Improving Moran's Index to Identify Hot Spots in Traffic Safety

Improving Moran's I to Identify Hot Spots in Traffic Safety

Elke Moons*, Tom Brijs, Geert Wets

Transportation Research Institute (IMOB)

Hasselt University

Science park 1/15

B-3590 Diepenbeek

BELGIUM

Tel. +32-11-26.91.26

Fax: +32-11-26.91.99

Email: elke.moons@uhasselt.be

* corresponding author

Abstract

This chapter aims at identifying accident hot spots by means of a local indicator of spatial association (LISA), more in particular Moran's I. A straightforward use of this LISA is impossible, since it is not tailor-made for applications in traffic safety. First of all, road accidents occur on a network, so Moran's I needs to be adapted to account for this. Moreover, its regular distributional properties are not valid under the circumstances of Poisson distributed count data, as is the case for accidents. Therefore, a Monte Carlo simulation procedure is set up to determine the correct distribution of the indicator under study, though this can be generalized to any kind of LISA. Moran's I will be adapted in such a way, that it can overcome all the previously stated problems. Results are presented on highways in a province in Flanders and in a city environment. They indicate that an incorrect use of the underlying distribution would lead to false results. Next to this, the impact of the weight function is thoroughly investigated and compared in both settings. The obtained results may have a large impact for policy makers, as money could be allocated in a completely wrong way when an unadjusted LISA is used.

1 INTRODUCTION

Over the past decades, traffic safety has become a topic of increasing interest in the media, as well as for policy makers. The States General of Traffic Safety [1] have set the ambitious goal to reduce the number of individuals killed in traffic per year from 1,000 in 2006 to 500 by 2015. As opposed to most of our neighboring countries, Belgium's score concerning traffic safety is still below par. The number of people that had a fatal accident per 1 billion vehicle kilometers equals 11.1 in Belgium in 2006 [2]. This figure is about 31% higher than the number in France, 44% more than in The Netherlands and even 50% higher when compared to Germany. Putting these figures in an international context only confirms Belgium's poor performance (The United States have a figure of 9 persons killed per 1 billion vehicle-kilometer, Australia has a value of 7.9 and Japan of 10.3). Therefore, it only seems logical that traffic safety has become top priority in the National Safety Plan.

A key issue in traffic safety analysis is determining the reason for a site to be hazardous, also referred to as hot spot analysis (HSA). In general, HSA can be split up into four phases. The first step is to identify the dangerous locations. Next, a ranking of these locations needs to be established. The severity of the accident, determined by the severity of the injuries, can be taken into account here [3, 4, 5, 6]. Consequently, one tries to come up with an explanation why some sites are hot spots and others are not (i.e. profiling of hot spots). This can be verified through an analysis of maneuver diagrams, information from traffic accident records, characteristics of the environment, of the infrastructure, etc. [7, 8]. And, finally, one needs to select the hot spots to be treated [9]. Very often, this turns out to be a policy decision and the choice may be based on different aspects: e.g. based on limited financial supplies, or on a cost-benefit analysis [10, 11]. Only the first phase, identification, will be discussed in this chapter, although the technique could be applied for the purpose of ranking as well.

There exists no univocal definition of a hot spot [12]. Sometimes the number of accidents per vehicle-kilometer driven (VKD) or per number of vehicles is used to identify hot spots, other researchers use an absolute figure (accidents per km/year or per year), and some use a combination of both. Since the definition of a hot spot is already very broad in itself, there also exists a wide range of methods and techniques in the domain of traffic safety to identify hazardous locations on a road network, ranging from simple models that are based on the observed number of accidents to more advanced statistical models that are based on the expected number of accidents. Hot spot safety research encapsulates localizing and treating crossroads and road segments with an unexpected high number of accidents. In order to reduce this number of collisions, it is important to know where concentrations of accidents occur. Therefore, the geographical aspect is highly important to determine and to handle the most unsafe traffic sites in a scientifically sound and practical way. Although one acknowledges the importance of this geographical

aspect, very often statistical – non-spatial – regression models are used to model the number of accidents.

Analyzing hot spots always occurs within in a certain time frame and a large number of locations will show no accidents for that period of time. This is recognized in the literature as sparseness. This abundance of zeroes causes estimation problems in most prediction models. Negative binomial models have been developed to solve this problem and in the recent past this was often countered by using Zero-Inflated Poisson (ZIP) models [13]. It is assumed that a location can find itself in two conditions: either the location is inherently safe (state of zero accidents), or there is a chance that an accident occurs at that location (i.e. the location has a strictly positive mean number of accidents, but the probability of having zero accidents at that location is larger than zero). Modeling accident data through this type of models often yields better results than using an ordinary Poisson regression model. Though, recently this was criticized in the literature [14, 15], because there is no theoretical underpinning to believe that there exists a location that is inherently safe. Namely, an accident is not necessarily caused by infrastructural characteristics, the state of the driver (inattention, drunk driving, etc.) often plays a very big role. Because of this, it is unrealistic to believe that there exists even just one inherently safe location. Very often will the abundance of zeroes be caused by a low exposure (low traffic volumes) and/or by an ill-considered selection of accidents in time and/or space. This can be solved on the one hand by enlarging the time frame or the geographical window or by using a better set of explanatory variables and/or by taking non-observed heterogeneity effects into account to explain the model or by applying methods for small area estimation (e.g. Poisson-lognormal models).

Next to applying a frequentist's approach, traffic safety researchers are inclined to use Bayesian models, since they can make use of prior information in an efficient way. An example that is widely used in traffic safety literature is the Poisson-gamma model [3, 4, 12, 16, 17, 18, 19]. Researchers tend to prefer it to the Poisson regression model, because this model can handle the problem of overdispersion [20]. The Poisson distribution, underlying the regression model, assumes that the mean and variance are equal to each other and since the mean number of accidents usually is very low, accident data often show a larger variance. Very recently [21,22], research was conducted on the effect of low means and small sample sizes in traffic safety and this has led to the conclusion that the Poisson-lognormal model often achieves better results than the Poisson-gamma model.

The advantage of using regression techniques is that one has a 'normal' number of accidents for a certain location at one's disposal and as a consequence one can determine the effect of treating that specific location (i.e. safety potential, [23]. This is expressed in terms of the difference between the expected number of accidents according to the (Bayesian) model and the number of accidents that is judged to be 'normal' for a similar location, i.e. the *potential of accident reduction (PAR)*. This leads us to what is judged to be the model-based definition of a hot spot [24]: a hot spot is a location with an observed number of accidents that is

higher than expected in comparison with similar locations as a consequence of local risk factors.

Most of the techniques discussed above ignore the existing geographical relationship between the different locations. However, it seems only logical that the structure of the underlying road network can play an important role in determining hazardous locations. For example crossroads, on and off ramps on a highway, the existence of one-way streets, it all may have direct implications on the number of accidents on a location nearby. Next to that, there is a recent trend to examine road segments instead of dangerous locations, because of the obvious spatial interaction between accident locations that are close to one another. Spatial techniques allow to account for this. These spatial methods usually exist in one dimension and in two dimensions, but often they are not suited to be used alongside a network. This chapter displays the use of Moran's I to identify hot spots on highways and on regional roads, hereby taking into account the structure of the road network. However, as indicated above, due to the nature of road accidents (sparseness), a straightforward use of the indicator has serious flaws and adaptations are required to apply it in this context. Section 2 gives a background on spatial autocorrelation in general and it explains the use of Moran's I. The second part of this section denotes the required adaptations and the changing distributional properties when Moran's I is applied to a traffic safety context. The impact of using different weight functions is also discussed. Section three gives a description of the data, together with the results on highways in Flanders on the one hand and on regional roads in a city context on the other hand. Conclusions and some ideas for future research are given in Section 4.

2 THE METHOD: MORAN'S I

2.1 Background on Spatial Association

Recently, there is a tendency to use spatial data analysis techniques next to statistical (Bayesian) regression models in HSA [25]. This enables to account for the spatial character of a location. In this chapter, a spatial autocorrelation index is used. It aims at evaluating the level of spatial (inter-)dependence between the values x_i of a variable X under investigation, among spatially located data [26]. If the idea of temporal autocorrelation is extended, then a simple representation of spatial dependence can be formulated as follows:

$$x_i = \rho \sum_j w_{ij} x_j + u_i \,,$$

where ρ measures the spatial autocorrelation between the x_i 's, w_{ij} are the weights that represent the proximity between location i and j and u_i are independent and identically distributed error terms with mean zero and variance σ^2. However, in contrast to temporal autocorrelation, the spatial neighborhood is multidirectional, making it more complex and leading to specific indices for spatial autocorrelation. Specifically, spatial correlation analysis determines the extent to which the value of the variable X at a certain location i is related to the values of that variable at contiguous locations. This assessment involves analyzing the degree to which the value of a variable for each location co-varies with values of that variable at nearby locations. When the level of co-variation is higher than expected, neighboring locations have similar values (both high or both low) and autocorrelation is positive. Opposite, when the level of co-variation is lower than expected, high values of the variable are contiguous to low values and the autocorrelation is negative. The lack of significant positive or negative co-variation suggests absence of spatial autocorrelation [25].

Global measures of spatial autocorrelation have been applied for several decades and mainly stem from the work of Moran [27][see e.g. 28, 29]. Moran's I is most often used and its usefulness for transport fluxes and traffic accident analysis has been thoroughly discussed in the literature [30, 31]. Next to the global measure that gives an idea about the study area as a whole, it may also be interesting to limit the analysis to a smaller part of it. It might happen that smaller parts of the study area show spatial autocorrelation, but that it has not been picked up by the global measure. Though, also when global autocorrelation is present, the local indices can be useful to point at the contribution of smaller parts of the investigated area. The use of these local indices is more recent [25, 32, 33]. Each location is now characterized by one value of the index that denotes the individual contribution of the location in the global autocorrelation measure.

These local indices are considered to be *Local Indicators of Spatial Association* (LISA's) 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 indices needs to be proportional to the global measure of spatial association.

2.2. General Use of Moran's I

The global version of Moran's I was first discussed in Moran [27], however, in this chapter its local version will be applied. The LISA version of Moran's I that satisfies the two requirements as stated in 2.1 can be written down as follows:

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

with

- x_i representing the value of interest of variable X for point i,
- $\bar{x}$ the average value of X,
- w_{ij} representing the proximity of point i's and point j's locations, with $w_{ii} = 0$ for all points,
- n representing the total number of points, and
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, the variance of the observed values.

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 it needs to be split into smaller parts. 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 (see also 2.3).

Anselin [33] 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 [34]:

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

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 [34]. Anselin [33] recommends randomization inference, e.g. by using a permutation approach. However, Besag and Newell [35] and Waller and Gotway [36] note that when the data have heterogeneous means or variances, a common occurrence with count data such as accidents, the randomization assumption is inappropriate. Instead, they recommend the use of Monte Carlo testing.

2.3 Adaptations

It can be observed that some kind of proximity measure w_{ij} is used to denote the distance between location i and location j in the calculation of Moran's I. In general, geo-referenced x and y coordinates are attached to each location and distances are determined by means of a bird's-eye view. However, accidents take place on a road network, and it may happen that locations are very close to each other in space, though, via the network, they cannot be reached easily (e.g. because one of them is located in a one-way street). A logical step is then to consider the distance traveled alongside the road network. Every location can be pinpointed at the road network map and distances can be determined via the network. This

also takes care of junctions and on and off ramps in a proper way and encapsulates the whole network structure in its measure. This is the first extension that is used in this chapter in comparison to the 'normal' use of LISA's.

Four other adaptations to previous uses of local Moran's I in traffic safety have been proposed here. First of all, it is important to use the index in a correct way. One needs to account for zero observations as well (instead of only taking into account locations with at least one accident, see e.g. [25]). Otherwise the average value would clearly be overestimated. Moreover, all locations with accidents would be judged to be of a too high importance.

Second, as already indicated in the previous paragraph, any reference value can be used for $\bar{x}$. In normal use of the index, this is just the average of the area under study, however, if - e.g. for computational purposes - the area needs to be split up, the average of the total area can still be used here, so that comparisons between small parts are straightforward. If one wants to compare different countries to each other, a global average can be computed and in this way all countries can be compared to that global average. From a traffic safety point of view, it might be interesting to compare e.g. to the average for that type of road, the average of a region or a country.

Third, this local measure of spatial association can be regarded as being a traffic safety index, since for each basic spatial unit (BSU) of road the local Moran index can be regarded as a measure of association between the BSU under study and the neighboring BSU's 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:

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

It might be argued 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 (i.e. a spike). In this case, very negative values of Moran's I would occur. However, although conceptually appealing, this gives very contradictory effects as illustrated by the following example. Suppose that the global average over a certain area equals one accident ($\bar{x}=1$). Then, if 7 accidents occurred at location i and none in its surrounding, this would lead to a negative value of Moran's I and possibly a significant negative autocorrelation. 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, indicating no significant autocorrelation. This would mean that a more dangerous location has a less significant Moran's I when compared to a more 'safe' location. This is really counterintuitive, so therefore it was opted to look only at points where a high number of accidents is contiguous with high values in the neighborhood (the location and its surrounding area reinforce each other in a positive way).

Finally, since the distributional properties of Moran's I are intangible, as, suggested by [35, 36] a Monte Carlo approach was applied to arrive at cut-off values for the local Moran's I above which the location can be considered to be a hot spot. To this end, the total number of accidents for the study area will be spread randomly over the total available locations. Note that locations are allowed to have more than one accident, otherwise, high concentrations of accidents cannot be determined. For each location, the local Moran's I is then calculated. This simulation will be repeated 500 times to end up with an approximate distribution of the local Moran index for the particular situation at hand. Next, to determine the hot spots, it was decided upon to filter out the locations with a high number of accidents contiguous with high neighboring values. For this subsample of locations, the 95% percentile (P_{95}) of the distribution of the remaining Moran values is determined. This value will be utilized as the cut-off value to determine an accident hot spot in the study area. If the local Moran's I value of a location of the true data also has similar high values between the location under study and its contiguous locations ànd it exceeds this 95% percentile (i.e. if $I_i > P_{95}$), then this location is considered to be hazardous, and hence a hot spot location.

A real world example for 506 accidents at 3,252 locations is shown in Figure 1. It is obvious that a Gaussian approximation would not work well in these circumstances. The black curve indicates the simulated density for Moran's I, while the red curve shows the Gaussian approximation with the mean and variance as they are expected to be under randomization.

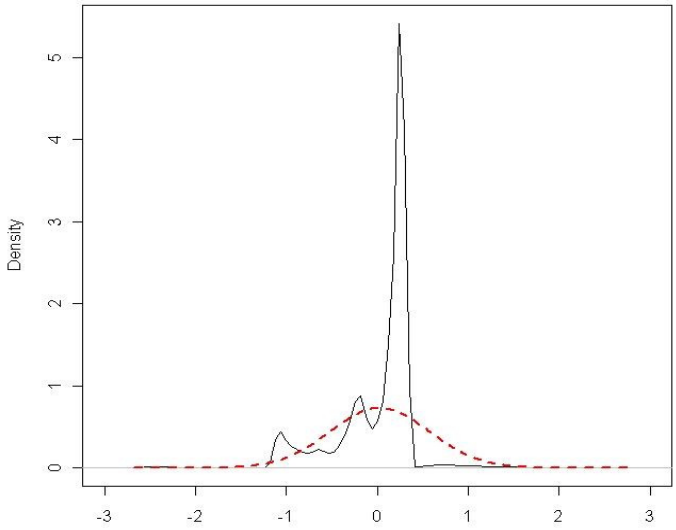

Fig. 1 Simulated density of local Moran's I.

2.4 Impact of the Weights

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 [37]. One needs to consider 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 BSU's for which both BSU's 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, the posted speed limit, etc.

Additionally, the choice of weights is not uniquely defined. Getis and Ord suggested [32] 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 it receives. Note that at the end of a road, one only accounts for the neighbors that

exist. In general, the weights are row-standardized, meaning that the sum of the weights at each location sums up to 1. Figure 2 shows the impact of using different weight functions. Note that the functions are truncated at zero, or else three of the four functions would go to infinity. The black solid line indicates an Epanechnikov-like kernel (E-like) which is often used in kernel density estimation. This is shown to point at the contrast when compared to powers of the distance between locations. The green dashed line equals 1 over the squared distance. One may observe that after 150m the weights are almost equal to zero. The blue dotted line is the inverse of the distance, this allows some weight to be given at locations up to about 750m from the site under investigation. The red line shows 1 over the square root of the distance. This clearly gives some weight at all contiguous locations. For the choice of weight function, just as for the choice for the number of neighbors, there does not exist one optimal choice. It preferably changes per setting, depending on the road configuration of the area under study and it interacts with the number of neighbors. Perhaps different simulation settings together with some expert knowledge can help to provide some more insight in this matter. Although not the main focus of this chapter, some results on different weight functions are shown in Section 3.

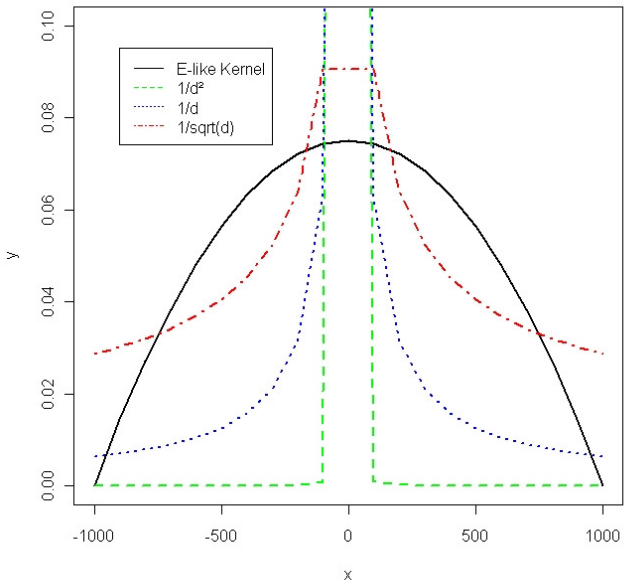

Fig. 2 Impact of different weight functions.

3 ANALYSES AND RESULTS

This Section illustrates the use of Moran's I for two different configurations. A first application comprises accidents on highways in Limburg, a province in Belgium. A second data set consists of accidents on regional roads in the city of Hasselt (capital of the province of Limburg) and its surroundings. Both data sets are provided by the Belgian Federal Police.

Variability, i.e. the fact that the yearly number of accidents on a road segment varies from year to year, is an important issue for accident analysis. 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 because of unpredictable factors, such as the weather. Therefore it is of great importance that the study period is long enough to ensure representative accident samples. Based on a large number of studies, it is generally agreed upon that the period of three to five years is sufficient to guarantee the reliability of the results [18]. For both analyses, data on accidents were collected from 2004 to 2006.

3.1 Data

The first analysis is carried out on the province of Limburg in Belgium. Figure 3 indicates the location of the province of Limburg within Flanders (the upper, Dutch speaking part of Belgium).

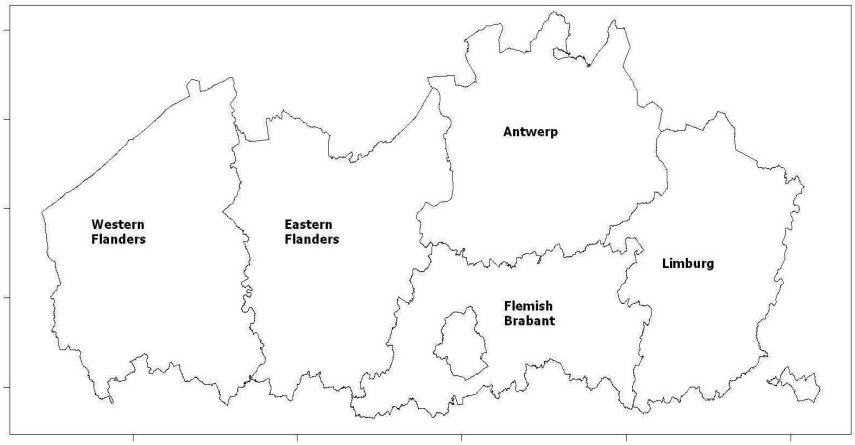

Fig. 3 Limburg within Flanders.

The second analysis is carried out on regional roads. Figure 4 indicates the road network of the city of Hasselt and it's surroundings, together with the BSU's

where accidents occurred. Note that many accidents occurred on the inner and the outer ring way of the city and at the arterial roads towards the city. In the upper left corner, one can observe the clover leaf junction of the two highways in Limburg, the E314 and the E313. This is expected to be a hazardous location, though one needs to take care in which setting. It may be true that this proves to be dangerous when analyzing highways separately, while on regional roads (they actually comprise of provincial roads, regional roads and highways), it may prove not to be a hot spot after all.

For both settings, the basic spatial unit is defined to be about 100m. Accidents occurring at highways are assigned to the closest hectometer pole, so they are regarded as BSU, both for highways as for regional roads. The initial weights that are used are the inverse of the squared distance, where the distance was determined from one BSU to the next one on the network. The number of neighbors is also distance based. For each BSU, BSU's within a 1km range of the BSU under investigation are included as neighbors. So each point, not located near the end of any highway, has approximately 20 neighboring points (more neighbors are possible for the city environment configuration). Nearby the junction of both highways, it may happen that BSU's from the second highway are within the predefined number of neighbors for a location at the first highway. To account for them in a proper way, distances need to be network-based. In the city environment, this becomes even more important since there are much more small roads within the neighborhood of each other.

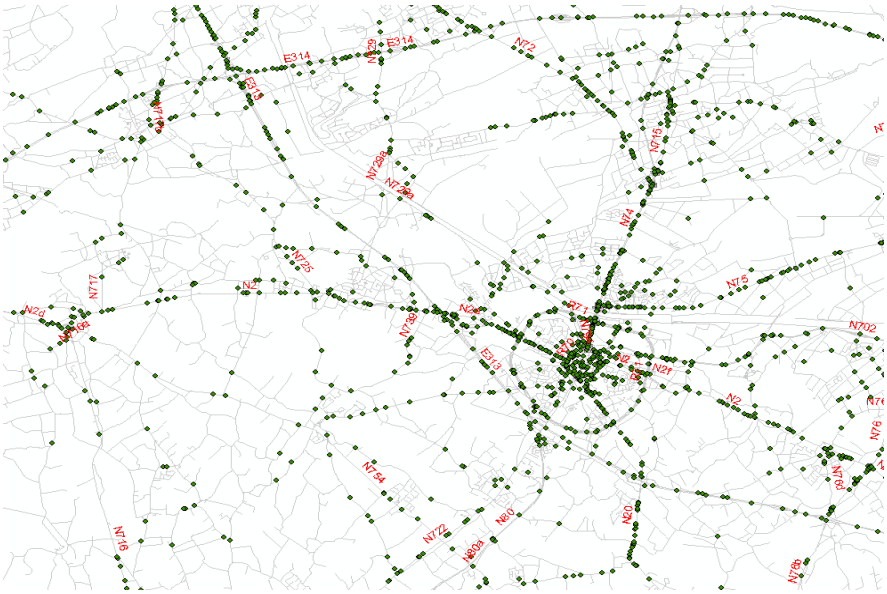

Fig. 4 Road network around Hasselt.

Because the idea is to compare the results of Limburg with other provinces in Flanders, the number of accidents for Flanders was set as a reference value ($\bar{x}$) in both analyses.

Limburg has 3,252 hectometer poles alongside its two highways (E313 and E314) and 506 accidents occurred on these highways between 2004 and 2006. In the second configuration, 1,678 collisions took place on one of the 3,856 possible hectometer sites.

As stated above, 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 statistics proves to be far from Gaussian. Moreover, count data often suffer from the problem of overdispersion and means and variances tend to be heterogeneous, so as stated in [34, 35, 36]. Therefore, the Monte Carlo approach is applied to derive the distribution of the local autocorrelation statistic.

3.2 Analyses and Results

For the setting on highways, the 506 accidents are spread randomly over the 3,252 hectometer poles to determine the distribution of the local version of Moran's I. This density is illustrated in Figure 1. Since we decided to look only at the locations that show a positive reinforcement with their contiguous locations in the calculation of the local autocorrelation index, these values need to be filtered out of the 500 x 3,252 values. From these remaining values, the 95% percentile was calculated and this value, i.e. $P_{95} = 4.32$ was utilized as cut-off value to determine which location is a hot spot and which not. Only 5 of the 3,252 locations appeared to be hot spots on highways in Limburg. For reasons of comparison, the standard used Gaussian approximation was also applied to investigate the difference in results. Using also the 95% percentile (of the Gaussian distribution!) as cut-off value, now 46 locations proved to be hot spots according to this method. The previous 5 are part of them, however, about 87% 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 the funds to ameliorate traffic safety and thus it indisputably shows the importance of using the right distributions.

For the more urban configuration, the 95% percentile proved to be much lower, P_{95} now yields 2.14. 48 sites are determined as hot spots by the local Moran index, while again more than twice as much (114) locations were pinpointed as hazardous if the Gaussian approximation would have been used.

Figure 5 shows the resulting hot spots for both configurations. The left figure shows the province of Limburg and its two highways, while the right figure shows Hasselt and its surrounding area. The hot spots are indicated in red, while the underlying road network is drawn in black.

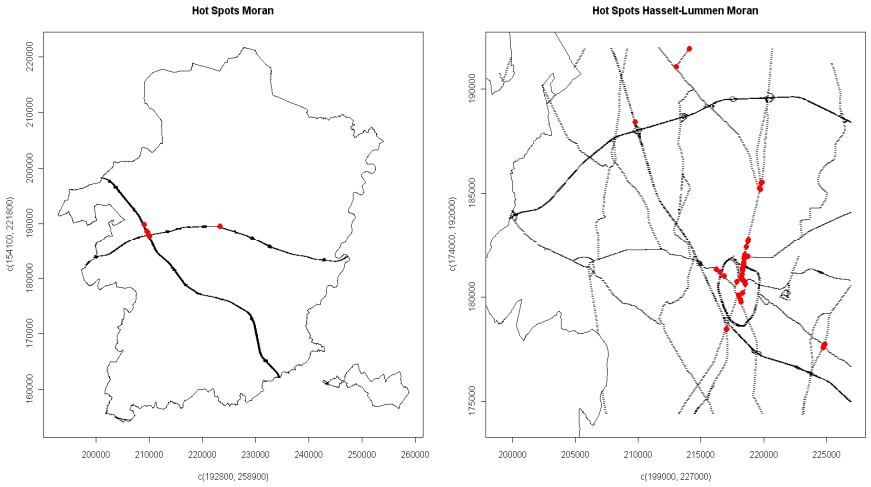

Fig. 5 Hotspots for the both configurations.

Next, four different weight functions are compared to each other. The determination of neighbors stays the same. The previously discussed results are denoted as setting 1. Setting 2 indicates the version where the inverse of the distance is used as weight function, in setting 3 the inverse of the square root of the distances is applied and in setting 4 the Epanechnikov-like function. Table 1 and 2 will display the number of locations and their corresponding accident figure over three years and how many of these locations are determined to be hot spots (HS) for each setting. Table 1 shows the results on highways, whereas Table 2 gives the results in the city environment.

Table 1 Results on highways for different weight functions.

Loc. with … accid.	Nr. of locations	HS in setting 1	HS in setting 2	HS in setting 3	HS in setting 4
0	2,871	0	0	0	0
1	318	0	4	8	6
2	50	3	8	10	10
3	5	0	0	0	0
4	2	1	1	1	1
5	1	0	0	0	0
7	2	1	1	1	1
8	1	0	0	0	0
14	1	0	1	1	1
24	1	0	0	0	0
Total	**---**	**5**	**15**	**21**	**19**

Most locations that proved to be hot spots for the inverse quadratic weight function remain hot spots for all other 3 settings. Furthermore, those locations that are hazardous in setting 2 are almost always HS in setting 3. The hot spots that are retrieved in setting 4 appear to be a mixture of those of setting 2 and 3. It has to be noted that the Gaussian approximation leads to at least 3 times as much 'so-called' hot spots in each of the applied settings. Often this difference is even much larger. This once again emphasizes the necessity to apply the Monte Carlo approach to end up with proper results. The largest difference in the number of hot spots occurs between setting 1 and 2. Although there is a remarkable difference in shape of the weight function (concave versus convex) between settings 3 and 4, the resulting hot spot locations do not differ a lot. This is probably due to the fact the differences occur predominantly at locations further away from the investigated BSU. It also shows that for the accidents on highways not all locations with a high number of accidents turn out to be hot spots, whereas this does happen to be the case for the city environment. It is obvious that the denser the road network is in the study area, the more these dangerous locations show up in the analysis. Only one highway location turns out to be a hot spot location for the analysis in the city environment, and this is a location at the junction of both highways in Limburg. This clearly indicates that the context plays an important role. When considering an urban environment of this type (city with ring ways and arteries), one might argue that it suffices to consider only accidents at provincial and regional roads to determine the most dangerous locations; However, one needs to be careful in generalizing this result, since it is only based on one particular example.

Table 2 Results in city environment for different weight functions.

Loc. with … accid.	Nr. of locations	HS in setting 1	HS in setting 2	HS in setting 3	HS in setting 4
0	3,064	0	0	0	0
1	494	2	1	0	0
2	163	4	6	11	9
3	62	10	11	12	14
4	23	5	6	6	6
5	15	8	8	7	8
6	9	3	4	4	4
7	10	3	4	4	4
8	1	0	0	0	0
9	2	1	1	1	1
12	1	1	1	0	0
14	3	2	3	3	3
17	1	1	1	1	1
20	1	1	1	1	1
21	2	2	2	2	2
24	2	2	2	2	2

30	1	1	1	1	1
67	1	1	1	1	1
78	1	1	1	1	1
Total	**---**	**48**	**54**	**57**	**58**

4 CONCLUSIONS AND DISCUSSION

The aim of this chapter was to apply a local indicator of spatial association, more in particular Moran's I, to identify hazardous locations on highways and on regional roads. First of all, it needs to be acknowledged that accidents occur on a network, and this should be accounted for by using the correct network-based distances between locations under study.

Moreover, accident data in general stem from a Poisson random process, rather than a Gaussian random process and locations with zero counts are very frequent, so the normal use of the indicators seemed very elusive. To account for these characteristics, a simulation procedure was set up to arrive at the distribution of Moran's I, so as to determine the 5% most extreme observations. Two different settings were regarded, the highway network of the province of Limburg in Belgium and the network around the city of Hasselt, the capital of Limburg. To construct the distribution of the improved 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 to arrive at the distribution of the local indicator. Sites that showed a local index above the 95% cut-off value of the density are then regarded as hot spots.

For comparison purposes, the same analysis was carried out using the Gaussian approximation for Moran's I instead of the simulated distribution. Now at least twice as much locations were defined as hot spots, including the ones obtained by a correct use. Blindly using the Gaussian approximation is certainly not an option, and one absolutely needs to take into account the nature of the data under study. This is a very relevant result for policy makers, 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 locations are true hot spots. The impact of different weight functions in both settings has also been investigated. This illustrates that the context and the density of the road network is very important when choosing a good weight function. Further research combined with expert knowledge seems required to come up with some rules of thumb to be used for analyses in the future.

A next step to be taken is to investigate how these hot spots can be combined into hot zones. A possible way forward has been suggested by Loo [38].

An important avenue for future research is to apply and compare the results of this and other (spatial) techniques (such as network-based K-function) to identify hot spots on other road types (e.g. local roads).

Acknowledgments

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

REFERENCES

[1] Staten-Generaal van de Verkeersveiligheid: Verslag van de Federale Commissie voor de Verkeersveiligheid (2007) (in Dutch)

[2] International Traffic Safety Data and Analysis Group (IRTAD): Selected risk values for the year 2006. http://cemt.org/IRTAD/IRTADPublic/we2.html. Accessed 20 August 2008

[3] 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)**, 457-476 (2006)

[4] Brijs T., Karlis D., Van de Bossche F., Wets, G.: A Bayesian model for ranking hazardous road sites. J. Roy. Stat. Soc. A **170**, 1-17 (2007)

[5] 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(4)**, 699-720 (2005)

[6] Vistisen, D.: Models and methods for hotspot safety work. Phd Dissertation, Technical University of Denmark (2002)

[7] Pande A., Abdel-Aty, M.: Market basket analysis: Novel way to find patterns in accident data from large jurisdictions. El. Proc. of the 86th Annual Meeting of the Transportation Research Board, Washington, D.C. (2007)

[8] Geurts K., Thomas I., Wets, G.: Understanding spatial concentrations of road accidents using frequent itemsets. Acc. An. Prev. **37(4)**, 787-799 (2005)

[9] Miranda-Moreno L.F., Labbe A., Fu, L.: Bayesian multiple testing procedures for hotspot identification. Acc. An. Prev. **39(6)**, 1192-1201 (2007)

[10] Banihashemi M.: EB Analysis in the micro optimization of the improvement benefits of highway segments for models with accident modification factors (AMFs). El. Proc. of the 86th Annual Meeting of the Transportation Research Board, Washington, D.C. (2007)

[11] Kar K., Datta, T.K.: Development of a safety resource allocation model in Michigan. Transp. Res. Rec. **1853**, 64-71 (2004)

[12] Hauer, E.: Identification of sites with promise. Transp. Res. Rec. **1542**, 54-60 (1996)

[13] McCulloch, C.E., Searle, S.R.: Generalized, Linear and Mixed Models. Wiley: New York (1989)

[14] Lord, D., Washington, S.P., Ivan, J.N.: Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle accidents: Balancing statistical fit and theory. Acc. An. Prev. **37(1)**, 35-46 (2005)

[15] Lord, D., Washington, S.P., Ivan, J.N.: Further notes on the application of zero-inflated models in highway safety. Acc. An. Prev. **39(1)**, 53-57 (2007)

[16] Hauer, E.: Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd, Oxford, (1997)

[17] Hauer, E., Douglas, W.H., Council, F.M., Griffith, M.S.: Estimating safety by the empirical Bayes method: A Tutorial. Transp. Res. Rec. **1784**, 126-131 (2003)

[18] Cheng, W., Washington, S.P.: Experimental evaluation of hotspot identification methods. Acc. An. Prev. **37(5)**, 870-881 (2005)

[19] Li, L., Zhang, Y.: A GIS-based Bayesian approach for identifying hazardous roadway segments for traffic accidents. El. Proc. of the 86th Annual meeting of the Transportation Research Board, Washington, D.C. (2007)

[20] Lord, D.: Modeling motor vehicle accidents 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(4)**, 751-766 (2006)

[21] 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 accidents: a Bayesian perspective. El. Proc. of the 86th Annual Meeting of the Transportation Research Board, Washington, D.C. (2007)

[22] Park, E.S., Lord, D.: Multivariate Poisson-lognormal models for jointly modeling accident frequency by severity. El. Proc. of the 86th Annual Meeting of the Transportation Research Board, Washington, D.C. (2007)

[23] Persaud, B., Lyon, C., Nguyen, T.: Empirical Bayes procedure for ranking sites for safety investigation by potential for improvement. Transp. Res. Rec. **1665**, 7-12 (1999)

[24] Sørensen, M., Elvik, R.: Black spot management and safety analysis of road networks - best practice guidelines and implementation steps. TØI report 919/2007, RIPCORD/ISEREST project (2007)

[25] 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(6)**, 991-1004 (2005)

[26] Levine, N.: CrimeStat: A spatial statistics program for the analysis of crime incident locations, vol. 1.1. Ned Levine and Associates/National Institute of Justice, Annandale, VA/Washington, D.C. (2000)

[27] Moran, P.: The Interpretation of statistical maps. J. Roy. Stat. Soc. B **10**, 243-251 (1948)

[28] Griffith, D.A.: Spatial autocorrelation: A primer. Association of American Geographers, Resource Publications in Geography, Washington D.C. (1987)

[29] Haining, R.: Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University press, Cambridge (1990)

[30] Black, W.R.: Network autocorrelation in transport network and flow systems. Geog. An. **24(3)**, 207–222 (1992)

[31] Black, W.R., Thomas, I.: Accidents on Belgium's highways: a network autocorrelation analysis. J. Transp. Geog. **6(1)**, 23–31 (1998)

[32] Getis, A., Ord, J.K.: The analysis of spatial association by use of distance statistics. Geog. An. **24(3)**, 189–206 (1992)

[33] Anselin, L.: Local indicators of spatial association-LISA. Geog. An. **27(2)**, 93–115 (1995)

[34] Schabenberger, O., Gotway, C.A.: Statistical Methods for Spatial Data Analysis. Chapman and Hall/CRC Press (2005)

[35] Besag, J., Newell, J.: The detection of clusters in rare diseases. J. Roy. Stat. Soc. A **154**,, 327-333 (1991)

[36] Waller, L.A., Gotway, C.A.: Applied Spatial Statistics for Public Health Data. John Wiley and Sons (2004)

[37] Anselin, L., Florax, R.J.G.M.: New Directions in Spatial Econometrics. Springer-Verlag, Berlin-Heidelberg (1995)

[38] Loo, B.P.Y.: The identification of hazardous road locations: A comparison of blacksite and hot zone methodologies in Hong Kong. Int. J. Sust. Transp. Forthcoming (2008)