PERGAMON

Accident Analysis and Prevention 35 (2003) 991–1004

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

The local spatial autocorrelation and the kernel method for identifying black zones
A comparative approach

Benoît Flahaut[a,c,*], Michel Mouchart[b], Ernesto San Martin[b,e], Isabelle Thomas[a,c,d]

[a] Department of Geography, Université Catholique de Louvain, Place Louis Pasteur 3, Louvain-la-Neuve 1348, Belgium
[b] Institute of Statistics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
[c] National Fund for Scientific Research, Brussels, Belgium
[d] Center for Operations Research and Econometrics, Louvain-la-Neuve, Belgium
[e] Departamento de Estadistica, Pontificia Universidad Catolica de Chile, Santiago, Chile

Received 28 February 2002; accepted 15 October 2002

Abstract

This article aims to determine the location and the length of road sections characterized by a concentration of accidents (black zones). Two methods are compared: one based on a local decomposition of a global autocorrelation index, the other on kernel estimation. After explanation, both methods are applied and compared in terms of operational results, respective advantages and shortcomings, as well as underlying conceptual elements. The operationality of both methods is illustrated by an application to one Belgian road.

Keywords: Black zones; Kernel estimators; Local spatial autocorrelation; Road accidents

1. Introduction

This article aims to present and compare two methods for identifying and delimiting road sections characterized by a concentration of road accidents (black zones): one method is based on spatial autocorrelation indices, the other one on kernel estimators.

In a previous article, it was shown that the choice of the length of the road sections has a substantial influence on the statistical description of accident count and density (Thomas, 1996). No reference was made to the spatial pattern of the accidents. By using the concept of black zone, this article tackles the problem of the length as well as the location of dangerous road sections, taking into account the contiguity structure of the basic individual spatial units. No attempt is made here to find out which factors explain the occurrence of accidents, or which countermeasures should be taken to reduce their number. The article focuses on an exploratory spatial data analysis problem: defining the location and the length of black zones.

Two statistical methods are used and compared. *Local spatial autocorrelation indices* are a decomposition of the global Moran *I* index. This method is applied to the observed number of accidents. It enables us to define the length for each black zone that best fits the reality observed, by choosing the length for which the local index is maximal. The *kernel method* is a non-parametric method that uses a density estimation technique; it enables the observer to evaluate the local probability accident occurrence, and consequently the probable dangerousness of a zone. Both methods differentiate local dangerousness and generate a smoothing of the empirical spatial process. They are applied to the same data set and the results are compared. This article also shows that although each method starts from different conceptual approaches, both may provide quite similar results under a specific choice of parameters and lead to a definition of non-contiguous black zones.

The format of the article is as follows: Section 2 presents an overview of the literature on methods for identifying hazardous road locations. Section 3 presents spatial autocorrelation and Section 4 refers to the kernel method and compares the methods. The two methods are then applied to road accidents that occurred on one Belgian road (Section 5). Section 6 concludes the paper.

[*] Corresponding author. Fax: +32-10-47-28-77.
E-mail address: flahaut@geog.ucl.ac.be (B. Flahaut).

0001-4575/$ – see front matter © 2003 Elsevier Science Ltd. All rights reserved.
PII: S0001-4575(02)00107-0

2. Background

We know that quantitative spatial data analysis is highly constrained by the limited availability of data: ideally, spatial data would be available on the finest possible spatial level in such a way that the researcher fully controlled the aggregation levels and procedures (see for instance Bailey and Gatrell, 1995, or Foterhingham et al., 2000). This is also true for road accidents (Thomas, 1996). Spatial aspects of road accidents have however often been neglected in the literature. They have recently aroused some interest, but several basic methodological aspects are still held in low esteem, such as the spatial aggregation of data and the definition of the concentrations of road accidents. This article aims to define the length and the spatial limits of road sections characterized by a concentration of accidents. Hence, a black zone is here defined as a set of contiguous spatial units taken together and characterized by a high number of accidents. The definition of these units depends to a very large extent on the finest spatial aggregation level for which data are available, i.e. the basic spatial units (BSU).

Methods developed for identifying accidents concentrations often apply to black spots, which are pinpoint concentrations of road accidents (see reviews by Silcock and Smyth, 1985; Nguyen, 1991; Joly et al., 1992; Hauer, 1996; Vandersmissen et al., 1996). In Belgium, for instance, a black spot is a 1 hm long road segment on which at least three severe road accidents are registered over 1 year. Hence, the number of accidents for these very small spatial units is likely to display a high random variation over time and/or space. Recently, identification of black zones has been considered in the literature, as arising from the awareness of the evident spatial interaction existing between contiguous accident locations. The existence of black zones reveals concentrations and hence may suggest spatial dependence between individual occurrences. Spatial concentrations may be due to one or several common cause(s). In this article, we intend to identify the location and the length of the black zones, but we do not aim at determining the causes of the accidents.

The most appropriate level of spatial aggregation for road accidents is clearly the road section, but in most studies its length is not justified and not controlled (see Thomas, 1996 for a review). No clear indication exists of what the best length of a dangerous road segment should be, nor of whether an optimal length can be defined. Deacon et al. (1975) make a distinction between "short" and "large" highway segments, respectively, called spots and sections; in their article, spots are road segments between 0.24 and 0.48 km long (0.15 and 0.3 miles), while sections are 4.8 km long (3 miles). These lengths are chosen in order to limit the heterogeneity within each road segment, but the authors recommend the use of a constant length because the interpretation of accident data would be more complicated for sections of variable length. Okamoto and Koshi (1989) propose seven ways of defining road segments: some are based on fixed lengths (100, 1200 and 2100 m) and others on variable length (one set is based on homogeneous geometric design, and three others on an error criterion). Stern and Zehavi (1990) divide the road studied into 1 km long sections, without any particular justification for this length. Elvik (1988) suggests defining dangerous road sections of a fixed length, by moving a "glider" of a specific length along the road. However, results are not found to be satisfactory with this method.

In this article, we do not define a priori the length of the road sections. Two methods are suggested for defining black zones: local spatial autocorrelation and kernel methods. Both methods tackle the problem of the location and the size (length) of black zones. They both aim to aggregate basic spatial units into black zones, taking into account the observed spatial structure of the data. These methods are respectively presented in Sections 3 and 4.

3. Spatial autocorrelation

The problem of identifying black zones consists in spatially aggregating contiguous BSUs (here, in Belgium: hectometers of roads); these should be similar in terms of geographical proximity and the number of accidents observed should be high.

Haggett et al. (1977) mention two types of spatial analysis: (1) the spatial structure analysis of the locations of spatial units i, and (2) the spatial structure analysis of the values x_i of a variable X. Spatial autocorrelation analysis belongs to the second type: it aims to evaluate the level of spatial (inter-)dependence between the values x_i of a variable X, among spatially located data, these locations being given. The concept of spatial autocorrelation is thus that there is spatial autocorrelation when the x_i's display interdependence over space. In other words, spatial autocorrelation is a spatial arrangement where spatial independence has been violated (Levine, 2000).

A simple representation of spatial dependence is the following:

$$x_i = \rho \sum_j w_{ij} x_j + u_{\mathrm{i}} \qquad (1)$$

where ρ measures the spatial autocorrelation between the x_i's, w_{ij} are the weights deemed to represent proximity relationships; they are often a function of distance, e.g. $w_{ij} = d_{ij}^{-a}$, u_i are independent and identically distributed errors, with variance σ^2.

Formulation (1) extends to space the idea of temporal autocorrelation (Tiefelsdorf, 2000, p. 2 ff.), but space makes the spatial neighborhood multidirectional and hence much more complex, leading to specific indices for autocorrelation.

Specifically, spatial autocorrelation analysis assesses the extent to which the value of a variable X at a given location i is related to the values of that variable at contiguous/

neighboring locations. The assessment involves analyzing the degree to which the value of a variable for each location co-varies with values of that variable at contiguous or nearby locations. When the level of co-variation is higher than expected, contiguous locations have similar values and autocorrelation is positive. When the level of co-variation observed is negative, high values of the variable are contiguous with low values and the autocorrelation is negative. The lack of significant positive or negative co-variation suggests the absence of spatial autocorrelation.

Global methods of assessing spatial autocorrelation have existed for several decades and mainly stem from the work of Moran (1948) (see Cliff and Ord, 1973, 1981; Griffith, 1987; Anselin, 1988; Odland, 1988; Haining, 1990; Tiefelsdorf, 2000). For quantitative variables, Moran's I and Geary's c are the indices most often used to assess the global level of spatial autocorrelation. Moran's I is preferred for its greater general stability, its testability, the flexibility of its conditional distribution as well as its usefulness in applied work. Its usefulness for transport fluxes and road accident analyses has previously been discussed (Black, 1991,1992; Black and Thomas, 1998). Let us call to mind the definition of Moran's I index:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2} \tag{2}$$

where w_{ij} is the weights representing proximity relationships between location i and neighboring location j; they form together a spatial contiguity matrix.

$$S_0 = \sum_i \sum_j w_{ij}$$

$$z_i = x_i - \bar{x}$$

$$z_j = x_j - \bar{x}$$

where x_i is the value of the variable X at the location I, $\bar{x}$ the mean of all $\mathrm{x_i}$'s, n the total number of locations, $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, n$.

For any variable X, the mean as well as the deviation of any one observation from that mean can be computed. The statistic then compares the value of that variable at any location i with the values at all other locations j. In Moran's initial formulation the weight variable is a contiguity 0–1 matrix; Cliff and Ord (1973) generalized these definitions to other types of weights.

3.1. Local indices of spatial autocorrelation

Local indices of spatial autocorrelation are more recent (Getis and Ord, 1992; Ord and Getis, 1995; Anselin, 1995; Boots and Tiefelsdorf, 1995, 1997). Each spatial unit i is now characterized by one value of the index; it gives the individual contribution of that location in the global spatial autocorrelation measured on all n locations. The local Moran's I index computed at location i is called a LISA (Local Indicator of Spatial Association, Anselin, 1995) when:

$$I_i = z_i \sum_j w_{ij} z_j \tag{3}$$

and

$$\sum_i I_i = \sum_i z_i \sum_j w_{ij} z_j \tag{4}$$

and

$$\sum_i I_i = \gamma I \tag{5}$$

with $\gamma = S_0 m_2$ and $m_2 = \sum_i z_i^2 / n$.

In other words, a LISA is an indicator of the extent to which the value of an observation is similar to or different from its neighboring observations. This (1) allows a value I_i to be associated with each basic spatial unit i, and (2) requires that the notion of neighborhood (w_{ij}) among the BSUs is specified. Neighborhood can be simply defined by contiguity (0–1 matrix), or by any other function of distance between the BSUs.

The global autocorrelation index I is broken down into n local components noted I_i (5). The local component I_i is the product of the centered local value (z_i) and the weighted mean of the centered neighboring values ($w_{ij}z_j$) (3). The sign of this component is an indicator of the agreement (positive) or disagreement (negative) of the signs of the local centered value and the weighted average of the neighboring values. Hence, similarly to the global index, I_i can be positive, negative, or equal to zero. It is negative when there is an association of opposite values at neighboring locations, and positive in the case of spatial association of similar values.

For the identification of the black zones, the product of positive z values for one BSU and for the weighted average of its neighboring BSUs is the only one taken into consideration. These positive values correspond to a large number of accidents, large being defined as greater than the mean ($x_i > \bar{x}$ and $\sum_j w_{ij} x_j > \bar{x}$). Such indices are noted I_i^* later in the text. The weights w_{ij} are here row-standardized ($\sum_j w_{ij} = 1, \forall i$). Based on the number of accidents per hectometer and on a spatial contiguity matrix, local spatial autocorrelation indices I_i^* are computed for the n BSUs and used for estimating the dangerousness of a road. The *intensity* of the dangerousness depends on the value of the local Moran's I_i^* and the *length* of the black zone depends on the w_{ij} matrix. The number of neighboring BSUs taken into account in the computation of the index (BSUs j for which $w_{ij} \neq 0$) determines the length of the zones. A different length can be defined for each black zone by successively computing the index I_i^* with different numbers of neighbors for a given i (for example 10 different neighborhoods corresponding to 10 different lengths). Then, for each i (center of a zone) the length is chosen for which the I_i^* index is maximal; it corresponds to the best measure of the association between one BSU and its neighbors. Hence, this method leads to the definition of black zones of various lengths, the length depending on the observed local spatial structure of the number of accidents. The application of such a method in road accidents analysis is quite novel.

4. Kernel methods

An alternative approach for identifying black zones is to evaluate an intrinsic local degree of dangerousness, i.e. to evaluate how a global level of dangerousness is distributed over space. This approach may receive a natural interpretation in the framework of a non-parametric model and the numerical evaluation is eventually based on a kernel method. The nature of the kernel methods is first briefly presented in the framework of a non-parametric density estimation. A simple model of the spatial distribution of dangerousness is then sketched.

4.1. A non-parametric density estimator: some generalities

Consider a set of observed data points assumed to be a sample from an unknown probability density function, say *f*. Density estimation is the construction of an estimator of the density function *f* from the observed data. The *kernel estimator* of an unknown density *f* is defined by:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

where *h* is the so-called smoothing parameter or bandwidth and the function *K*, called a "kernel", is usually a symmetric probability density function. The kernel estimator is therefore an average of "bumps" placed at each observation point; the kernel *K* determines the shape of the bumps while the bandwidth *h* determines their width. For a motivation of this definition and more details on the method in general, see Silverman (1986, Chapters 1 and 2); for short presentations oriented toward spatial data analysis see Bailey and Gatrell (1995, Section 3.4.2) or Foterhingham et al. (2000, Sections 4.5 and 6.6); for an example of using such a tool in crime analysis, along with a specialized package called Crimestat, see Levine (2000).

Thus, the kernel estimator depends on two parameters: the bandwidth *h* and the kernel density *K*. It may be shown that the density kernel estimator is generally robust with respect to kernel choices; this eventually justifies the usual choice of a Gaussian kernel (for details, see Silverman (1986, Chapter 3). For a given kernel *K*, the kernel estimator critically depends on the choice of the smoothing parameter *h*. An appropriate choice of the smoothing parameter should be determined by the purpose of the estimate. Silverman (1986, Section 3.4.1) suggests a subjective choice of the smoothing parameter if the purpose of the estimation is to explore the data in order to propose possible statistical models and hypotheses. In addition, he suggests an automatic choice of the smoothing parameter, which may be considered as a starting point for subsequent subjective adjustments (Silverman, 1986, p. 44). Indeed, an *optimal smoothing parameter* h_{opt} may be obtained by minimizing the approximate integrated mean square error; such an optimal bandwidth is proportional to $n^{-1/5}$, where *n* is the sample size. The constant of proportionality depends on the unknown density *f*; for computing it, iterative methods are typically used (see Silverman, 1986, p. 40). The initial iteration often makes use of a *reference bandwidth* h_{ref}, defined by both the kernel *K* and the unknown density *f*; when *f* is Gaussian with variance σ^2 the reference bandwidth is obtained by:

$$h_{\mathrm{ref}} = 1.06\,\sigma\,n^{-1/5} \tag{7}$$

In sum, a reference bandwidth h_{ref} as well as an estimation of the ideal bandwidth h_{opt} may be computed from a given set of data; these smoothing parameters may be used as starting points for subsequent subjective adjustments, which in turn depend on some specific requirements motivated by a particular application.

4.2. A simple model for dangerousness distribution

As previously pointed out, a natural use of density estimation is a description of some properties of a given set of data: density estimation may indeed give valuable indications on such features as skewness or multimodality (i.e. the presence of several local maxima in the density) in the data. In this road safety application, density estimation is used to describe the dangerousness of a given road. The first problem to be considered consists in proposing a statistical model, which suggests a statistical meaning of the dangerousness concept. The descriptive analysis of the data is thus operated with a reference model in mind; by so doing the arbitrariness of a descriptive analysis is controlled. For details concerning the role of a model in the statistical interpretation of a descriptive technique, see Dempster (1971, Section 3), Cox (1995) and the introductory section of Mouchart and San Martin (2002).

A road is identified by an interval of the real (positive) line; its length, measured in hectometers, is denoted *L*. Let *N* be the number of accidents that occurred on that road and $i \in \{1, \ldots, L\}$ be the possible locations of these accidents on that road. Let us further define a random variable $Y_n \in \{1, \ldots, L\}$ as the location of the *n*th accident, that is:

$$\{Y_n = i\} \Leftrightarrow \text{the accident } n \text{ has occurred at the hectometer } i \tag{8}$$

The number of accidents that occurred in any interval $]a, b]$ (that is, a section of the road) is therefore given by:

$$\sum_{n=1}^{N} I(]a, b])\ (Y_n) \tag{9}$$

where $I(]a, b])(Y)$ is equal to 1 when $Y \in]a, b]$, and to 0 otherwise.

For a given road, the data generating process is hierarchically specified as follows. A first random process generates the number of accidents, namely N. This marginal model generating N characterizes the global dangerousness of a (fixed) road. A second random process generates the location of each accident on that road given the number of accidents N. This second process may be assumed to take the form of a non-homogeneous Poisson process.

More explicitly, let us assume that (1) the number of accidents on any interval, say J, of the road under consideration is Poisson distributed with intensity parameter $\lambda(J)$, where $\lambda(\bullet)$ is a measure on the road giving a measure N to the entire road[1]; (2) the number of accidents of any two disjoint intervals are independent random variables, and (3) the density of the intensity function, representing the instantaneous dangerousness, exists and is not a constant (hence the non-homogeneity of the process) but a smooth function of the location. Here, the clustering of the accidents is interpreted as an effect of the smooth variation of the intensity function of the process. From a formal point of view, the (non-parametric) estimation of the intensity function may then be shown as being similar to the (non-parametric) estimation of a density function.

4.3. Comparing kernel estimators with spatial autocorrelation

The kernel method operates a smoothing of the empirical process, and is characterized by two features: the choice of the kernel and that of the window; together they determine the structure and the length of the neighborhood. The computation of the Moran's I_i can also be interpreted as a smoothing of the empirical spatial process: indeed, the spatial autocorrelation method is based on the idea that accidents tend to be concentrated on certain segments of the road, and the sequence of local indices may be viewed as a sequence of moving averages of the deviance. Thus, the observed clustering of accidents is interpreted as an effect of the presence of spatial autocorrelation. The smoothing is also characterized by the weighting system and the definition of the neighborhood, which in turn determine the structure of the black zones (location, dangerousness, and length). In both methods, the smoothing parameter can be controlled: a wider window corresponds to a wider concept of a dangerous zone whereas a narrower window corresponds to a more peaked approach.

Spatial autocorrelation aims to detect the operation of an unknown underlying spatial process. Following Tiefelsdorf (2000), the spatial structure is regarded as a linking substance and functional connection between interrelated spatial objects (here: the hectometers of a road) on which a spatial process (here: the occurrence of road accidents) evolves with a given strength (evaluated with a measure of the spatial autocorrelation). A possible interpretation of local indices is to consider that they identify those spatial objects which have a significant impact on the global spatial process. As they belong to the class of local indicators of spatial association, they allow one to localize spatial clusters (black zones).

The kernel method is also based on the idea of an underlying model of dangerousness. The spatial process is here specified a priori: the (local) intensity of the underlying Poisson process is an explicit formalization of the (local) concept of dangerousness in the context of a random measure (which is the natural formalism for the random location of indistinguishable objects). The clustering of the accidents is here interpreted as an effect of the shape of the intensity function of the process, this shape being assumed to be smooth. The intensity of the Poisson process gives its expectation; thus the expected number of accidents in an arbitrary interval is given by the intensity of that interval (remember that the intensity is a measure) and similarly to a probability measure, its density gives a heuristic "instantaneous" expectation.

Finally, with the kernel method, the data are observed from a modeling perspective that is specified a priori (a Poisson process). In this sense, the Poisson process may be viewed as a possible explanation of the spatial structure. Spatial autocorrelation indices are more related to an exploratory analysis: they detect the existence of a spatial process in the data. Nevertheless, there is no hypothesis related to the subject of the spatial process, namely, the distribution of accidents and/or the distribution of their explanatory causes. The identification of these potential explanatory variables can be analyzed in a further step.

Although the two methods are based on different approaches, they may produce quite similar numerical results, understood as local indices of dangerousness. Empirical results, in the next section, show how the similarity of results can be controlled by a judicious choice of the parameters. Interpretability of the empirical results rests therefore on a proper understanding of the meaning of the selected parameters.

5. Application to one Belgian road

5.1. The data set

In Belgium, any road accident that occurs on a public road and that involves casualties is officially reported. Its

[1] A more explicit expression for "probability" would be "a measure of probability"; mathematically, a (positive) "measure" is characterized by the same properties as a measure of probability except that the measure of the universe is any positive real number or +infinity; in particular, the measure of two disjoint sets is the sum of the measure of each set. We use this concept to measure the dangerousness of a section road given the global level of the complete section: this is a problem similar to "distribute" a probability over a segment of line, except that the total dangerousness of the section, i.e. the number of accidents, is not equal to one. It is furthermore natural to assume that the distribution of dangerousness is "smooth" and accordingly represented by means of a density.

location is accurately known on numbered roads because there is a stone marker every hectometer; numbered roads are motorways, national, and provincial roads linking towns. Hence, this case study is limited to accidents with casualties on numbered roads; the hectometer (hm) is accordingly the smallest spatial unit for which accident data are spatially available (called basic spatial unit, or BSU). The period under study is 1992–1996, a period long enough to limit random fluctuations, and short enough to limit changes in road and traffic conditions. In Belgium, 137,213 accidents with casualties occurred on the 15,359 km of numbered roads during that period.

The study was originally applied to all numbered roads in Belgium (Flahaut, 1999), but this paper is limited to one road: the *N29*, a quite dangerous road which goes from Charleroi (0 hm) to Jodoigne (527 hm) (Fig. 1). It is a two-lane road, joining several (small) towns, with an average daily traffic of 10,600 vehicles. Five hundred and ten accidents were recorded between 1992 and 1996 and at least one accident was recorded on 259 hm out of 527. Restricting the data to one road allows one to remove a large part of the variation in driving conditions, and hence to better control the tested effects.

5.2. The autocorrelation method

5.2.1. Defining black zones

When computing spatial autocorrelation, one of the most important problems is the representation of the contiguity structure between the BSUs (Anselin, 1988, 1995). As no information is available on the relationships between adjoining hectometers (that would be ideal according to Cliff and Ord, 1973, 1981), the "true" structure of the weights is unknown. Therefore, the choice of the weighting structure is never objective.

Two elements define the contiguity structure: the number of neighbors (or level of contiguity) and the weights representing the proximity.

In this application, the *number of neighbors* determines the length of the black zones. Local spatial autocorrelation indices I_i^* are successively computed for 10 different levels of contiguity. More specifically, the number of neighbors varies from 2 to 20 with a path of 2 (2, 4,. . . , 20); they are symmetrically distributed on both sides of each studied hectometer *i*, and lead to corresponding potential black zones of 3–21 hm. For instance a 21 hm long black zone is made out of 20 neighbors, 10 on both sides of the studied hectometer. For each individual hectometer *i*, 10 Moran's I_i^* indices are computed and compared. This comparison is important to ensure some objectivity in the subjective choice of the weighting structure. The length of the black zone is then defined as the length for which the Moran's I_i^* is maximized; this value corresponds to the highest measure of the spatial association of large numbers of accidents between 1 hm and its neighbors. This process allows one to choose the level of contiguity that is best adapted to the local spatial structure of the accidents observed: the length of a black zone is not fixed a priori, but differs for each black zone.

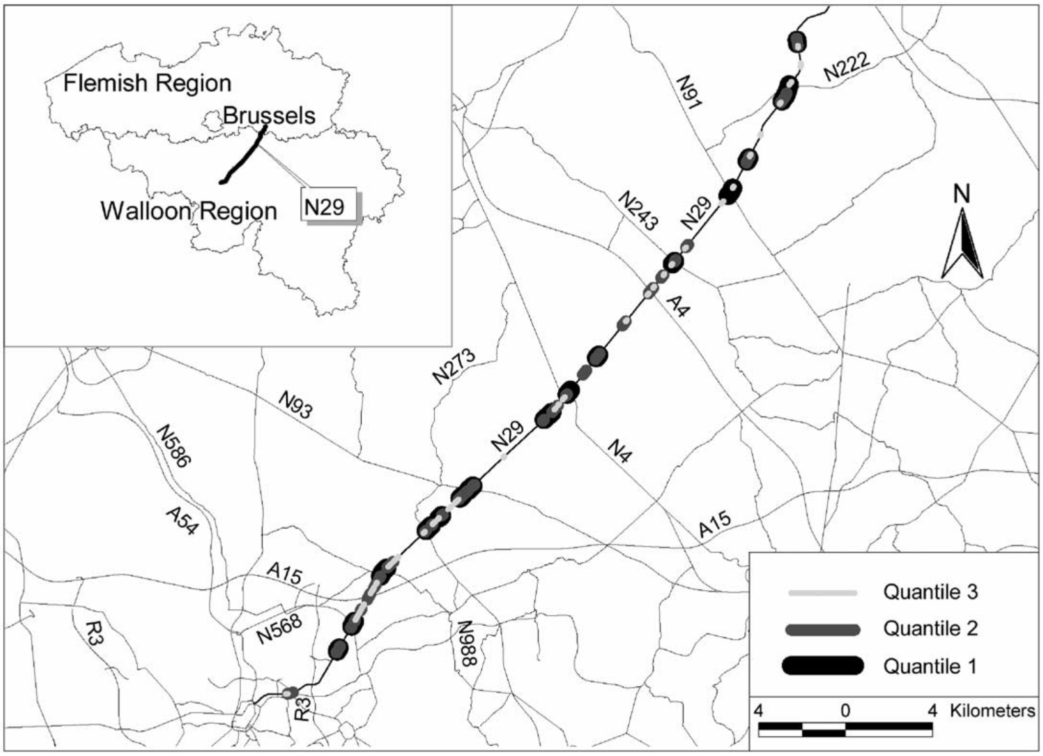

Fig. 1. The *N29*: its location in Belgium and the black zones defined by the local autocorrelation indices I_i^* (weights function of d_{ij}^{-2}).

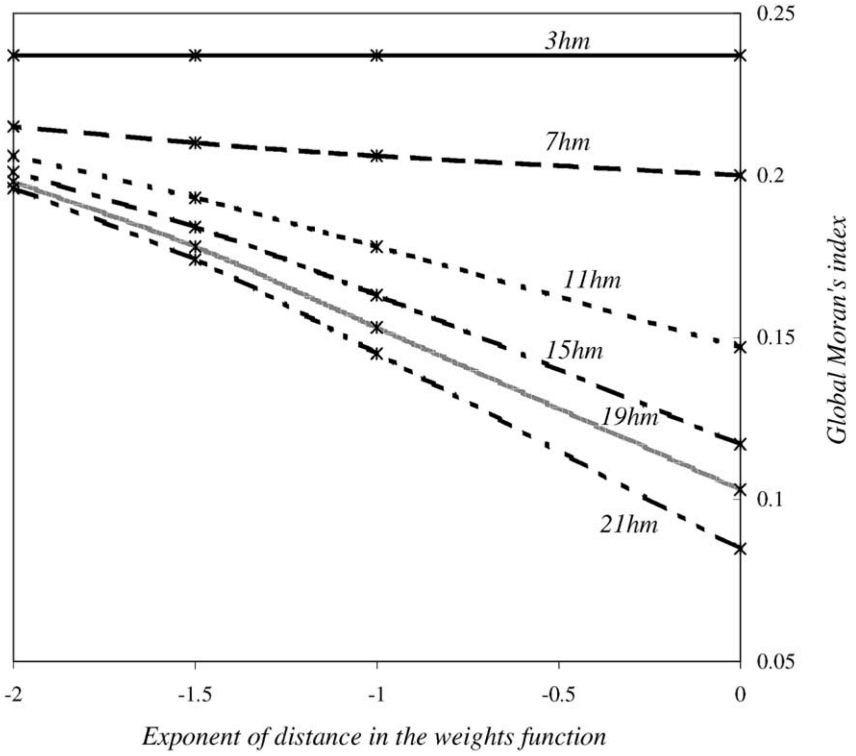

Fig. 2. Variation of the global Moran I index with different exponents for d_{ij}.

For a given number of neighbors, the *weights* w_{ij} represent the relations of proximity between the hectometers. As mentioned earlier, the simplest way to define them is a 0–1 matrix where $w_{ij} = 1$ when i and j are contiguous, and otherwise 0. Using other values for w_{ij} enables us to better distinguish between "close" and "not so close" locations. Such weights should decrease with distance (Cliff and Ord, 1973, 1981; Haining, 1990). In the first application, weights are inversely proportional to the squared distance (d_{ij}^{-2}); sensitivity analyses to other values of the exponent to distance are reported in Section 5.2.2.

Results are illustrated in Fig. 1. For cartographic reasons, local Moran's I_i^* indices are classified into five quantiles; black zones belonging to the first three quantiles are mapped.

5.2.2. Sensitivity analyses

In a first set of sensitivity analyses, several *levels of contiguity* (number of neighbors) are successively considered leading to fixed lengths for black zones (3, 5 hm, etc.). Each length leads to a different data set. For each data set, a global index of spatial autocorrelation is computed with different distance functions. Fig. 2 shows the increase in the global I values when decreasing the exponent of distance in the weight function, namely d_{ij}^{0}, d_{ij}^{-1}, $d_{ij}^{-1.5}$, and d_{ij}^{-2}. The largest spatial association is obtained with d_{ij}^{-2}. Fig. 3 shows the variation of the global Moran I index with the level of contiguity, for d_{ij}^{-2}. A value of -2 for the exponent of distance means a stronger reduction in the importance of the hectometers located further away, making the spatial association stronger with closest neighbors. This is confirmed by a tendency of the global index to decrease when increasing the number of neighbors, for a given weight function. These variations show the usefulness of not fixing a priori the number of neighbors: the local spatial structure of the accidents is taken into consideration by choosing a different length for each black zone.

In a second set of sensitivity analyses, optimally[2] defined black zones are considered as one data set; in this case, the level of contiguity is not fixed a priori, but is optimized locally. An increase in the global I index with a decreasing exponent of distance in the weight function is also observed. These observations justify the choice of d_{ij}^{-2} for this study, in addition to the consistency with the literature on spatial interaction. Therefore, the use of the exponent of -2 induces a reduction in the number of black zones as well as in their mean length (Fig. 4).

To conclude, a dangerousness index based on local spatial autocorrelation appears to be a quite satisfactory method for identifying the location and the length of dangerous road sections. The variable number of neighbors taken into consideration in the computation of the local Moran's I_i^* indices allows us to choose a length that is locally optimized for each

[2] Optimal black zones correspond to these for which the local index I_i^* is maximized among 10 different spatial structures (equivalent to 10 different lengths, see Section 5.2.1).

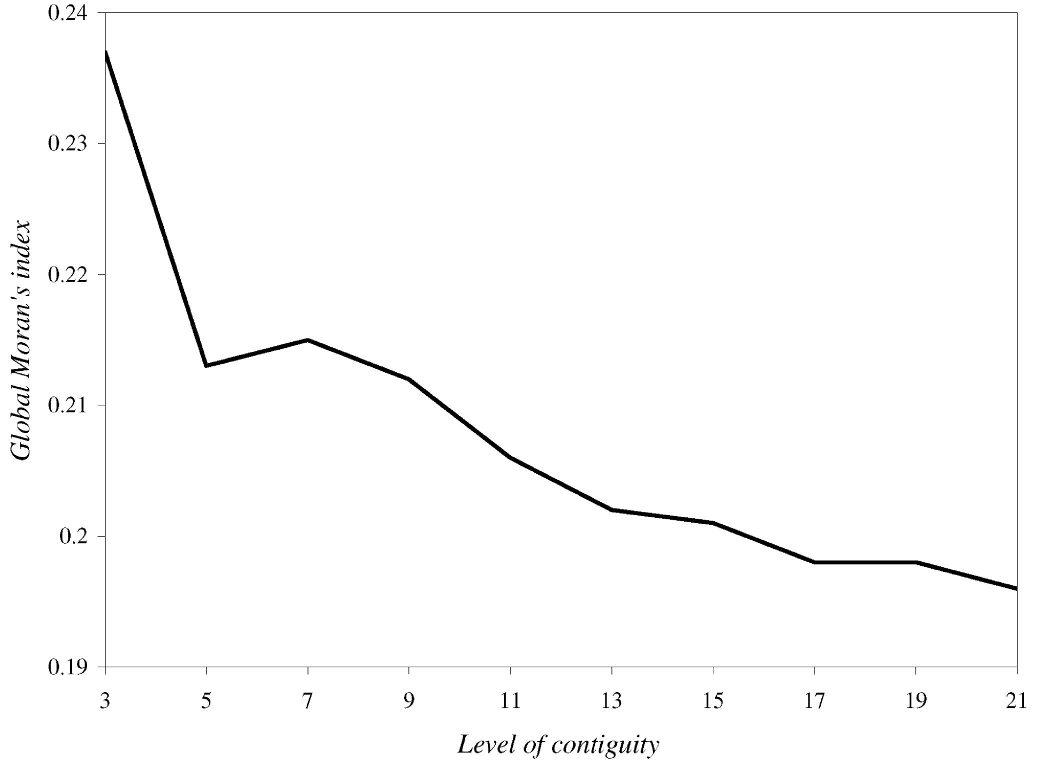

Fig. 3. Variation of the global Moran's I index with the level of contiguity (for d_{ij}^{-2}).

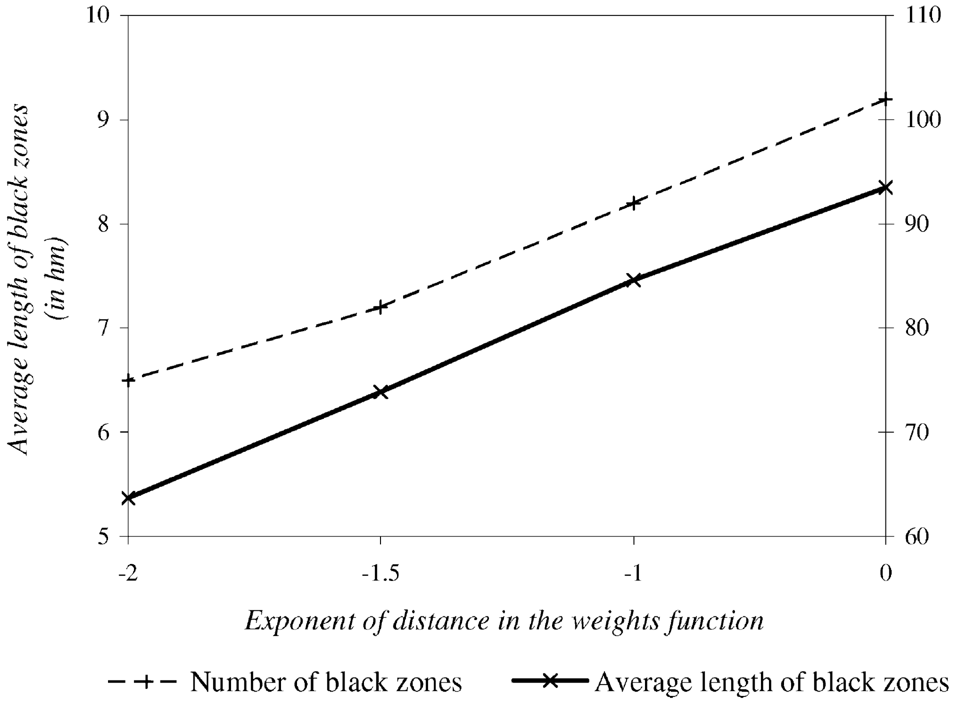

Fig. 4. Variation in number and average length of black zones according to weights function.

black zone, according to the local spatial structure observed. In this way, by proposing a priori several different spatial structures for the weights, the subjectivity of one particular structure is minimized. Not all identified black zones need to be treated in order to reduce their dangerousness. Among these zones, all are dangerous because they are identified as black zones, but some are more dangerous than others. In Section 5.3, the kernel method is applied to the same data set and the results are compared in Section 5.4.

5.3. *The kernel method*

5.3.1. *Choices of parameters*

As explained in Section 4, the intensity of the Poisson non-homogeneous process, conditional on the total number of accidents N, is equal to N times a probability measure. Therefore, a local index of dangerousness may be obtained as N times a density estimator:

$$\hat{r}(i|N) = N\hat{f}(i|N)$$

and may be heuristically interpreted as the expected number of accidents at hectometer i, given the total number N. A kernel estimator for the density produces a (non-negative) smooth function, whose integral, over the entire road section under analysis, is equal to N. The definition of black zones is based on a truncation of this smooth estimator. In other words, a dangerous zone is defined by neighboring hectometers sharing a value of this index higher than a given threshold. Such a threshold may be defined either by means of an a priori given acceptable level or by means of a more statistical approach. Indeed, a hectometer may be considered dangerous if its local intensity is higher than the level corresponding to a given quantile of the distribution (for instance, the median). Such a concept of dangerousness eventually depends on two choices: that of the degree of smoothing of the kernel estimator and that of the truncation procedure.

After fixing a Gaussian kernel, this estimator only depends on the smoothing parameter h. We have at least two “automatic choices” for this parameter. The first one is given by the reference bandwidth h_{ref} (see (7)); the second one, h_{opt}, is provided by a minimization criterion, namely minimization of the integrated mean square error. As the density estimator is sensitive to the bandwidth-parameter choice, so is the $\hat{r}(i|N)$ estimator.

In this case study, the following procedure is used to estimate the expected number of accidents at hectometer i. For each hectometer i, the density estimator $\hat{f}(i|N)$ is computed by using the reference bandwidth h_{ref}. We use the method implemented by Gasser et al. (1991) because it is adequate (1) when the data observed belongs to a bounded interval (in our case, the accidents observed belong to a bounded road), and (2) when the data have a multimodal density function, a feature to be expected and actually observed, for an analysis of dangerousness (see Fig. 5).

The sensitivity of $\hat{f}(i|N)$ is then evaluated by considering two other bandwidth parameters: the first one, an

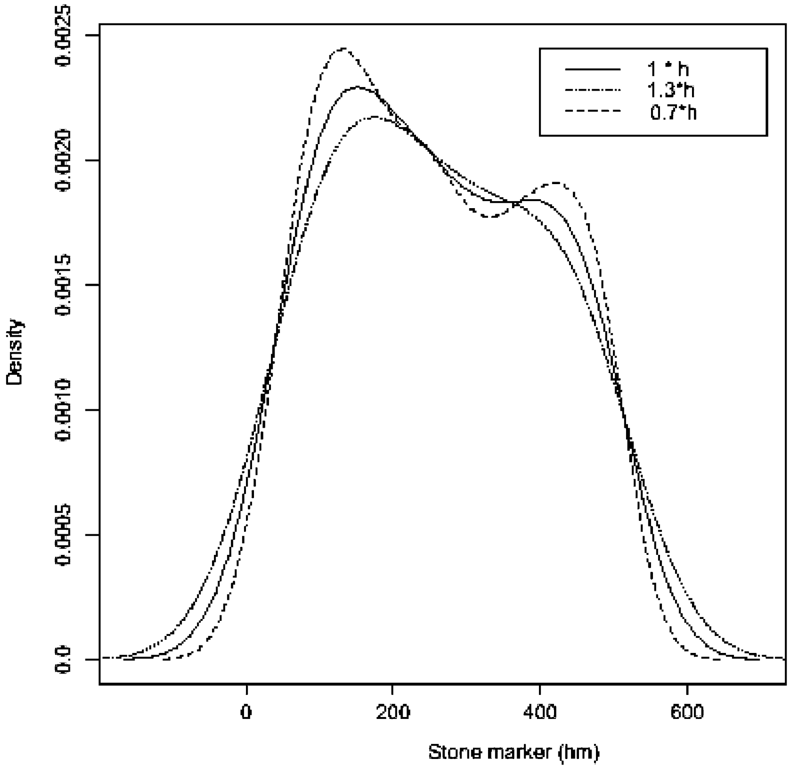

Fig. 5. Density estimator for three bandwidths ($h = h_{\text{ref}} = 59.6$ hm).

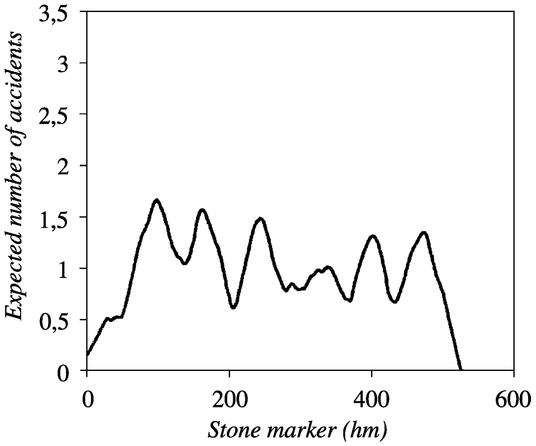

Fig. 6. Estimator of the expected number of accidents for $h_{\text{opt}} = 26.5$ hm.

over-estimation of h_{ref}, namely, 1.3 h_{ref}, the second one, an under-estimation of h_{ref}, namely 0.7 h_{ref}. The value of this step consists in controlling the multimodality of the unknown density. Moreover, we have also considered the following weights in the framework of an automatic choice of the bandwidth parameter, not only h_{opt}, but also 0.7, 0.5, 0.3, and 0.2 h_{opt}.

5.3.2. *Interpretation of the results*

5.3.2.1. Global estimator. For the *N29* data set, the reference smoothing parameter h_{ref} is 59.6 hm. Fig. 5 shows the density estimator by using h_{ref} as well as 0.7 h_{ref} (41.7 hm) and 1.3 h_{ref} (77.5 hm). Conditional on the number of accidents on the *N29*, these global density estimators allow us to determine how accident locations are distributed. Consequently, a global vision of the dangerousness of the *N29* is obtained: the first 250 hm of the *N29* are more dangerous than the last ones. Furthermore, Fig. 5 also suggests bimodality in the data. It should be noted that the overall pattern of the density estimates is rather insensitive to the three different values of h_{ref}.

5.3.2.2. Automatic estimator. For the *N29*, the optimal (or automatic) smoothing parameter h_{opt} is equal to 26.5 hm. Fig. 6 shows the expected number of accidents at each hectometer.

Given that h_{opt} is significantly lower than h_{ref}, the sensitivity of this estimator to small values of h is explored by multiplying the optimal smoothing parameter h_{opt} by 0.7 (18.5 hm), by 0.5 (13.3 hm), by 0.3 (8.0 hm) and finally by 0.2 (5.3 hm). Fig. 7 shows the corresponding family of estimators of the expected number of accidents at each hectometer for the *N29*. It may be seen that density estimators (and therefore expected numbers of accidents) corresponding to smaller smoothing parameters display more peaks than those corresponding to larger smoothing parameters. Thus, this family of estimators actually provides an alternative picture of the dangerousness distribution. With a small smoothing parameter, attention is focused on local conditions and if h is too small, the resulting pattern represents a minor smoothing of the data, keeping more structural features hidden. Nevertheless, it is important to remember that these peaks may be produced not only by the actual presence of a dangerous zone, but also by the variance of the estimator, the difficulty being that the variance increases when h decreases.

The choice of a smoothing parameter depends on the objective of the analysis. For a global vision of the dangerousness of a given road, a global estimator may be a good starting point. A smaller window allows for a narrowing of the global description. From the automatic choice, a family of smaller smoothing parameters may be computed; that is, if we need to describe the dangerousness of *small zones*, small smoothing should be used.

5.4. *Comparing the empirical results*

Although the two methods are basically different, they may provide similar numerical results under specific choices of the parameters required by each approach.

5.4.1. *Definition of the dangerous zones*

For spatial autocorrelation, the definition of neighborhood and proximity govern the structure of the black zones (contiguity level, weights discussed previously). Moreover, this paper is limited to the delimitation of dangerous zones; positive autocorrelation (product of high values) is the only one taken into account (see Section 3.1). This leads to the definition of a black zone: a high number of accidents in hectometer i and a high number of accidents in its neighboring hectometers. Indices resulting from other combinations have here been ignored.

In the kernel method, the definition of the zones depends on the choice of the smoothing parameter (window) and on the choice of a truncation point in the estimated intensities. The choice of window can be approached in two different ways. One is to consider a well-specified process, such as a non-homogenous Poisson process, and to choose the smoothing parameter based on the statistical properties of the implied estimator (both asymptotic and small sample properties). The other concentrates on the descriptive features and considers that different smoothing parameters correspond to different types of dangerousness. Thus, a small value corresponds to a more local approach and is more appropriate for detecting black spots, whereas a larger value corresponds to the “fuzzier” concept of a zone to which the attention of people in charge of road safety should be drawn. Note that the distribution, and hence the interpretation, of the autocorrelation coefficients may also be evaluated under a non-homogenous Poisson process specification (i.e. the distribution of a statistic under a particular model) but this is beyond the scope of this paper.

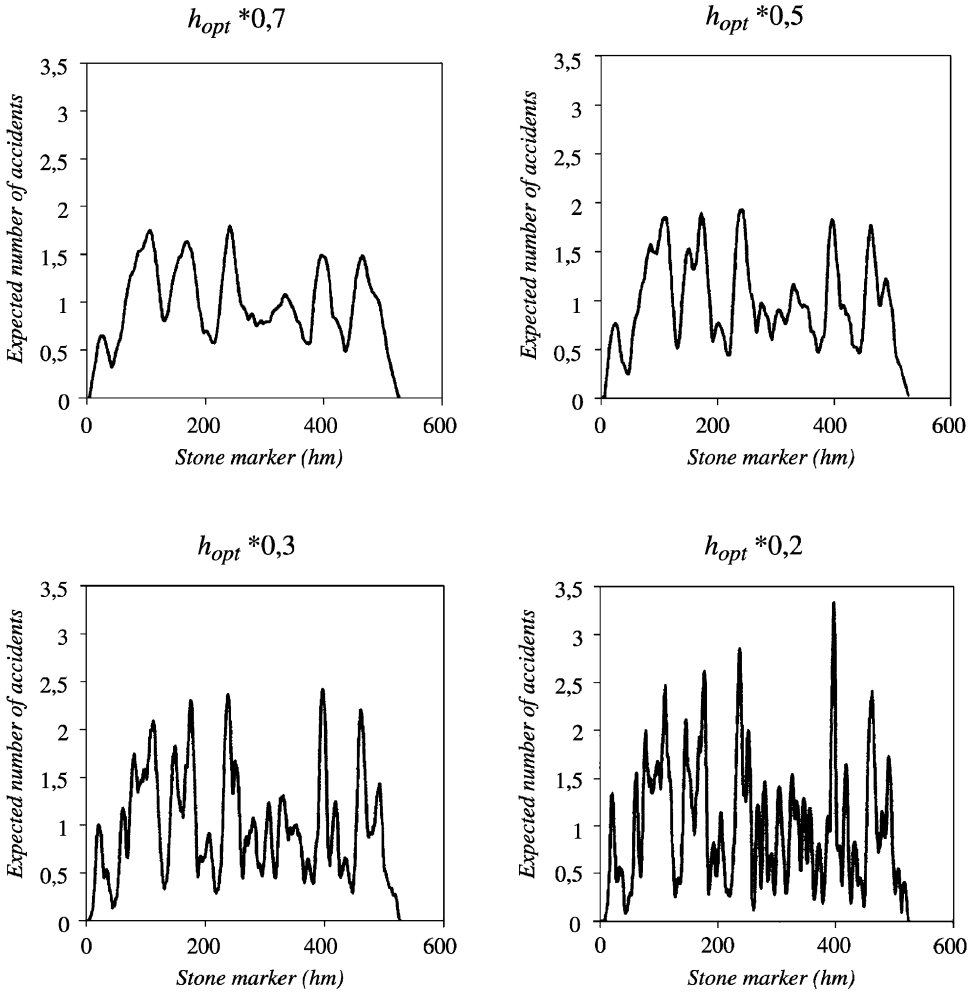

Fig. 7. Sensitivity of the estimator of the expected number of accidents to the smoothing parameter.

5.4.2. Comparing the values of the indices

Comparing empirical results means using comparable parameters in the formulation of both indices.

In a first step, the kernel estimators used in Section 5.3 are maintained and local spatial autocorrelation indices are computed with a number of neighbors corresponding to the length h_{opt} of the kernel method (26.5 hm for the *N29*); weights w_{ij} of these neighbors are computed with the kernel function *K* (see Section 4). Pearson's correlation coefficient between both indices is positive and significant ($r = 0.53$, $\alpha = 0.001$) but rather low compared to those obtained later in this section. This weak value can be explained by the fact that such a length of black zones ($2h = 53$ hm) is too large for a local approach: road segments with few accidents and road segments with more accidents are associated in a same black zone. A local approach breaks up such a zone into several zones (black and not black).

In a second step, the parameters of the kernel method are aligned with those of the local spatial autocorrelation. The average length of black zones identified with the spatial autocorrelation method varies between 5 and 9 hm depending on the decreasing function of the weights. Hence, a bandwidth *h* of 2.5 and 5 hm is used successively to compute the kernel estimators. A window *h* of 10 hm is also considered in order to obtain a better perception of the relations. The kernel estimators are compared with the autocorrelation indices I_i^* computed with four weight functions (d_{ij}^0, d_{ij}^{-1}, $d_{ij}^{-1.5}$ and d_{ij}^{-2}). Pearson's coefficients indicate a high and significant positive relationship between both types of indices, the correlation being higher when decreasing *h* from 10 to 5 and 2.5 hm (Table 1). The highest correlation is obtained with a window *h* of 2.5 hm for the kernel method and with d_{ij}^{-2} weights for the spatial autocorrelation method ($r = 0.86$).

This comparison shows a good resemblance between both indices when these are made comparable by using similar values for the parameters.

To conclude these comparisons, Fig. 8 illustrates the location of black zones identified by both methods. The smoothing parameter *h* for the kernel estimator is successively 26.5 and 2.5 hm. Only the upper 50% of kernel estimators are considered to allow a comparison with the autocorrelation

Table 1
Pearson's correlation coefficients between the autocorrelation index computed with different weights functions w_{ij} and the kernel method index computed with different sizes of the bandwidth h (significance $\alpha = 0.001$)

Size of bandwidth (kernel) (hm)	Weights function (autocorrelation)			
	d_{ij}^{0}	d_{ij}^{-1}	$d_{ij}^{-1.5}$	d_{ij}^{-2}
$h = 2.5$	0.76	0.84	0.85	0.86
$h = 5$	0.74	0.79	0.77	0.80
$h = 10$	0.63	0.61	0.55	0.62

results, because these identify only the dangerous zones while kernel indices give a level of dangerousness for all hectometers (from the most secure to the most dangerous). Hence, it is more judicious and coherent to consider only the dangerous hectometers of both approaches. Indices are classified in five quantiles.

5.4.3. Differences observed

Up to this point, this section has been limited to similarities between the results obtained by both methods; differences also exist.

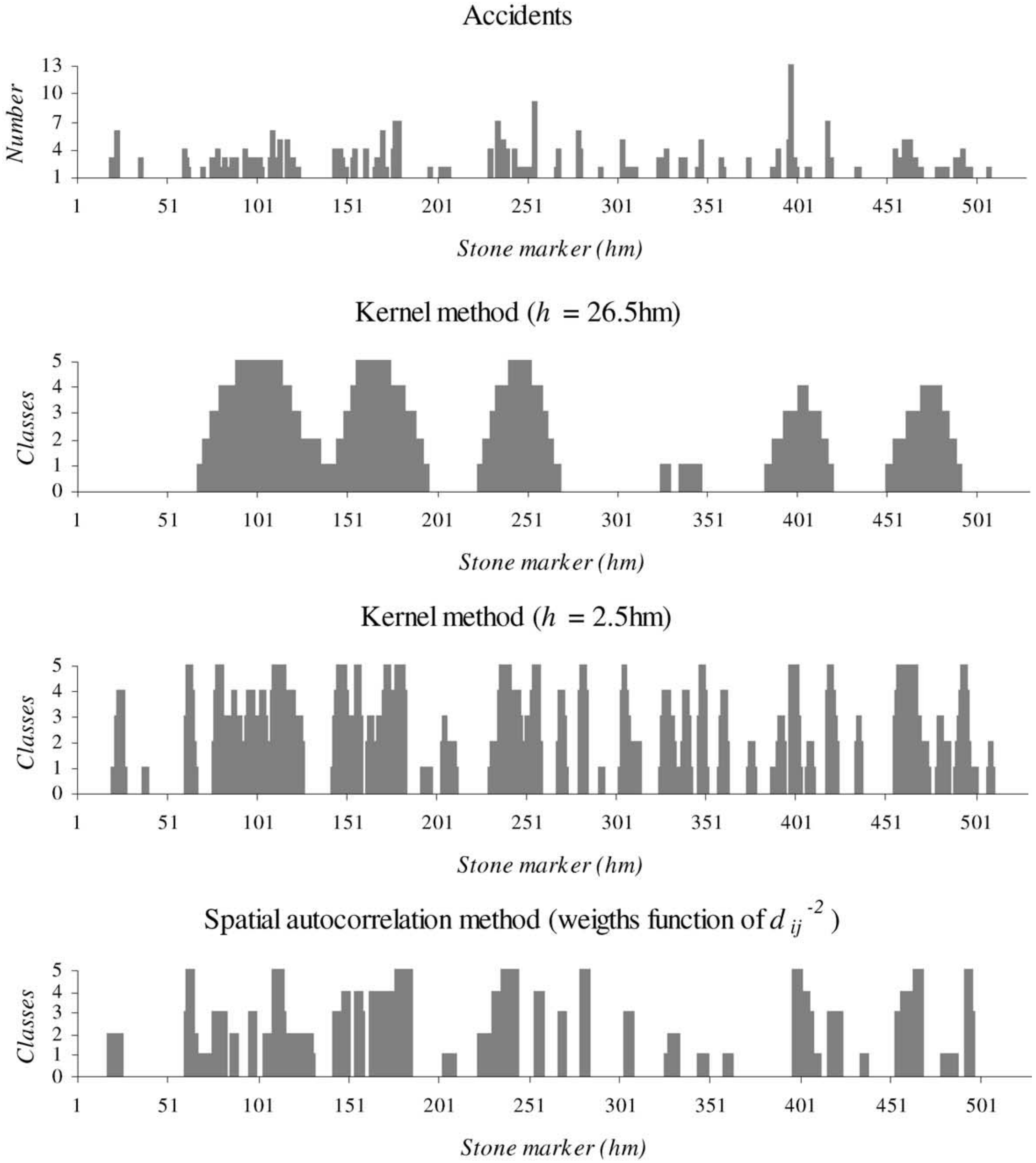

Fig. 8. Comparison of the dangerousness of the *N29* determined by both methods.

The presentation of the results is different in the sense that a kernel estimator is associated with each hectometer, while for the local Moran's index, a dangerousness index is attributed to each center of a black zone and is extended to the entire black zone.

An interesting characteristic of spatial autocorrelation is the variability of the length of the black zones; this is not the case for the kernel method where the length of the window is constant for the entire road. Hence, the spatial autocorrelation method could allow a better adaptation to the local spatial structure for a given road by giving two types of results (the dangerousness of a black zone and its length), while the kernel method gives the dangerousness of each hectometer (the length of the window is chosen a priori, before the computation of the kernel estimators).

6. Conclusion

In this article, two different statistical techniques are used to solve a geographical problem associated with road accidents: the definition of black zones, that is to say road sections on which the number of accidents is high. The local autocorrelation index and the kernel method have both their advantages and their drawbacks, which have been methodologically and empirically discussed. Several parameters have been used for both techniques. Under specific choices for these parameters, the two methods lead to quite similar results. Both methods have strong statistical foundations; the interpretation of the empirical results rests however on a proper understanding of the meaning of the selected parameters. Autocorrelation seems to take the local spatial structure into better consideration as it enables the length of the black zones to vary locally.

The choice of these two methods is based on the knowledge of the accidents location. Their utilization requires quite accurate positioning of each accident. Today, this is unfortunately not always the case in practice as many countries still have quite inaccurate positioning systems. However, a wider use of satellite (*gps*, *Galileo*) is expected and will provide more accurate accident registrations. This is encouraging for the future.

This paper is limited to one data set: a 59 km long numbered Belgian road. Further analyses should include the road network as a whole, including crossings between roads where contiguity takes on another meaning. Moreover, in the studied data set, the accidents have both a statistical and a geographical distribution. It is our intention to let the geographical distribution vary given the statistical distribution and see how far it influences the autocorrelation measures. A great number of permutations are possible; it is important to show and to understand their effect on the proposed indices.

Finally, we would point out that this paper is limited to one aspect of descriptive exploratory data analysis (ESDA). However, the final objective is to understand the spatial occurrences of road accidents and their concentration in black zones. The next step will be the construction of an explanatory model to relate the location of black zones to some explanatory variables. For instance, do black zones just reflect high traffic volumes? Or is it possible to identify other causal relations, such as physical features of roads or the characteristics of the immediate environment of roads ? Examining such a modeling including spatial autocorrelation is a promising perspective.

Acknowledgements

The authors are indebted to the Federal Office for Scientific, Technical and Cultural Affairs (OSTC, contract MD/10/042) and to the Belgian National Fund for Scientific Research (FNRS) for financing parts of this research.

References

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.

Anselin, L., 1995. Local indicators of spatial association-LISA. Geogr. Anal. 27 (2), 93–115.

Bailey, T., Gatrell, A., 1995. Interactive Spatial Data Analysis. Longman, Harlow, Essex, England, pp. 413.

Black, W.R., 1991. Highway accidents: a spatial and temporal analysis. Transport. Res. Rec. 1318, 75–82.

Black, W.R., 1992. Network autocorrelation in transport network and flow systems. Geogr. Anal. 24, 207–222.

Black, W.R., Thomas, I., 1998. Accidents on Belgium's motorways: a network autocorrelation analysis. J. Transport Geogr. 6 (1), 23–31.

Boots, B., Tiefelsdorf, M., 1995. The exact distribution of Moran's *I*. Environ. Plann. 27A (6), 985–999.

Boots, B., Tiefelsdorf, M., 1997. A note on the extremities of local Moran's *I*(i)s and their impact on global Moran's *I*. Geogr. Anal. 29 (3), 248–257.

Cliff, A.D., Ord, J.K., 1973. Spatial Autocorrelation. Pion, London.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes. Models and Applications. Pion, London.

Cox, D.R., 1995. The relation between theory and application in statistics (with discussion). Test 4, 207–261.

Deacon, J.A., Zeeger, C.V., Deen, R.C., 1975. Identification of hazardous rural highway locations. Transport. Res. Rec. 543, 16–33.

Dempster, A.P., 1971. An overview of multivariate data analysis. J. Multivariate Anal. 1, 316–346.

Elvik, R., 1988. Some difficulties in defining populations of "entities" for estimating the expected number of accidents. Accid. Anal. Prev. 20, 261–275.

Flahaut, B., 1999. Influence de l'aménagement du territoire sur la sécurité routière durable. Analyse de la situation belge. Concentration spatiale des accidents de la route: délimitation des zones noires, Unpublished Research Rapport, Louvain-la-Neuve.

Foterhingham A., Brunsdon, C., Charlton M., 2000. Quantitative Geography. Perspectives on Spatial Data Analysis. Sage, London, p. 270.

Gasser, T., Kneip, A., Kolher, W., 1991. A flexible and fast method for automatic smoothing. J. Am. Stat. Assoc. 88, 643–652.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24 (3), 189–206.

Griffith, D.A., 1987. Spatial Autocorrelation: A Primer. Association of American Geographers, Resource Publications in Geography, Washington DC.

Haggett, P., Cliff, A.D., Frey, A., 1977. Locational Analysis in Human Geography. Arnold, London.

Haining, R., 1990. Spatial Data Analysis in the Social and Environmental Sciences. Cambridge University press, Cambridge.

Hauer, E., 1996. Identification of sites with promise. Transportation Research Record 1542. 75th Annual Meeting, Washington DC, pp. 54–60.

Joly, M.-F., Bourbeau, R., Bergeron, J., Messier, S., 1992. Analytical approach to the identification of hazardous road locations: a review of the literature. Centre de recherche sur les transports, Université de Montréal.

Levine N., 2000. CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations, vol. 1.1. Ned Levine & Associates/National Institute of Justice, Annandale, VA/Washington, DC.

Moran, P., 1948. The interpretation of statistical maps. J. R. Stat. Soc. 10b, 243–251.

Mouchart, M., San Martin, E., 2002. Specification and identification issues in models involving a latent hierarchical structure. J. Stat. Plann. Inference, unpublished data.

Nguyen, T.N., 1991. Identification of Accident Blackspot Locations, An Overview. VIC Roads/Safety Division, Research and Development Department, Australia.

Odland, J., 1988. Spatial Autocorrelation, vol. 9. Scientific Geography Series, Sage, Newbury Park.

Okamoto, H., Koshi, M., 1989. A method to cope with the random errors of observed accident rates in regression analysis. Accid. Anal. Prev. 21, 317–332.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: distributional issues and applications. Geogr. Anal. 27, 286–306.

Silcock, D.T., Smyth, A.W., 1985. Methods of Identifying Accidents Blackspots. Transport Operations Research Group, Department of Civil Engineering, University Of Newcastle Upon Tyne.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

Stern, E., Zehavi, Y., 1990. Road safety and hot weather: a study in applied transport geography. Trans. Inst. Br. Geogr. 15, 102–111.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. Accid. Anal. Prev. 28, 251–264.

Tiefelsdorf, M., 2000. Modelling Spatial Processes. Lecture Notes in Earth Sciences, vol. 87, Springer, Berlin.

Vandersmissen, M.H., Pouliot, M., Morin, D.R., 1996. Comment estimer l'insécurité d'un site d'accident: état de la question. Recherche Transports Sécurité 51, 49–60.