Spatial clustering of traffic accidents using distances along the network

K. Aerts\textsuperscript{a,b}, C. Lathuy\textsuperscript{a}, T. Steenberghen\textsuperscript{a} and I. Thomas\textsuperscript{b}

\textsuperscript{a} Spatial Applications Division Leuven (SADL), K.U.Leuven R&D, Celestijnenlaan 200E, BE-3001 Leuven, Belgium, koen.aerts@sadl.kuleuven.be

\textsuperscript{b} Department of Geography, Université Catholique de Louvain, Place Louis Pasteur 3, BE-1348 Louvain-La-Neuve, Belgium

Abstract

Many existing geostatistical methods for defining concentrations, such as the kernel and the local spatial autocorrelation method, take into account the Euclidian distance between the observations. However, since traffic accidents are typically located along a road network, it is assumed that the use of network distances instead of Euclidean distances could improve the results of these clustering techniques. A new method is proposed here, taking into account the distances along the road network. This methodology is applied to and tested on the historic city of Brussels (Belgium).

Most existing accident clustering methods start from the actual distribution of the accidents over the road network. However, to guarantee the independence between the results and the actual accident locations, a close distribution of points of measurement is randomly distributed over the road network. A cost matrix is calculated, containing the distances between each individual accident and the points of measurement. Next, a dangerousness index is calculated for each point of measurement, taking into account the weight of the accidents. Nearby accidents have a greater influence on the dangerousness index than accidents further away. Finally, a network Voronoi diagram is used to distribute the dangerousness index over the road network.

The proposed methodology allows computing spatial concentrations of road accidents based on distances along the road network, including distances between accidents on different, but intersecting roads. By using randomly distributed points of measurement instead of accident locations, the independency between the accident locations and the resulting dangerousness segments is guaranteed, allowing for easy comparison of results in different analyses (for example in different time periods).

Introduction

Traffic accidents tend to be concentrated in clusters in geographic (e.g. Yamada & Thill, 2004); accidents are more likely to occur at dangerous locations. Concentrations of traffic accident occurrences suggest spatial dependence between accidents and common causes. These “black zones”, or zones with significantly high accident numbers, can be detected by several geostatistical techniques. This identification and analysis of locations producing more accident than the average, is hence an important step in traffic accident prevention.

Previous research focused on the use of the kernel and the local spatial autocorrelation (LISA) method to identify black zones. Steenberghen et al (2004) made a comparison between these two methods for a dense city network. They pointed out that the kernel method gives better results on dense road networks, while the local spatial autocorrelation method performs better on networks where the maze is large. However, the use of a kernel based on Euclidean distances can lead to overestimations, especially in dense city networks.
Traffic accidents are the consequence of traffic movements, which take place uniquely along a road network. The existing methods for detecting dangerous locations, such as the kernel or the local spatial autocorrelation method (Flahaut et al., 2003; Steenberghen et al., 2004), are based on Euclidean distances, and thus disregard the specific nature of traffic movement. However Yamada & Thill (2004) found a significant chance of over-detecting clustered patterns in planar K-function analysis and proposes a network K-function to resolve this problem. Lu & Chen (2006) arrive to the same conclusion when analyzing point patterns of vehicle thefts. The K-function analysis only examines whether a given point distribution is different from a random distribution, but it does not reveal the location of clusters or black zones in the distribution.

The existing spatial clustering techniques have disadvantages when applied to a road network instead of a single road segment. The kernel method results in a grid over the total study area with a dangerousness measure for each grid cell, even if there is no road in the grid cell. This gives the false impression of large extents for the black zones. The local spatial autocorrelations method requires the aggregation of accidents in Basic Statistical Units (BSU’s) (Flahaut and Thomas, 2002). For single road segments, the BSU’s are usually divided according to the kilometer markers. For road networks this segmentation is not so straightforward, how to continue at intersections?

In this study, a new methodology for detecting dangerous location, based on distances along the road network, is proposed, which improves the disadvantages of the existing techniques. The paper is structured as follows: Section 2 proposes network distance weighted clustering method. Section 3 describes the study area under investigation and the input data. Section 4 gives the empirical results, and finally section 5 gives a conclusion, some final comments and some issues for further research.

**Network distance weighted clustering of accidents**

The goal of accident clustering techniques is to find dangerous locations or black zones (road segments, intersections), characterized by a higher number of traffic accidents then expected from a pure random distribution of accidents over the road network. The desired result is a dangerousness map which indicates the dangerousness of a road segment or intersection by means of a dangerousness measure.

**Points of measurement**

Existing clustering techniques start from the observed accident locations, resulting in a dangerousness measure for these accident locations. However, this dependency between the input accident locations and the results impedes the comparison between different analyses (for example for different time periods).

To guarantee independency, a dangerousness measure will be calculated for randomly distributed locations on the network, further referred to as points of measurement, instead of for the actual accident locations. The points of measurement are composed of the road intersections and a random distribution of points over the intermediate road segments. To ensure the complete coverage of the road network, the maximum distance between two points of measurement needs to be smaller than two times the influence distance.

Next, a network Voronoi diagram (Okabe, 1992) is created from the points of measurement. Each infinitesimal road segment is assigned the value of its closest point of measurement. The result is a dangerousness map which covers the whole road network.
Network distances

Proximity is used as a basis for most spatial clustering techniques (Steenberghen et al., 2004). A cluster can be defined as a localized excess incidence rate that is unusual in that there is more of some variable than might be expected. This implies that observations close to each other in space are more correlated to each other than far spaced observations. Often observations are presumed to be only correlated until certain a maximum distance, here called the influence distance (D).

Distance weighing functions are used to model this phenomenon. Hence the distance between observations needs to be known. Traditionally, the Euclidean distance between accidents is used in spatial clustering techniques. Figure 1(a) shows an example road network with 10 accidents and 2 points of measurement. In the remaining figures the influence range, or the road segments within the influence distance of a point of measurement, is indicated.

Figure 1(b) shows the influence range using Euclidean distances for the two points of measurement. A circle with radius D is drawn around each point of measurement. The parts of the road network inside these circles are within the influence range. The Euclidean influence range for A measures 222 m, for B 240 m. Figure 1(c) shows the network influence range for A (184 m) and Figure 1(d) for point of measurement B (217 m). Notice that the influence range is larger (15%) when using Euclidean distances.

The Euclidean distance method tends thus to overestimate the influence ranges, especially in dense network neighborhoods. This has consequences for the accident clustering, for example: accidents 4 and 10 fall between the Euclidean influence range of both points of measurement, while in case of network distances they are only accounted at one point of measurement. A clustering technique which uses distances along the road network instead of Euclidean distances will thus have more accurate influences ranges and give more realistic results.

Dangerousness Index

The dangerousness index is a measure for the dangerousness of a point of measurement and is calculated from the weighted number of accidents that occurred in a certain distance along the road network from the point of measurement:

$$Dl_i = \sum_{j=1}^{n} w_j I_D (d_{ij}) \quad \text{for} \quad i = 1 \text{K} \ m$$

where $Dl_i$ is the dangerousness index at point of measurement $p_i$, $n$ the number of accidents, $m$ the number of points of measurements, $w_j$ is the weight of accident $a_j$ at $p_i$, and $I_D (d_{ij})$ is a function to define whether the accident lies within the influence distance of $p_i$:

$$I_D (d_{ij}) = \begin{cases} 1 \quad \text{if} \quad d_{ij} \leq D \\
0 \quad \text{otherwise} \end{cases}$$
Figure 1: Comparison between influence ranges using Euclidean distances versus network distances. (a) Example road network with accidents and points of measurement; (b) Influence range for A and B in Euclidean distances; (c) Influence range for A in network distances; (d) Influence range for B in network distances.

The weight for each individual accident on a particular point of measurement can be dependent on several parameters, such as distance from the accident location to point of measurement, time, size of the accident, number of casualties, etc. As the focus in this paper is on the distance, the accidents are only weighted in terms of distance, other parameters are not investigated.

Several distance weighing functions are tested (see Figure 2):

1. **Distance Band**, each accident within the distance range has the same weight:
   \[ w_{ij} = 1 \]  \hspace{1cm} (3)

2. **Inverse Distance**, the weight of an accident is dependent on the inverse of the distance until the point of measurement:
   \[ w_{ij} = \frac{1}{d^{*}} \]  \hspace{1cm} (4)

3. **Linear Decrease**, the weight decreases linear within the distance range
   \[ w_{ij} = \frac{D-d^{*}}{D} \]  \hspace{1cm} (5)
Figure 2: Distance weighing functions, influence distance (D) = 50 m.

Figure 3 shows an example road segment with three accident locations with together 11 accidents and the points of measurement on the accident locations. The accidents are located at 26 m from each other, so within the distance range of 50 m, the outer accident locations have no influence on each other. From the three tested weighing functions, the Linear Decrease method gives the most homogeneous result. For the outer points of measurements, the dangerousness index is similar for each method. For the intermediate point of measurement, however, great differences occur. The Distance Band method seems to overrate the dangerousness index, while the Inverse Distance method makes an underestimation.

Study area and data

The dense road network of the historic city of Brussels, the capital city of Belgium, was chosen as the study area. As many other cities, the urban area of Brussels extends over its administrative boundaries. In this paper the central area of the 19 communities of the Brussels Capital Region (BCR) are investigated. They correspond to the CBD and the first suburbs (Van der Haegen et al, 1996).

The study area consists of 162 km², about 1.018.029 inhabitants, 658.788 jobs (BISA, 2005) and 1600 km of public road. In Brussels, many jobs are still located in the city center (e.g. Riguëlle et al, 2005). Half of the working population lives in the BCR itself; the other half commutes between home in the large influence area and office, hence causing large traffic movements. The socio-economic characteristics of the inhabitants is quite structured: better-off people are located at the outskirts, especially in the eastern-southeastern part (Goffette-Nagot et al, 2000)
Road network
The road network of the BCR consists of approximately 1600 km of public road, divided in about 12,000 road segments by 7,616 road intersections. The average road segment length between two intersections measures $138.5 \pm 123.7$ m.

The road network is modeled as a planar graph because neither relative height information, nor information about driving directions is currently available. This means that road intersections occur wherever two or more roads cross, even if there is a bridge or a tunnel, and every road segment can be accessed in both driving directions. A topological model with relative height information and driving directions could improve the overall methodology, however, at the moment no such model was available.

Accident data
This study is limited to accidents with casualties, accidents with only material damage are not considered by a lack of sufficient data. The accident data are registered by the police, and centralized at the National Institute of Statistics. The data covers a three year period, from 1997 until 1999, which is considered as the shortest period to derive statistical significant results. During this period, 8,762 accidents with casualties occurred. The elaborate accident dataset contains detailed information for each individual accident: temporal information (date and time), environmental conditions (weather, light conditions), road conditions (road surface, road infrastructure), and nature of wounds (lightly, heavily, fatally wounded).

In the urban agglomeration in Belgium, the accident locations are determined by corresponding address, noted on the official police reports. For numbered roads, the accidents are located by stone markers every 100 meters along the road. The accidents are located on the road network by dynamic segmentation and address matching techniques. Because of incomplete addresses, only 5144 (59%) of the accidents could be accurately located on the road network. Only the located accidents are restrained for further analysis.

The number of unique accident locations (1 m accuracy) gives already a first indication of spatial clustering of traffic accidents: 5,144 accidents occurred at 3250 distinguishable accident locations. Figure 5 shows the distribution of the accidents over the unique accident locations in the CDB of Brussels. Larger circles stand for locations with more accidents.
Results

The network weighted cluster method is used to compute spatial black zones for the Brussels Capital Region. The influence distance for the BCR was set to 50 m, resulting in an influence range (the actual length of road segments in a certain influence distance from a point) of 100 m on a single road segment (in Belgium the accuracy of accident location is 100 m on numbered roads). On road crossings, the influence range may be higher.

The network distance from an accident location to the points of measurement is calculated with the Network Analyst Extension from ArcGIS 9.1 and stored in an origin-destination cost matrix (OD cost matrix). The calculation of this matrix is a time consuming operation (multiple hours to days, depending on the size of the dataset). To speed up the analysis, a cut-off distance can be set at the influence distance (D).
A close distribution of points of measurement is placed randomly over the road network. Each road intersection is included (N = 7.616) and the intermediate road sections are populated with random points of measurement at an average distance of 25 m (N = 66.987), resulting in a total number of 74.603 points of measurements. A network Voronoi diagram is calculated from the points of measurements.

The distances between the accidents and the points of measurement are calculated and stored in a cost matrix from which the weights are calculated. The dangerousness indices are determined for each point of measurement and assigned to the corresponding road segment in the Voronoi diagram.

The resulting dangerousness map for the CBD is depicted in Figure 6. Darker road segments are more dangerous then lighter ones. The black zones are mainly situated at the intersections of the major roads and the inner ring road (Pentagon, see Figure 4) and on a major road running from North to South inside the inner ring. The smaller, local roads tend to be less dangerous.
Conclusions and further research

This paper presented a new methodology to identify black zones of traffic accidents on a road network using distances along the network. This method has some major improvements over already existing spatial clustering techniques. The use of network distances instead of Euclidean distances gives a more realistic measure of spatial correlation for events that are the result of movements uniquely along a network such as traffic accidents. The use of points of measurement ensures the independency between the actual accident locations and the output, allowing for easy comparison between analyses, and avoids an aggregation of accidents.

The method was tested on the dense road network of the historic city of Brussels. In the next phase, the traffic accidents of the fast growing city of QingDao, China, will be investigated. A comparison between the results of the two cities will be carried out in order to identify causal factors for traffic accidents at specific locations.

Current research also focuses on estimating the statistical significance of the dangerousness indices by Monte Carlo simulations. Dependent on the influence range for a point of measurement, the expected dangerousness index may be higher. The Monte Carlo simulation takes into account this variation in expected dangerousness index, so a more accurate result may be obtained. Next to variations in influence range, also variations in traffic flow, road category, etc. can be incorporated in the Monte Carlo simulations.
Acknowledgements

The research presented in this paper is part of a broader research program on “Time-Space patterns of traffic accidents” (2005-2007), funded by the Chinese Ministry of Science and Technology (MOST) and the Belgian Federal Science Policy Office (SPO).

References


