traffic accidents, the randomization assumption is inappropriate. Instead, they recommend the use of Monte Carlo testing.

Within the field of application, this local measure of spatial association can be regarded as being a traffic safety index, since for each hectometer (the basic spatial unit for analysis on highways in Belgium) of road the local Moran index can be regarded as a measure of association between the hectometer under study and the neighboring hectometers that are similar to the one under study concerning the number of accidents. A negative value of the local autocorrelation index at location i indicates opposite values of the variable at location i compared to its neighboring locations. A positive value, on the contrary, points at similar values at location i and its neighborhood. This means that location i and its weighted neighborhood can both have values above the average value or both can have values below the average. In the application area of traffic safety, however, one is only interested in locations that have

a. a high number of accidents in regard to the total average number of accidents ($x_i - \bar{x} > 0$),
b. and where the neighborhood also shows more accidents than was expected on average ($\sum_j w_{ij}(x_j - \bar{x}) > 0$).

One might argue that it is also important to look at locations with a high number of accidents at location i and a very low number in the surrounding area. In this case, very negative values of Moran's I would occur. However, this gives very contradictory effects as will be illustrated by an example. E.g. suppose that the global average equals one accident ($\bar{x} = 1$). If 5 accidents occurred at location i and none in its surrounding, this would lead to a negative value of Moran's I. However, adding one accident to every surrounding point of location i, hence making the surrounding area more hazardous, would lead to a Moran's I of zero. This is really counterintuitive, so therefore it was opted to look only at points where the location and its surrounding area reinforce each other in a positive way.

An important disadvantage of spatial autocorrelation in general is that this measure is not uniquely defined. There is no optimal specification for the weights and this proves to be one of the most difficult and controversial methodological issues in spatial econometrics [2]. One needs to account for two different aspects, i.e. the *number of neighbors* (level of connection) and the *value of the weights*. Concerning the level of connection, it seems impossible to define an optimal distance between two hectometer poles for which both poles would still show any connection. This optimal distance will vary with the type and the characteristics of the road under investigation, but probably also with the road configuration and the posted speed limit. Additionally, the choice of weights is not uniquely defined. A suggestion was made to assign all locations in the neighborhood of a certain location a weight equal to one and the remaining locations a weight value of zero, though this does not account for the fact that the locations are not uniformly spread. It seems only natural to account for the distance between the locations to determine the local autocorrelation. Often the inverse of the squared distance is used. This entails that the less nearby a location is to the location under study, the less weight is given to that 'distant' location when computing the autocorrelation. Note that at the end of a road, one only accounts for the neighbors that exist.

3 Application: Accidents on Highways

This Section gives an illustration of the use of the local Moran index for accidents on highways in the province of Limburg in Belgium. An important issue for accident analysis is the variability, i.e. the fact that the yearly number of accidents on a road segment varies for each year. This can be explained by the inherent accident risk of a road segment. The randomness in the number of accidents is typical, because of the nature of accidents and it depends on factors that cannot always be predicted. Therefore it is of great importance to study the road network for a sufficient amount of time. The study period must be long enough to ensure representative accident samples. Based on a large number of studies, in general, it is agreed upon that the period of three to five years is sufficient to guarantee the reliability of the results [8]. For the analysis, data on accidents on highways were collected from 2003 to 2005. Analyses were carried out on the province of Limburg in Belgium. Figures 1 and 2 indicate the location of the province of Limburg within Flanders (the upper, Dutch speaking part of Belgium) and the highways in Limburg, respectively.

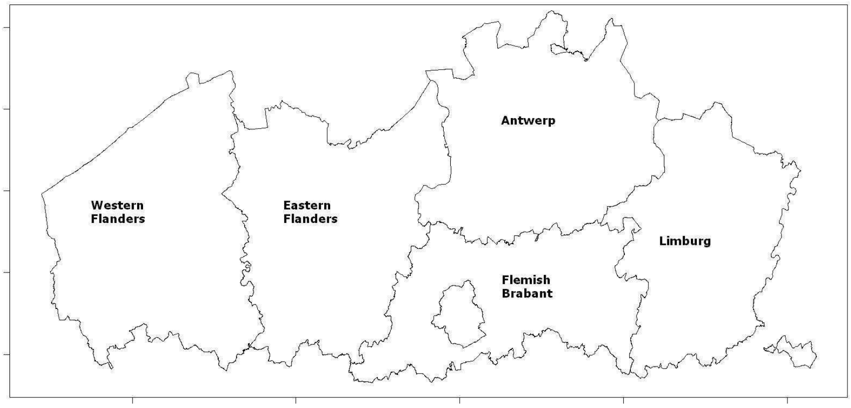

Fig. 1. Limburg within Flanders

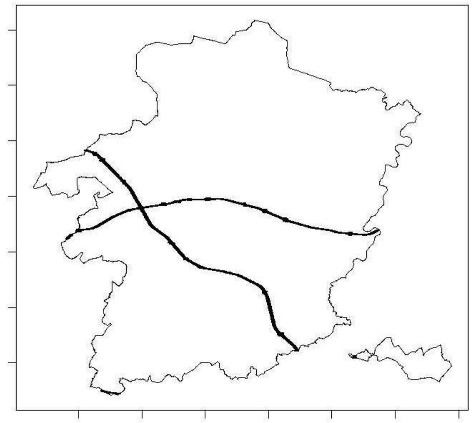

Fig. 2. Highways in Limburg

Because we want to compare the results of Limburg with other provinces in Flanders, the average number of accidents for Flanders was set as a reference value in the analysis. Limburg has 3252 hectometer poles alongside its two highways (E313 and E314). Each accident that occurs on a highway is registered at the closest hectometer pole. The distance matrix that is used for the weights in Moran's I was completely determined based on distances from one point to another along the road network (so not a bird's-eye view). These network distances are used to account for the curvature in the highways and to take care of the junction in a proper way. Note that the inverse of the squared distance was used to determine the weights for the computation of Moran's I. The number of neighbors for this analysis is also determined based on distance. For each hectometer pole, poles at 1km from the left and at 1 km from the right are included as neighbors. So each point, not located near the end of any highway, has approximately 20 neighboring points. Nearby the junction, it happens that hectometer poles from the second highway are within the predefined number of neighbors from a location at the first highway. To account for them in a proper way, distances are calculated based on the network.

470 accidents occurred on Limburg highways from 2003 to 2005, which makes it the safest province in Flanders for this period. Since accidents form a Poisson process instead of a Gaussian process and because of the sparseness (most locations have a zero accident count), the distribution of the local autocorrelation statistic proves to be far from Gaussian. Moreover, count data often suffer from the problem of overdispersion and means and variances tend to be hetereogeneous, so as stated in [5], [32] and [30]. Therefore, a Monte Carlo approach is recommended to derive the distribution of the test statistic.

To this end, the 470 accidents are spread randomly over the 3252 locations and the local Moran index is then calculated for each location. Note that locations are allowed to have more than one accident (sampling with replacement). Otherwise, high concentrations of accidents cannot be determined. This simulation was repeated 500 times to end up with an approximate distribution of the Moran index for this situation (see Fig. 3). It is obvious that a Gaussian approximation would not work well in these circumstances.

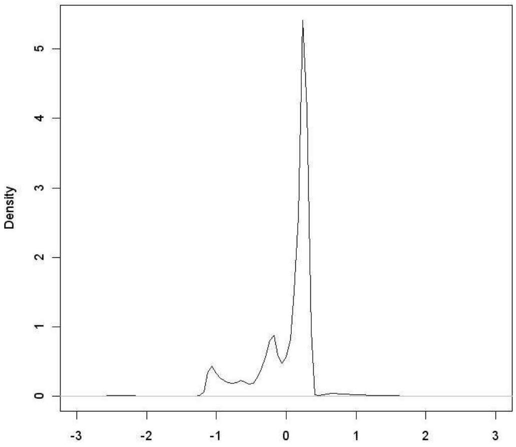

Fig. 3. Simulated density of local Moran index

To determine the hot spots, we decided to look only at locations that show a positive reinforcement with their neighbors in the calculation of the local autocorrelation index. So, these values were filtered out of the 500 × 3252 values (i.e. 53,407 out of the 1,626,000 index values) which results in the following density (see Fig. 4).

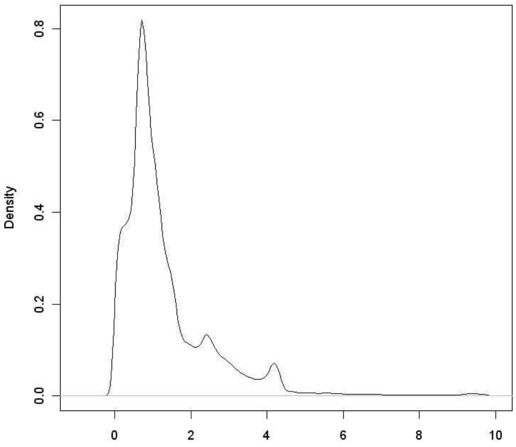

Fig. 4. Simulated density of positive reinforced Moran values

From this distribution, the 95% percentile was calculated, and this value, i.e. $P_{95} = 3.93$, was utilized as the cut-off value to determine an accident hot spot. Therefore, first, the Moran index was calculated for the true number of accidents on each location. Next, the Moran indexes where the location under study and surrounding area positively reinforced each other were selected and each of these index values was compared to the cut-off value. If that local value exceeded the cut-off value (i.e. if $I_i > P_{95}$), then location i was considered to be hazardous. Of the 3252 locations, 15 proved to be hot spots. They are shown in Fig. 5.

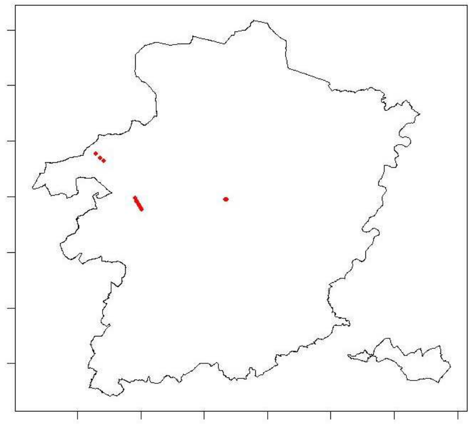

Fig. 5. Hot spots in Limburg

For comparison purposes, the Gaussian approximation was used to investigate the difference in results. Using also the 95% percentile (of the Gaussian distribution!) as cut-off value, now 59 locations proved to be hot spots according to this method. They are shown in Fig. 6. The previous 15 are part of them, however, about 3/4 (!) of the points are falsely identified as belonging to the 5% most extreme Moran index values. From a policy point of view, this might lead to a wrong allocation of funds to ameliorate traffic safety and thus it indisputably shows the importance of using the right distributions.

Recently, some other spatial analysis techniques have been applied to road accidents, e.g. the use of local indicators of network-constrained clusters [34] or the use of network K-functions [33], [23], [17]. A comparison of the results of these techniques with the result presented in this paper provides a promising pathway for future research.

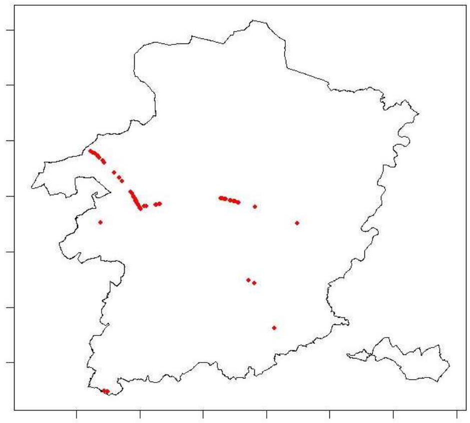

Fig. 6. Hot spots in Limburg when using the Gaussian approximation

4 Conclusion

The aim of this paper was to apply a local indicator of spatial association, more in specific Moran's I, to identify hazardous locations on highways. Because of the nature of traffic accidents, locations with zero counts are very frequent. Moreover, accident data in general stem from a Poisson random process, rather than a Gaussian random process, so the normal use of the indicator seemed very elusive. To account for these characteristics, a simulation procedure was set up to arrive at the distribution of the local indicator, so as to determine the 5% most extreme observations. The study area of application was the highway network of the province of Limburg in Belgium. To construct the distribution of Moran's I, a Monte Carlo simulation experiment was set up where the reported number of accidents was spread randomly over the population of possible locations. This was repeated 500 times and each time the Moran index was calculated for each location. Locations with a local Moran index above the 95% cut-off value of the density are then regarded as hot spots. 15 locations ended up as being hazardous. For comparison purposes, the same analysis was carried out using the Gaussian approximation instead of the simulated distribution. Now 59 locations were defined as hot spots, including the previous 15. This illustrates that blindly using the Gaussian approximation certainly is not an option, and one absolutely needs to take into account the nature of the data under study! For policy makers, this is a very relevant result, since they usually do not have access to an unlimited budget to treat hot spots. To allocate their funds in the best possible way, it is important to know which location are true hot spots.

A generalization of this result is possible, when more provinces are analyzed. Applying this and other (spatial) techniques to identify hot spots on other road types (regional/provincial and local roads) and to compare the different results is certainly an important avenue for future research.

Acknowledgments

The authors would like to thank the Strategic Analysis Department of the Federal Police for providing the data used in this study.

References

1. Anselin, L.: Local indicators of spatial association – LISA. Geog. An. **27(2)** (1995) 93–115.
2. Anselin, L., Florax, R.J.G.M.: New directions in spatial econometrics (1995) Springer-Verlag, Berlin-Heidelberg
3. Australian Transport Safety Bureau: International Road Safety Comparisons: the 2005 report - A comparison of road safety statistics in OECD nations and Australia. ATSB Research and Analysis Report Road Safety **Monograph 19** (2007)
4. Banihashemi M.: EB Analysis in the micro optimization of the improvement benefits of highway segments for models with accident modification factors (AMFs). (2007) El. proc. of the 86^{th} Annual Meeting of the Transportation Research Board, Washington D.C., USA
5. Besag,J., Newell, J.: The detection of clusters in rare diseases. J. Roy. Stat. Soc. A **154** (1991) 327–333
6. Brijs T., Van den Bossche F., Wets G., Karlis D.: A model for identifying and ranking dangerous accident locations: a case study in Flanders. Stat. Neerl. **60(4)** (2006) 457–476
7. Brijs T., Karlis D., Van den Bossche F., Wets G.: A Bayesian model for ranking hazardous road sites. J. Roy. Stat. Soc. A **170** (2007) 1–17
8. Cheng, W., Washington, S.P.: Experimental evaluation of hotspot identification methods. Acc. An. Prev. **37** (2005) 870–881
9. Cressie, N.A.C.: Statistics for spatial data (1993) John Wiley and Sons.
10. Sorensen, M., Elvik, R.: Black spot management and safety analysis of road networks - Best practice guidelines and implementation steps. TOI report **919/2007** (2007) RIPCORD/ISEREST project
11. Flahaut, B. Mouchart, M. San Martin, E., Thomas, I.: The local spatial autocorrelation and the kernel method for identifying black zones - A comparative approach. Acc. An. Prev. **35** (2003) 991–1004
12. Geurts K., Thomas I., Wets G.: Understanding spatial concentrations of road accidents using frequent itemsets. Acc. An. Prev. **37(4)** (2005) 787–799
13. Hauer, E.: Identification of sites with promise. Transp. Res. Rec. **1542** (1996) 54–60
14. Staten-Generaal van de Verkeersveiligheid: Verslag van de Federale Commissie voor de Verkeersveiligheid (2007) 38p. (in Dutch)
15. Jianming M., Kockelman K.M.: Bayesian multivariate Poisson regression for models of injury count, by severity. (2006) El. proc. of the 85^{th} Annual Meeting of the Transportation Research Board, Washington D.C., USA

16. Jianming M., Kochelman K.M., Damien P.: Bayesian multivariate Poisson-Lognormal regression for crash prediction on rural two-lane highways. (2007) El. proc. of the 86^{th} Annual Meeting of the Transportation Research Board, Washington D.C., USA
17. Jones, A.P., Langford, I.H., Bentham, G.: The application of K-function analysis of the geographical distribution of road traffic accident outcomes in Norfolk, England. Soc. Sci. Med. **42(6)** (1996) 879–885
18. Kar K., Datta T.K.: Development of a safety resource allocation model in Michigan. Transp. Res. Rec. **1865** (2004) 64–71
19. Li, L., Zhang, Y.: A GIS-based Bayesian approach for identifying hazardous roadway segments for traffic crashes. (2007) El. proc. of the 86^th Annual meeting of the Transportation Research Board, Washington, D.C., USA
20. Li, L., Zhu, L., Sui, D.Z.: A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes. J. Transport Geogr. **15** (2007) 274–285
21. Lord, D.: Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Acc. An. Prev. **38** (2006) 751–766
22. Lord, D., Miranda-Moreno L.F.: Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modelling motor vehicle crashes: a Bayesian perspective. (2007) El. proc. of the 86^th Annual Meeting of the Transportation Research Board, Washington, D.C., USA
23. Lu, Y., Chen, X.: On the false alarm of planar K-function when analyzing urban crime distribution along streets. Soc. Sci. Res. **36** (2007) 611–632
24. McGuigan, D.R.D. The use of relationships between road accidents and traffic flow in 'black spot' identification. Traffic Eng. Contr. **22** (1981) 448–453
25. Miaou S.-P., Song J.J.: Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Acc. An. Prev. **37** (2005) 699–720
26. Miranda-Moreno L.F., Labbe A., Fu L.: Bayesian multiple testing procedures for hotspot identification. Acc. An. Prev. **39(6)** (2007) 1192–1201
27. Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika **37** (1950) 17–23
28. Pande A., Abdel-Aty M.: Market basket analysis: novel way to find patterns in crash data from large jurisdictions. (2007) El. proc. of the 86^{th} Annual Meeting of the Transportation Research Board, Washington D.C., USA
29. Park, E.S., Lord, D.: Multivariate Poisson-Lognormal models for jointly modeling crash frequency by severity. (2007) El. proc. of the 86^th Annual Meeting of the Transportation Research Board, Washington, D.C., USA
30. Schabenberger, O., Gatway, C.A.: Statistical methods for spatial data analysis (2005) Chapman & Hall/CRC Press
31. Vistisen D.: Models and methods for hotspot safety work. Phd Dissertation (2002) Technical University of Denmark
32. Waller, L.A., Gotway, C.A. Applied spatial statistics for public health data (2004) John Wiley and Sons
33. Yamada, I., Thill, J.-C.: Comparison of planar and network K-functions in traffic accident analysis. J. Transport Geogr. **12** (2004) 149–158
34. Yamada, I., Thill, J.-C.: Local indicators of network-constrained clusters in spatial point patterns. Geogr. An. **39** (2007) 268–292