SCIENTIFIC SUPPORT PLAN FOR A SUSTAINABLE DEVELOPMENT POLICY

SPSD II

INNOVATIVE SPATIAL ANALYSIS TECHNIQUES FOR TRAFFIC SAFETY

T. STEENBERGHEN, I. THOMAS, G. WETS

GENERAL ISSUES

AGRO-FOOD

ENERGY

PART 1

SUSTAINABLE PRODUCTION AND CONSUMPTION PATTERNS → TRANSPORT

SCIENTIFIC SUPPORT PLAN FOR A SUSTAINABLE DEVELOPMENT POLICY
(SPSD II)

Part 1:
Sustainable production and consumption patterns

FINAL REPORT

Innovative Spatial Analysis Techniques for Traffic Safety

CP/34

Thérèse Steenberghen – KULeuven/R&D Division SADL (coordinator)
Isabelle Thomas – UCL/Unité de Géographie
Geert Wets – Universiteit Hasselt (vroeger Limburgs Universitair Centrum) /Vakgroep Verkeerskunde & Beleidsinformatica)

Research contracts n° CP/01/341, CP/10/342, CP/07/343

January 2005

BELGIAN SCIENCE POLICY

SADL

universiteit hasselt

UCL Unité de Géographie

BELGIAN SCIENCE POLICY

D/2006/1191/5
Published in 2006 by the Belgian Science Policy
Rue de la Science 8
Wetenschapsstraat 8
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11 – Fax: +32 (0)2 230 59 12
http://www.belspo.be

Contact person:
Mr Georges Jamart
Secretariat: +32 (0)2 238 37 61

Neither the Belgian Science Policy nor any person acting on behalf of the Belgian Science Policy is responsible for the use which might be made of the following information. The authors are responsible for the content.

Table of contents

Part I: General introduction 5
1 Context........ 5
2 Objectives 5
3 Research framework 6
4 About this report 6

Part II: Research parts 7
1 SADL KUL R&D 7
1.1. Refinement of research goals 7
1.2. Building the accident database 7
1.2.1 Locating accidents: methods........ 7
1.2.2 Locating accidents: results........ 8
1.2.3 Conclusion 9
1.3. Building a land use map 9
1.3.1 Introduction: from traffic indicators to a land use map 9
1.3.2 Pre-processing of the imagery 11
1.3.3 Classification to land cover and land use........ 13
1.3.4 Post-processing and results........ 16
1.3.5 Conclusion 17
1.4. Delimitation of black zones 17
1.4.1 Introduction 17
1.4.2 Clustering of accidents: Ripley's K-function 18
1.4.3 Kernel density maps 19
1.4.4 From densities to probabilities........ 22
1.4.5 Discussion of traffic safety in Brussels........ 28
1.4.6 Conclusion 30
1.5. Explanatory model for accident frequency 31
1.5.1 Introduction 31
1.5.2 Data description........ 31
1.5.3 Regression methods 32
1.5.4 Model results 34
1.5.5 Conclusion 39
2 LUC 40
2.1. Research goals........ 40
2.2. Quality assessment of the Belgian traffic accident data........ 40
2.2.1 The characteristics of the data 40
2.2.2 Non-applicable fields........ 41
2.2.3 Missing and inaccurate values 41
2.2.4 Inconsistency 42
2.2.5 Validity 43
2.2.6 Reliability 43
2.2.7 Conclusion 44
2.3. Relevance assessment of the traffic accident variables........ 45
2.3.1 Traffic accident data and association rules 45
2.3.2 Empirical study........ 45
2.3.3 Results 47
2.3.4 Conclusions 52

2.4. Model based clustering on traffic accident data 53
2.4.1 Purpose 53
2.4.2 Cluster techniques and applicability 53
2.4.3 The traffic accident data 54
2.4.4 Empirical results for the Walloon Brabant region 55
2.4.5 Empirical results for the Brussels Capital region 57
2.4.6 Conclusions 61
2.5. Understanding accidents in black zones using frequent item sets 64
2.5.1 The technique of frequent item sets 64
2.5.2 The data set 65
2.5.3 Conclusions 66
3 UCL 68
3.1. The spatio-temporal distribution of road traffic accidents in a periurban environment. The case of Brussels 68
3.2. Spatial nested scales for road accidents in the periphery of Brussels 70
4 Overall conclusion 70
5 Bibliography 73
5.1. SADL KUL R&D 73
5.2. LUC 75
5.3. UCL 77

Part III: Appendix 78

Part I: General introduction

1 Context

During the last decades, mobility of goods and persons –particularly the share of road transport– has increased significantly. This was accompanied by a dramatic number of traffic accident victims (table 1). Compared to the European average, Belgium is confronted with a poor record in terms of traffic safety (table 2).

Table 1. Evolution of road transport an of traffic (un)safety.

Year	Victims (deceased + wounded)	Vehicle km (billion km)
1970	107.777	Approx 2.500.000
1980	84.700	3.753.745
1990	88.160	4.594.058
2000	69.431	5.735.034

Source: BIVV 2001

Table 2. Number of fatalities per 100.000 inhabitants in Belgium and Europe.

	1997	**1998**	**1999**
Belgium	13,4	14,7	13,7
European Union (15 countries)	11,6	11,3	11,1

Source: IRTAD - International Road Traffic and Accident Database (OESO)

This (un)safety is frequently blamed on poor spatial planning. In a previous research project (PADDI, "Impact of spatial planning on traffic safety", 1998-2000), the impact of the spatial characteristics on traffic safety was examined. Much effort was spent to develop adequate tools for proper location of accidents on both numbered roads (highways and major roads), and secondary roads. This research resulted in new methods for the identification of statistically sound black zones (Flahaut 2001, Flahaut et. Al. 2002, Flahaut and Thomas, 2002), and in a number of case studies where relations were identified between type of urbanization and traffic safety (Steenberghen and Dufays, 1999, 2000, 2002, Steenberghen et al. 2003).

2 Objectives

The original objective of this research was to improve the explanatory model for traffic safety, in order to clarify the interactions between safety factors. The previous research was based on GIS and spatial statistics. In order to reach the objective, this project explored the potential of new data sources and analysis techniques. Three innovative approaches are used, carried out by the three research partners (**KUL**, **LUC** and **UCL**). But very soon, it became clear that a single explanatory model was difficult to develop, even with new techniques. It is even questionable whether one single traffic safety model exists.

Three innovative approaches were explored:

High resolution satellite imagery**: (KUL)**: This part of the project focused on the identification of land use, infrastructure and traffic characteristics by means of remote sensing techniques applied on high resolution satellite imagery. Due to technical difficulties this research goal was not fully carried out and research goals shifted towards spatial statistical exploration and modelling of the accident pattern.

***Spatial Data Mining* (LUC)**: determination of significant combinations of causal factors (LUC). The accident data used for this research consists of all the road accidents in Belgium for a period of 9 years (1991-1999), with approximately 100 attributes for each accident. Knowledge Discovery and Data Mining are explored here as tools to detect structures, patterns and relations in this large data set.

Multi Level Analysis**: (UCL)**: Explanatory factors of traffic (un)safety appear to be very scale-dependent. Rather than developing an explanatory model for different scales, recent developments through multilevel frameworks provide the opportunity of examining interactions at different levels and of integrating interactions between scale levels.

3 Research framework

Every research team will elaborate an own analysis technique and conformity will be brought by successive iterations. Some results of each team will be used as input for following steps in the research of all other teams. Fundamental data for each research are the accident data: the actual database developed in the former research project containing this data will be actualized so that the accident data covers a period of 9 years from 1999 to 1999, including accident locations. Based on these accident locations, clustering techniques can identify the existence and location of black zones which can then serve as additional input in the spatial data mining process. Also, the multilevel modelling can make use of these spatial accident clusters.

Although there are several ways for the three research partners to interact with each other and exchange provisional results, there is a limitation because different study areas are being used by the partners. KUL and UCL focus on the Brussels Capital Region and on Walloon-Brabant respectively but LUC is able to test their data mining techniques on different regions including Brussels and Walloon Brabant. But the lack of one common study region still stays a disadvantage

4 About this report

Part two of this document reports the three researches. The research of the KUL and LUC is reported in several sections while that of the UCL is reported very brief as an abstract but the full report of UCL is found in the appendix as a paper.

.

Part II: Research parts

1 SADL KUL R&D

1.1. Refinement of research goals

During the two years of research we had to adjust the original research goals concerning the remote sensing. In our attempts to extract new explanatory indicators for traffic safety from Ikonos images we experienced several difficulties (cfr. 1.3.1.1 Evaluation of Ikonos' usefulness for traffic safety indicators) which forced us to abandon the goal of automatic extraction of traffic safety indicators from the Ikonos image. Therefore we reoriented the study goals towards spatial statistics: optimization of two-dimensional black zone definition and (spatial) modelling of accident frequency came to play an important role and methodological issues were given more attention. However this doesn't mean that the remote sensing objectives were completely forgotten: a new classification approach was designed to classify the very high resolution image into a land use map.

The results of four research topics are described in the following:

- Building the accident database
- Extracting land use from Ikonos imagery
- Determination of accident black zones
- Statistical modelling of accident frequency

The study area for all research parts is the Brussels Metropolitan Region while the accident database was built for Belgium.

1.2. Building the accident database

1.2.1 Locating accidents: methods

Raw accident data are provided by the National Institute of Statistics (NIS) in ASCII format. The raw ASCII files are imported in MS Access where a relational database is designed (Appendix A.1). The regional road administrations in Flanders and the Walloon Region check and correct the location attributes (since the mid 1990's). The location information in the database is either a combination of a street name and house number, or a road number and hectometre number, or, in case of a crossroad accident, two roads. The location of accidents can be done rather easily in GIS with different techniques. We used *address matching* and *dynamic segmentation* techniques developed in the previous research project (Steenberghen et al. 2000).

Accident location based on road and hectometre number

Dynamic segmentation is a GIS technique which uses a digital road network where every road has its identification number (road number) and a measuring system (hectometres). Based on this route system accidents can be located in GIS as a route event.

Accident location based on street name and house number
Address matching is another GIS technique that locates accidents based on the street name and house number. This technique was only applied in the Brussels region because of the availability of an accurate base map containing streets and buildings. This information allowed us to locate many accidents with precision.

Finally, crossroad accidents can be located at the intersection of the two streets. We finished this part of the research with some post processing steps and checks of the accident locations.
An example: Given that an accident has been located in the GIS system in municipality X, then we can check if this is the same municipality X that is mentioned in the accident database. Another kind of post processing is the grouping of accidents near crossroads on the crossroads themselves and the alignment of accidents on the street axes.

1.2.2 Locating accidents: results

In a previous research project, accidents from the period 1991-96 were already located and now they were updated with another three years so that we have accidents (with injuries or worse) for the 9-year period 1991-1999. During this period there occurred in Flanders, Walloon and Brussels respectively 307.161, 138.285 en 27.480 accidents. For the first two regions, only the accidents on numbered roads (representing 57%) are located while for Brussels all accidents had to be located.
Not all accidents could be located due to inconsistencies in the database. For Flanders and Walloon (only numbered roads) the main reason is the absence of a value for the hectometre or a value of 0 (0 = default value). Nevertheless, about 80% of all accidents on numbered roads could be located but the percentage has increased over time and it is higher for crossroad accidents then for road segment accidents (Figure 1-1). The increasing trend is due to the intensified corrections on the database that have been carried out by the regional administrations and crossroad accidents are always easier to locate than none-crossroad accidents (no need of a hectometre). In Brussels there were many problems with street names and missing house numbers which explains the rather moderate 60% of localized accidents, representing 27.000 located accidents in the past 9 years.

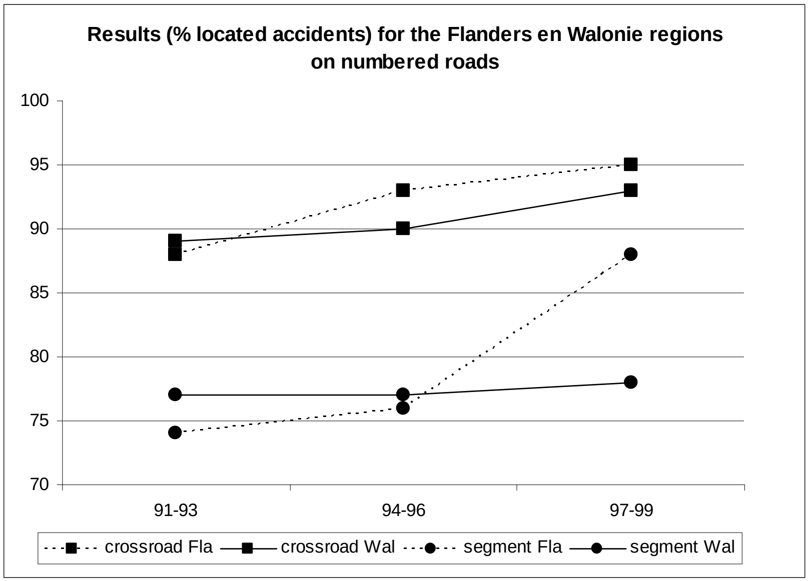


Figure 1-1: accidents location results for Flanders and Walloon region

1.2.3 Conclusion

Both the attribute and spatial database were successfully constructed for the three regions in Belgium: Flanders and Walloon regions and Brussels. There are however some remarks to be made.

One major drawback is that many attributes are not available or have inconsistent values for many accident records but this is due to misspecifications and input errors in the source data which is delivered to us by the NIS[1]. One of the important consequences of these errors is the frequent occurrence of "0" values for the hectometre which is filled in by the NIS operator when the hectometre value is not known (it should be a null value instead!). This is the major reason that not all accidents on numbered roads could be located (about 80%) but this problem is gradually being corrected by the regional authorities. The situation in Brussels is worse, only 60% of all accidents could be located because, in addition to accidents on numbered roads, all other accidents were located and this was done by means of address matching. Misspelled street names and lacking house numbers are the main causes of the rather low share of located accidents. Finally it must be said that the NIS accident statistics are far from error free and that there is a serious underrepresentation.

1.3. Building a land use map

1.3.1 Introduction: from traffic indicators to a land use map

1.3.1.1 Evaluation of Ikonos' usefulness for traffic safety indicators

There are several obstacles when using Ikonos imagery for (automatic) collection of traffic related information. The initial main goals of this research part (SADL) were highly focused on extracting (new) traffic and road environment indicators from satellite imagery. Indicators could be: density of moving vehicles, parked vehicle densities, visibility on crossroads, streetside obstacles, etc. However, some constraints and technical problems show up making it practically impossible to achieve this main research goal.

Traffic related indicators

Estimating the traffic volume based on a satellite image was one of the original goals in the study. The idea is to derive the vehicle density from the image and then linking this density to existing traffic volume data, establish a relationship between traffic density on the image and actual traffic volume. We will point out the difficulties one can expect to deal with, especially in identifying the vehicles – automatically – on the image.

A very important constraint is the fixed timing of the image capture for our study region, which is always around 11.00 am, every three days (approximately). This is due to the satellite's orbit characteristic and cannot be changed. This is already a severe limitation for the calculation of all traffic related indicators as traffic is a highly dynamic phenomenon which must be monitored during the course of a day to understand it properly. Thus, the static nature of the imagery, in particular the time of image capture, is not compatible with the highly dynamic road traffic.

At the technical level, automatic detection of vehicles also poses some problems. A standard multispectral classification into vehicles and road tarmac is all too simple. The road tarmac doesn't have the same spectral signature for all roads and overlaps with certain dark-colour vehicles. And even if this spectral problem could be solved there is another problem with this technique. As the multispectral bands are only 4x4 meters resolution (which is not appropriate to identify vehicles), fusion with the panchromatic band (1x1 meters) is necessary. But due to a

[1] National Institute for Statistics

slight time delay between the capture of the pan and multispectral images, the vehicles are not anymore at the same location (on a freeway they move relatively fast) and thus on the fused image, in the neighbourhood of the vehicles, the fused pixels are spectrally invalid[2]. We found a fairly easy technique in the literature, which only needs the panchromatic band but requires several images of the same area (Dial et al. 2001. The problem with this method is not of spectral but of geometric nature because the images need to be perfectly geo-registered. As will be shown in the next section, such a high geometric precision is very difficult to achieve with Ikonos imagery. Surely, there must exist highly intelligent techniques for object recognition and extraction which may give good results without imposing strict spectral and/or geometric assumptions but these techniques are part of the complex discipline of computer vision and require appropriate training and knowledge. So, a first conclusion is that the automatic detection of vehicles is technically a very complicated matter and in our case would be of low relevancy due to the fixed time of capture. Moreover, many of the methods assume a pure road, free of spatial clutter (e.g. a freeway in open area) but the reality is in most of the cases different (presence of roadside bushes, buildings, light poles, bridges preventing a full view of the road). These effects are increasingly disturbing in dense urban environments.

Road environment indicators

Because the automatic detection of moving vehicles seems not feasible, other potential indicators were evaluated if they could be extracted from the image. Here, again the conclusion is that relevant indicators are extremely difficult to extract with sufficient precision from the image. Extracting parked vehicles seems more difficult than extracting moving vehicles because the view is disturbed by buildings, shadows from buildings and trees. Another potential indicator is "visibility on crossroad", to be measured by the offset of buildings to the street side but in order to measure this in a correct way, one needs an orthogonal view. The fact that our image is an oblique image (as most images) makes that building footprints are hidden behind the buildings themselves. Besides, in many cases the border between road tarmac, footpath and building is diffuse.

All these difficulties concerned with the identification and automated extraction of road traffic and road environment indicators forced us to abandon the initial goals of the remote sensing part. But meanwhile we had built a considerable amount of knowledge and experience with very high resolution imagery that we had to consolidate. Therefore we applied this knowledge to build a land use map out of the Ikonos image.

1.3.1.2 Very high resolution images and traditional classification methods

We will discuss two issues to deal with when classifying a very high resolution image to a land use map, demonstrating that the standard multispectral per-pixel classification methods are not suitable in this context.

A major problem with high resolution images is the fact that too much detail complicates classification. A street is no longer uniquely characterized in the spectral domain as it is in lower resolution imagery. On an Ikonos image, a street is an amalgam of sub-objects such as the pavement surface, vehicles, pedestrian crossings, light poles, etc., each having distinct spectral characteristics. The same is true for other land cover objects. Such high detail interferes with the relationship between land cover objects and the spectral response of their pixels, unlike the better relation in Landsat imagery. This will lead to a huge salt'n pepper effect when only applying a

[2] In the future, when the resolution will be higher this time delay effect may be used to estimate the speed of the vehicle based on the difference in position between the vehicle on the pan and on the multispectral image.

per-pixel classification on high resolution images. Thus, high resolution imagery urges for an expansion of the classic per-pixel image analysis[3] and draws attention for the spatial context. Specific and rather expensive software exists (e.g. *eCognition*, Definiens Imaging) which allows complex object-oriented classification tasks. Another more simple approach is possible when there is some ancillary information of the study area present, such as a vector layer with land parcels. Land parcels are meaningful objects which can serve as the basis for an image classification. The so-called per-parcel approach (Aplin et al. 1999) was applied on the Ikonos image of Brussels and includes a classification of the land parcels.
A second problem concerns the identification of the *land use*. Whereas *land cover* is related to the physical characteristics of the earth's surface and can rather easily be classified (see above), land use is related to the socio-economic occupation of the earth's surface and its classification is more problematic. Land use is first of all defined in terms of function but it can be to some extend inferred from its form (Barnsley & Barr, 1996). Spatial patterns and relations (between land cover objects) must be taken into account to infer the land use. We propose an extension of the per-parcel approach presented by Aplin by classifying land use through a statistical analysis of several land cover parameters. In this way 5 land use classes were determined: low density residential, medium density residential, high density built area, offices and/or industrial area and green area.

1.3.2 Pre-processing of the imagery

1.3.2.1 Geometric pre-processing: rectification

Ikonos imagery exists at different geometric quality levels and our image was of moderate geometric quality. When overlaid with a high-precision vectormap several displacements could be observed and these errors must be corrected before all other processing can take place. The vectormap, a very accurate base map (URBIS v2, C.I.B.G.[4]) is based on orthophotos and has a very high accuracy[5]. Our Ikonos image is one of the "Geo" type, the cheapest Ikonos product, and has an absolute planimetric accuracy of 50 metres (CE90), terrain displacements (due to terrain height differences) not included (Appendix B.1). More precise Ikonos products exist but they are 5-10 more expensive than Geo. For this reason we thought it was worthwhile to invest considerable time in an orthorectification method to improve the image at the geometric level. First, we applied a standard polynomial rectification but we noticed that the accuracy was still far from a sufficient level.

In the area covered by the image the maximum elevation difference is more than 100 meters which leads to a planimetric displacement of up to 44 metres given the elevation of the satellite at the time of image capture. These important terrain displacements should be corrected. A so called "relief corrected affine transformation" has been successfully applied on several Ikonos images to correct errors due to elevation differences (Baltsavias, Pateraki, Zhang, 2001). With this technique one can easily rectify an image including terrain displacement errors. First, knowing the satellite's position in space and every pixel's height we can reproject each pixel to a reference plane at a constant height and second, we apply polynomial transformation in order to produce the final image (Appendix B.2B.2).

[3] The standard per-pixel classification consists of classifying each pixel in a spectral class according to the pixel's value on the spectral bands (for Ikonos: red, green, blue and near infrared).
[4] Brussels Urbis Adm v2.0 (Brussels Urban Information System), C.I.B.G. Kunstlaan 20, Brussel
[5] Distinct objects like poles etc. have a positional error of less than 20 cm.

To check whether the above method works fine and to confirm that terrain height differences are an important error source in our image, we did two preliminary tests. A total of 14 GCP's (Ground Control Points) were selected on the image and the reference map. First, the image GCP's were orthorectified using the above method and an accuracy test revealed that after orthorectification the RMSE (Root Mean Square Error) dropped from 5.9 to 2.8 meters. Second, we did a standard polynomial rectification, based only on GCP's at approximately the same elevation level (80 meters). Next, we applied this rectification model once on the remaining set of checkpoints and once again on these checkpoints after reprojection in the reference plane at elevation 0. As expected, the RMSE for the original checkpoints gets higher with increasing elevation difference (Table 1-1). Such a trend is not present in the set of reprojected checkpoints indicating again that terrain height differences are an important source of geometric errors.

Table 1-1: Influence of relief on planimetric accuracy

Checkpoint's elevation (meter)	Checkpoint's elevation difference (meter)*	Without elevation correction (RMSE, meters)	With elevation correction (RMSE, meters)
97	17	1,44	4,91
50	30	16,94	3,87
20	60	33,279	5,83
All checkpoints	**All checkpoints**	**22,9968**	**5,195**

* The image was rectified with a first order polynomial using 5 ground control points, all located at about **80 meters** height.

For the study we used a DEM (grid spacing of 30 m) and collected precise coordinates and elevations for the ground control points located at the centroids of roundabouts, recognizable in both the reference map and the image (Appendix B.3). The algorithm was implemented using VBA ArcObjects in the ArcGIS environment.

Although the preliminary test confirmed that the method improved accuracy a lot, our final results were disappointing. The differences between the predicted and true point locations for all final GCPs are quite high with some extremes of up to 12,9 meters. Several orthorectifications were done with different sets of GCPs but none of these resulted in a marked increase in accuracy (Appendix B.4). As a matter of fact, we can get the same accuracy with a standard polynomial rectification (without correcting the relief displacements). With this information it is evident that there must be some other major source of error than the elevation differences, probably the satellite's tilting movements that the method does not take into account. Correcting these tilting movements requires specific software and knowledge that is beyond the scope of this research project.

As long as the image is not correctly rectified, there will be no overlap between image and reference map, hence this is an important loss of complementary information. The solution for a correct rectification appeared to be a *'rubber sheet'* transformation based on a large number of GCPs, in particular in the area where the orthorectification produced the largest errors. With this well established technique (standard in many image analysis tools) we were able to get a good overlap between image and reference map in the whole study area. An example of the required match between image and map is found in appendix B.5.

1.3.2.2 Spectral pre-processing: Data Fusion

Next an appropriate method for merging the multispectral bands with the high resolution panchromatic band was applied. Adaptive Image Fusion is a data fusion method using filter techniques that acts as a pre-segmentation on the image (Steinnocher, 1999). The goal is not only to sharpen the multispectral image but also to get the objects in the new image more homogeneous. The more the objects are homogenous the less salt'n pepper will show up in the classification. Adaptive Image Fusion also preserves the spectral characteristics of the original low resolution image to a high extent (Steinnocher, 1999). The fused image is sharpened as can be noticed on the edges of buildings but the buildings themselves are far more homogenous than on the original multispectral or on the PCA sharpened[6] image (Figure 1-2).

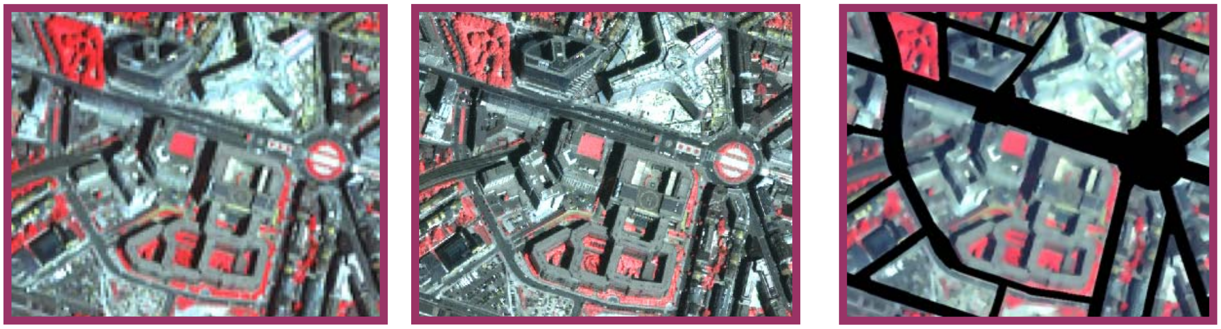


Figure 1-2: Data fusion: 4 x 4 m MS (left), 1 x 1 m pansharpened MS (middle), 1 x 1 m AIF MS (right)

1.3.3 Classification to land cover and land use

1.3.3.1 Per-pixel classification: land cover

A supervised per-pixel maximum likelihood classification was performed on the basis of several land cover class signatures that were derived by a preliminary unsupervised classification and by human interpretation. First an unsupervised classification was run on a representative subset of the city centre – which contained most of the land cover classes – resulting in about 30 classes. Some of these classes were merged with the aid of separability measures and visual interpretation. Then, from another subset, situated near the peri-urban transition zone some additional classes were selected to complete the existing classes. Finally the image was classified using 12 land cover classes: a shadow class, several "built"- classes, soil, water, vegetation, roofs, grass-arable land etc. Salt'n pepper was eliminated to some extent with a majority filter and Erdas' elimination module (elimination of small contiguous blocks of pixels). The resultant land cover map contains many land cover objects and has consequently only a moderate readability (Figure 1-3). The quality of the land cover map is crucial because it is the basis for the second, land use classification. Accuracy assessment on this map was not performed but we certainly expect errors. An example is the presence of several water objects in the middle and top left part of the image. In this case, water has been confused with shadow. Indeed shadow is difficult to classify and sadly enough shadow is omnipresent in cities due to the high buildings. Shadow pixels were eliminated to some degree by means of shadow-specific majority filter.

[6] PCA (principal components analysis) method is one of the standard data fusion algorithms.

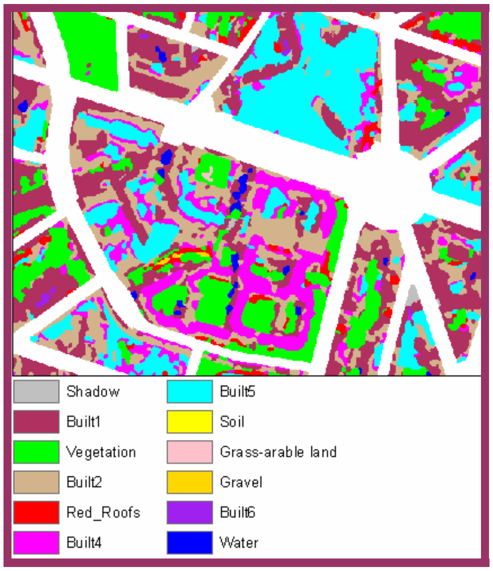


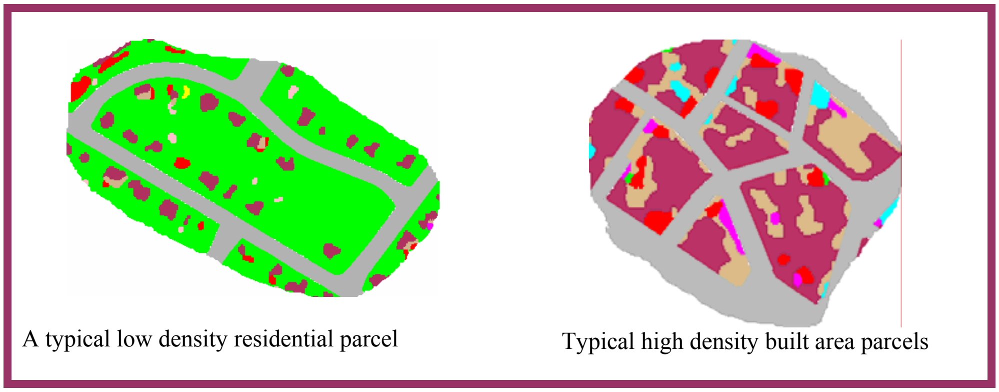
Figure 1-3: land cover classification (detail of European district)

1.3.3.2 Per-parcel classification: land use

For every parcel several land cover parameters were summarized:

- total area per land cover class
- mean area per land cover class object
- number of land cover class objects
- standard deviation for perimeter and area for all land cover objects

This resulted in 35 land cover parameters for every parcel upon which a discriminant analysis was conducted with the statistical software package Statistica. Discriminant analysis is used to determine which variables discriminated between two or more groups. The "best" variables are then used to predict group membership (of new cases): 24 out of the 35 variables were used in the discriminant analysis model. In fact these variables are transformed to fewer new variables – canonical roots - in such a way that these roots discriminate most between the groups. In our case the groups are the land use classes: high density built area or central city land use, medium density residential area, low density residential area, offices and industrial area and 'green' area. The last class is not really land *use* but more land cover, however since we are interested in urban land use we didn't focus on differentiating 'green land use'. For each of the land use classes some typical reference parcels were selected on which the discriminant and canonical correlation analysis was done. Figure 1-4 shows examples of a typical low density residential class and a high density built area class. It is clear that these two parcels differ from each other on a number of land cover parameters such as total green area, number of built land cover objects, mean area of built land cover objects etc. We expect these different land cover structures to be discovered in the canonical correlation analysis and the resulting canonical roots will describe these structures. The final result is a set of classification functions which allows us to classify all other parcels (over 2000 parcels) according to their scores on the roots.



A typical low density residential parcel

Typical high density built area parcels

Figure 1-4: Typical land use classes

Four canonical roots are calculated of which the first two explain the different land use very well (Figure 1-5). There are however two reference parcels of the type "offices and industrial" which tend more to type of "high density built" parcels. This distortion is also present in the third and fourth roots and possibly points out that these parcels represent an extra class between the two other classes. Another hypothesis is that these two parcels are merely outliers and should be neglected, thus accepting that they will be classified wrongly.

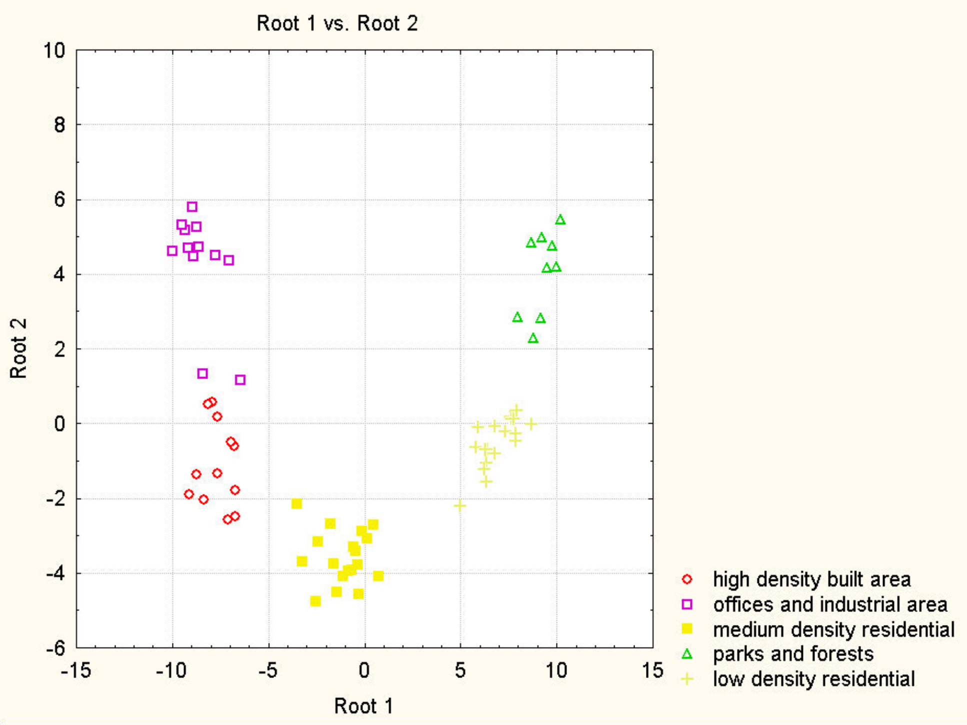


Figure 1-5: Root graph showing the five types of land use reference parcels

We can draw some conclusions concerning the interpretation of the first two roots. It is obvious that root 1 correlates with the level of morphological urbanization because on this root the land use classes can be ordered with decreasing level built up area. Thus, root 1 can be characterized as a "greenness axis": the higher its value, the more "green" the land parcel is or the less urbanized it is. The second root contrasts "offices and industrial area" and "parks and forests" on the one hand with "high density built area" and both residential classes on the other hand. In fact this root distinguishes between non-residential land use (the former) and residential land use (the latter) and represents a "residential axis". This is further confirmed because the root is mainly correlated with the "red roofs" land cover parameters which are typical for (suburban) residential land use. The third and fourth roots discriminate less between the land use classes and are more difficult to interpret.

1.3.4 Post-processing and results

After automatic classification of the five land use classes, two more land use classes, railway infrastructure and the canal were manually added. These two classes were not automatically classified because there were too few parcels involved making it much faster to add these classes manually. Then, some parcels with a low probability of being correctly classified were reclassified according to the land use of their neighbouring parcels. Suppose an office/industrial land parcel with low classification probability is surrounded by medium density residential parcels, then the central parcel would have been reclassified to medium density residential. This 'cleaning' process improved the map's readability but the effect on the classification accuracy was not tested. Finally a vector road layer was added to complete the map (Figure 1-6).

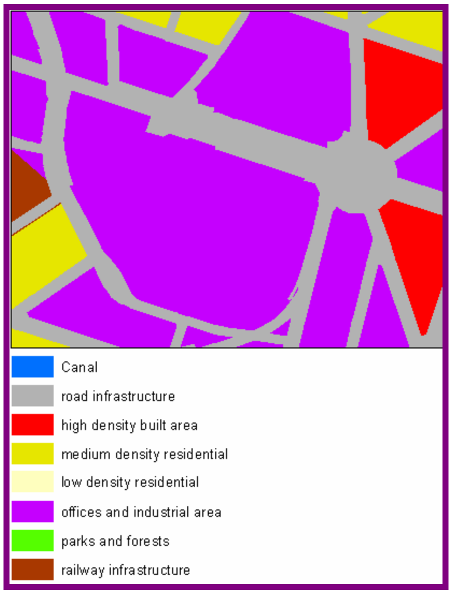


Figure 1-6: Land use map, per-parcel classification, European District

Classification accuracy is low, only 73%, which we believe is due to the fact of typical urban land use, which is more difficult to classify than non-urban land use (Table 1-2). In the accuracy analysis low and medium density built classes were merged because they were not separate classes in the reference layer (Corine Land Cover Map). Although the producer's accuracy (the percentage correctly classified reference parcels) is rather the same for all classes this is not the case for the user's accuracy which indicates that high density built area and offices and industrial area are the least well classified. This stresses again the difficulty of classifying urban areas where it is difficult to 1) produce a good quality land cover classification and 2) distinguish between offices/industrial land use and high density built areas. The distinction between those two land use classes is not completely clear as could be seen on the canonical root graph. Classifying urban land use is in many studies problematic and gives low accuracies (e.g. Aplin, 1999). Care should be taken when addressing all classification errors to the per-parcel classification technique because that is probably not the case. Improvements in the initial land cover classification and image quality (e.g. less oblique view angle, less shadows etc.) will have a substantial effect on the final land use classification. In addition the reference data's accuracy is questionable as the Corine Land Cover Map is small scale data compared to our land use classification. For the study region exists a more accurate large scale map but it is less applicable because the land use classes do not overlap well with our classes.

Table 1-2: Producer's and user's accuracies for per-parcel classification

	Producer's accuracy (%)	User's accuracy (%)
High density built area	72	60
Low & medium density built area*	73	83
Offices and industrial area	72	55
Parks and forests	76	89

**** low and medium classes are grouped because they are not separate classes in the reference layer***

1.3.5 Conclusion

For several technical reasons and because of the impossibility of monitoring traffic during the course of day with Ikonos images, the idea of extracting traffic safety indicators and traffic volume out of the image was abandoned and replace with the construction of a land use map. In order to make a land use classification, some pre-processing had to be done.

First, geometric pre-processing, to align the image with accurate reference maps, was attempted by means of an orthorectification procedure involving reprojection to correct for errors due to terrain height differences. However, the slight movements of the sensor during image capturing seemed also to play a role and this was too difficult to correct. So, instead of using the orthorectification procedure, an ordinary rubber sheeting transformation with many GCPs was done and resulted in an accurate image. A second part of the pre-processing was image fusion: Adaptive Image Fusion was used to fuse the bands into an input image which was an optimal input for the land use classification.

Land use classification requires a more complex approach than just land cover classification, which can fairly easy done by means of multispectral classification. In order to classify land use (e.g. residential area, industrial area) an external layer with land parcels was overlaid on the land cover image and several parameters were calculated for each parcel. Statistical analysis of this information allowed classification of the parcels into land use classes. Although the methodology seemed quite promising the overall accuracy is low (about 75%) with the specific urban land use classes being the least well classified. However, this does not mean that the technique is wrong because much depends on the quality of the image and the initial land cover map.

1.4. Delimitation of black zones

1.4.1 Introduction

During the past years, spatial accident research has evolved from the definition of black spots to linear black zones and recently it has been proven that linear zones are very well suited in situations with clear and distinguished traffic flows but in areas with a dense urban network and diffuse transport flows, it is more interesting to search for two-dimensional black zones. In such an environment accident locations are frequently based on proximity characteristics and two-dimensional clusters may suggest causal relations (Dufays et al. 2003). We present an application of a spatial clustering technique to road accidents, in the city of Brussels, capital of Belgium. We further discuss the effect of the incorporation of the actual accident exposure to calculate black zones representing accident risk.

The input data consist of 8.839 road accidents in the Brussels Capital Region covering 3 years (1997-1999). The accidents involve all kinds of road users but accidents involving vehicles are the most frequent. About 60% of all accidents in this database could be located on a detailed map by means of address matching and dynamic segmentation resulting in 5.148 located accidents.

1.4.2 Clustering of accidents: Ripley's K-function

The K-function is an exploratory method to describe the second order properties (spatial autocorrelation properties, related to clustering of points) of a point pattern (Bailey & Gattrell, 1996). We used the K-function to check whether the accident pattern could be characterized as clustered, regular or random. Further geographical accident research is in fact useless when accidents are not clustered in the vicinity of some spatially located "accident generator".

The K-function assumes that there are no large scale first order effects or *trends* present in the point pattern, which is not really the case (cfr. 1.4.3 Kernel density maps). Normally, K-values are rescaled to L-values and plotted in function of the distance. Such an L-function was calculated for roadsegment accidents in Brussels, for the 3-year period 1997-99 (Figure 1-7). In this plot, positive peaks indicate clustering and troughs indicate regularity, at corresponding scales of distance in each case. A random point pattern would result in L-values around 0. The L-plot for road segment accidents (accidents not on crossroads) clearly indicates clustering from the smallest scale up to circles with radius 3500 meters. The fact that accidents are bound to the road network could also cause the accidents to cluster (around the road network). Therefore we also plotted an L-curve for the road network[7] to compare with the one for the accidents. The road network doesn't show much evidence for clustering so we can exclude its influence on the accident clusters, although we will come to that point again. In the next sections shall we focus more deeply on this effect where it will be shown that there are differences in black zones (clusters) when incorporating accident exposure. But essentially, we can state that the road accidents are clustered.

[7] The road network was transformed from a line layer to a point layer representing the road density. In this way a K-function could be calculated for the road network.

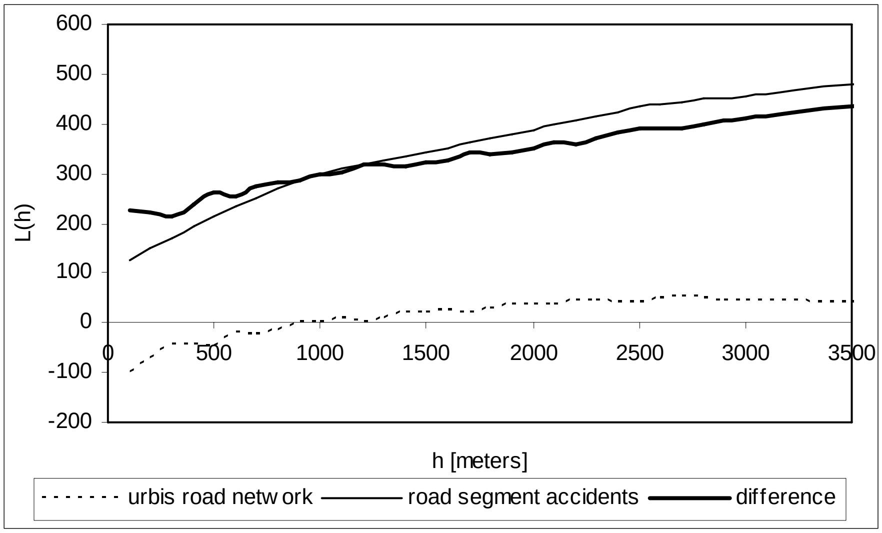


Figure 1-7: L function for road segment accidents and road network in Brussels

1.4.3 Kernel density maps

A density map is an exploratory spatial analysis tool to visualize first order effects or trends of a point pattern. They are useful to reveal the spatial structure of a point pattern and many GIS packages provide this tool[8]. We will briefly discuss density maps.

A smoothed density map is a continuously varying surface where each grid cell represents the density of the underlying point process and it offers a quick overview of the spatial pattern of the point events. Such a map can be calculated by overlaying the event datasets with a grid and calculating the event density (*f*) within a search region with radius h around the center of every grid cell. Below is the general formulation of such a density function:

$$f = \frac{1}{h^2} \sum_{1}^{n} W_i K(\frac{x - x_i}{h})$$

with x_i the location of an event and x the grid center point
K() is the kernel function
W_i is the event weight (optional)

The circles of neighbouring grid points are set to overlap in order to allow neighbouring grid points to share observations. The resulting grid is called a kernel estimation of the original pattern and it is a smoother, more continuous grid. Although the spatial distribution of accidents is bound to the road network, it is justified in a dense urban region to smooth the accident data over the whole space. This is because *proximity* is thought to play a major role in the spread of accidents within cities and in this way the accidents interact with the region or the regional characteristics (Dufays et al. 2003). Another advantage of smoothed maps is that the estimated densities are more stable than the original values because noise and outliers are filtered out. In this way, true spatial variation – if present – shows up (Talbot et al., 2000).

[8] ESRI ArcGIS provides density map functionality in the Spatial Analyst extension.

There are some parameters that control the way point patterns are smoothed which will now be discussed briefly. First, the *grid cell size* of the density map controls the detail and visual quality but has furthermore no important effects. We used a cell size of 50 meters, small enough for sufficient quality but not too small to allow fast processing in the subsequent simulations. The most important factor is the *bandwidth* or the search radius of the circular region, centred at each grid cell, in which events are summed and smoothed. The bandwidth size is proportional to the amount of smoothing and there exists some optimal value which gives an adequate representation of the regional distribution of the phenomenon under study (Bailey and Gatrell 1995). We made maps with several bandwidths and then choose two bandwidths of 250 and 150 meters because the maps provided good overview of the spatial structure at these two scales. A last parameter is the *kernel function, K*. This function determines the weight for each point within the search region and it is mostly a distance decay function from the grid centre point. A constant kernel function will result in less smoothed maps than a distance decay kernel function but the kernel choice has relatively little influence on the results.

Examples

Below are two illustrations of the effect of the bandwidth (Figure 1-8, Figure 1-9). Both maps show accident densities for the same set of accidents but with different bandwidths (150 and 250 m). First, the large bandwidth map has a better readability, less detail and allows easier interpretation of the overall pattern although essentially the pattern is the same. Second, the total area of high accident density zones is much larger for the large bandwidth while we find much higher densities on the small bandwidth map. Lastly, the same black zones show up in both maps but large zones break up into many smaller ones aligned with the main roads. When a large bandwidth is used, black zones tend to overestimate the reality. Examples are parts of the R0 and some of the arterials penetrating into the city which should be better off with a linear approach to black zones.

The notion of "black zones" should be clarified more clearly, in particular the delimitation of the black zones or how can we define a *suitable density threshold* to delimit black zones. In the two examples we used arbitrary boundaries to delimit black zones, respectively 43 and 75 accidents/km². It is however very questionable whether the resulting zones truly represent black zones. The next section will focus on this issue.

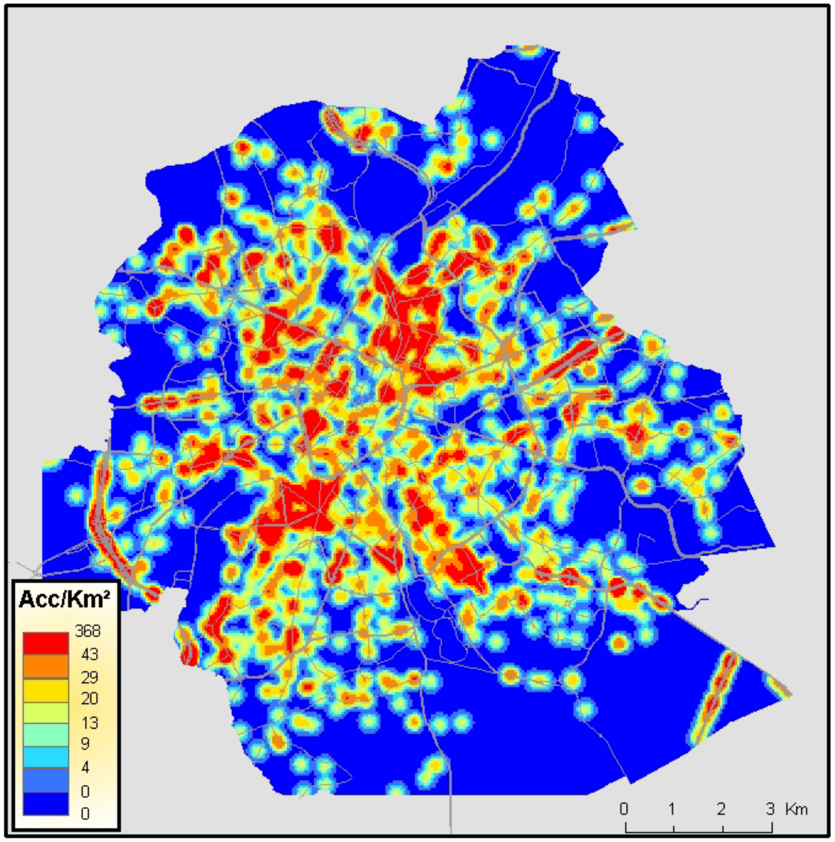


Figure 1-8: density map for road segment accidents in Brussels (1997-99), 250 m bandwidth

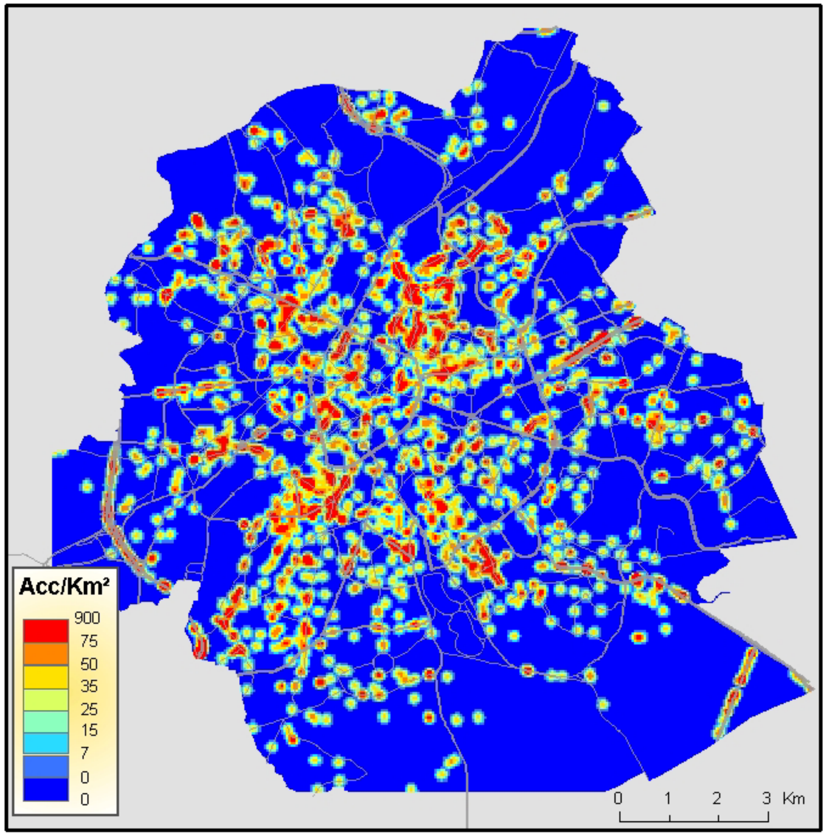


Figure 1-9: density map for road segment accidents in Brussels (1997-99), 150 m bandwidth

1.4.4 From densities to probabilities

1.4.4.1 Poisson probability

Whereas a density map is useful to gain more understanding of the underlying structure of a point pattern, it is however problematic to use this method for delimiting black zones. The problem is the definition of a suitable boundary or threshold value to delimit significant clusters. There is no rule or method to define such a value for a density map and this may lead to false results. The density maps presented in the previous section were simply classified into 8 quantile classes whereby we assumed that the last class would adequately represent black zones. In general, this classification method is known to produce good cartographic results but there is no evidence that the black zones truly represent real black zones. Suppose we apply the same method to a random pattern of points. Intuitively we wouldn't expect black zones to show up as the point pattern is random but a quantile classification will by definition classify the map producing "black" zones (Figure 1-10) which are in fact false positives. The histogram of the density values neither shows any clues for defining a threshold (Figure 1-11). Using a hard threshold is also wrong because the density values are highly dependent on the density bandwidth, higher bandwidths producing much higher densities (cfr. 1.4.3).

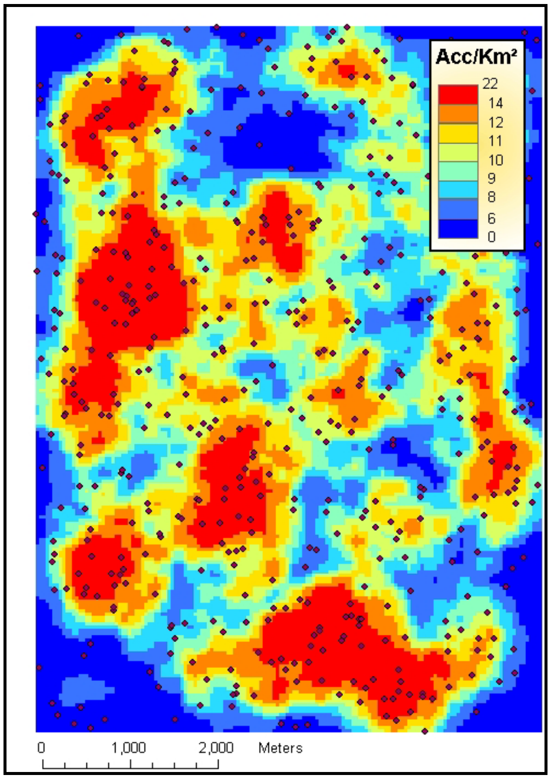


Figure 1-10: Density map of random point pattern (bandwidth = 600 m)

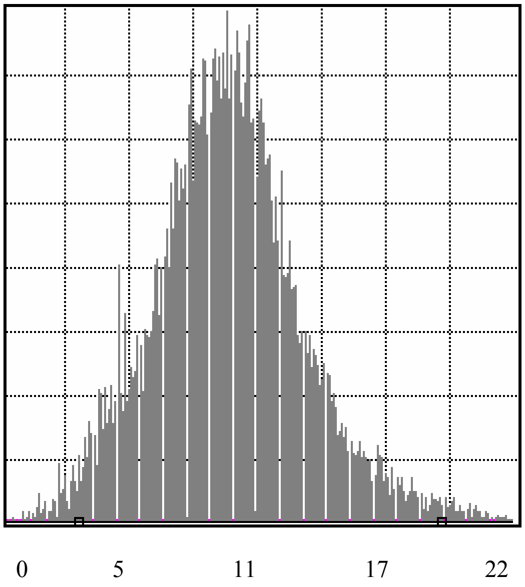


Figure 1-11: Histogram for density map

To overcome this problem and delimit true spatial clusters or significant black zones we need a statistical approach. One solution is the so-called probability map which uses Poisson probabilities to estimate the probability of the observed number of events to occur under the hypothesis complete spatial randomness. Such a probability map has been calculated for the same point pattern thereby revealing only some smaller clusters (Figure 1-12). The fact that there are still significant clusters (though small) can be interpreted as type I errors or false positives. Also, the histogram of probability values is very different then the one for the density values.

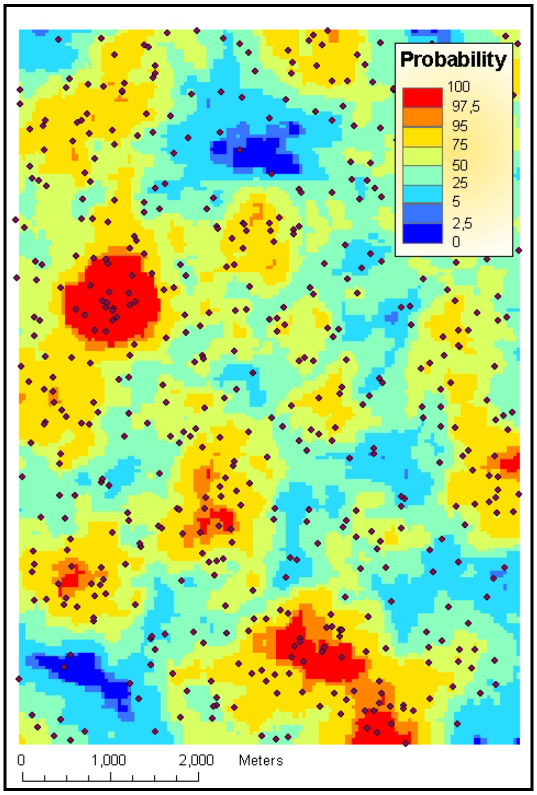


Figure 1-12: Probability map for random points (bandwidth = 600 m)

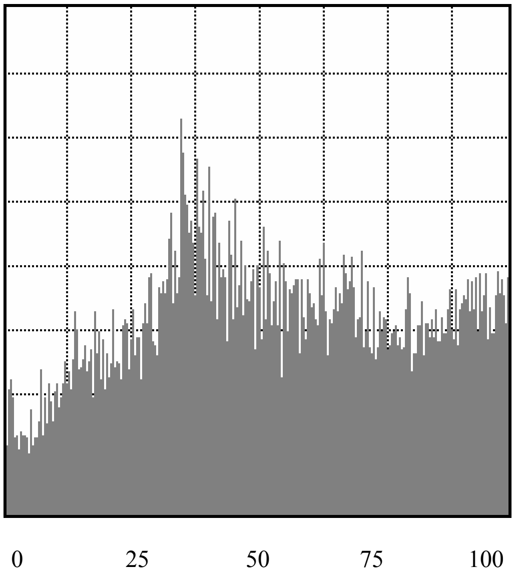


Figure 1-13: Histogram for probability map

We find extra evidence for the random nature of this point pattern by constructing Ripley's K-function (Figure 1-14). The maximum L-value for this pattern (about 35) is many times than the maximum L-value for the accidents (about 500), see Figure 1-7.

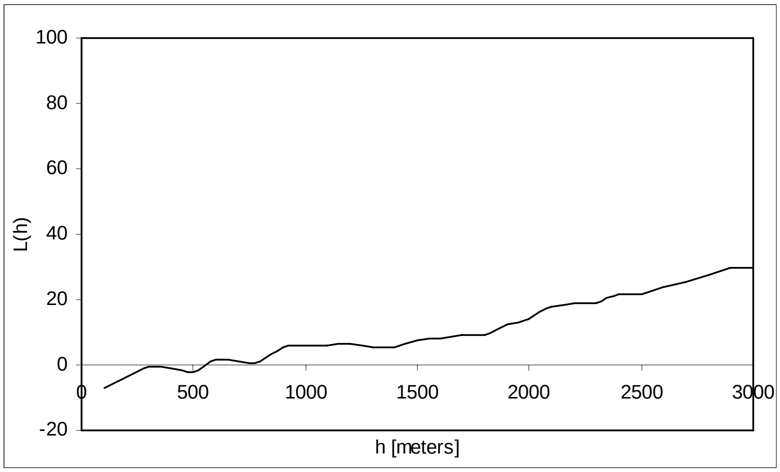


Figure 1-14: L-function for random pattern

We will now focus on the methodology behind the significance map. This method explicitly assumes that the point pattern is generated by a Poisson process. The Poisson test calculates the probability whether the observed number of accidents for a given grid cell (in fact within the search region around the grid cell) is significantly different from the expected number of accidents. Consider the following null hypothesis:

H_0: $Acc_{exp} = Acc_{obs}$

With Acc_{exp} = expected number of accidents
And $Acc_{exp} = \lambda_a$ = the Poisson distribution parameter.

λ_a can be approximated with the following equation:

$\lambda_a \cong n * p$

with n = total number of accidents
p = the probability for an accident occurring in the grid cell

and $p = \dfrac{Population_{loc}}{Population_{tot}}$

with $Population_{loc}$ the local population at risk, this is the population at risk within the search region around the grid cell
$Population_{tot}$ the population at risk for the whole study area

If n is large enough and p small enough then (n * p) gives a good approximation for λ_a. Now we can easily calculate $\lambda_a(x)$ for every location x in the study area (this for every grid cell and its surrounding search region). Knowing λ_a we can then calculate the probability that the observed number of accidents occurs under the null hypothesis. If this probability is less than 0.025, then the null hypothesis can be rejected and there is a significant excess of accidents in the vicinity of the grid cell[9].
An example of a probability map is shown in Figure 1-15.

[9] In order to compare the Poisson probabilities with Monte Carlo probabilities (cfr. 1.4.4.2) the Poisson probabilities were linear transformed: P' = (1 – P) * 100. In this way, the new probabilities P' are in the range 0..100 and those above 97.5 are significant.

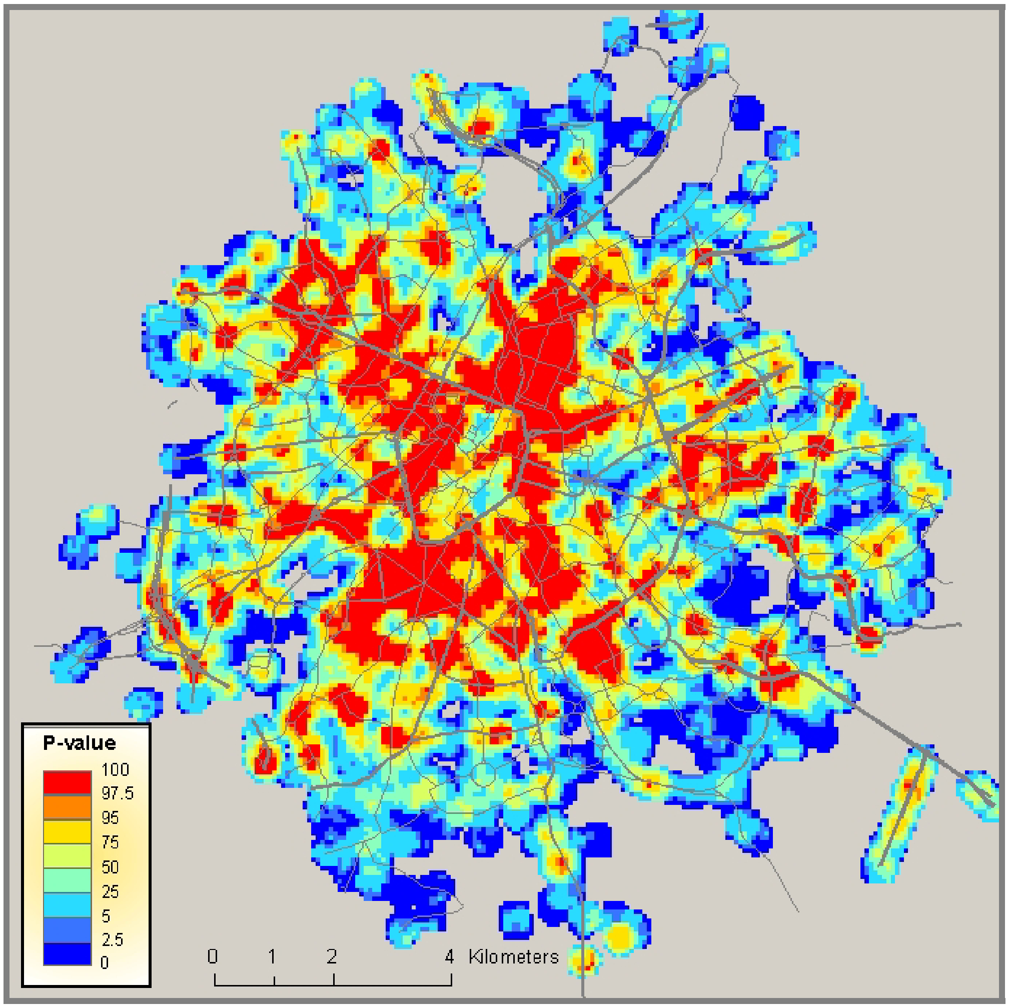


Figure 1-15: probability map, all accidents (1997-99), bandwidth 250 m

1.4.4.2 Comparison with Monte Carlo method

The Monte Carlo procedure is a very convenient procedure because it is not constrained by any assumptions concerning the distribution of the data; it is an assumption-free statistical test. So without any prior knowledge of the distribution process behind the accident point pattern we can apply the Monte Carlo test. In each simulation loop we randomly locate accidents in the study area and calculate a density map. The standard Monte Carlo test involves ranking the observed value of test statistic – the density value – amongst several simulated values. Once finished the proportion of simulated densities that are less than the observed density is computed for every grid cell and these results constitute the P-value surface. The disadvantage of the method is the long computing time which is proportional with the number of gridcell and simulation loops.

The comparison of two maps must be done with care and can be misleading depending on the method. Visual comparison, although not a really scientific method, confirmed immediately that both maps are the same (this is the reason why the Monte Carlo map is not shown, it looks the same as Figure 1-15). However, when a paired t-test[10] was done on the corresponding (paired) gridcells' probability values, it was concluded that the maps significantly different. The t-test is obviously not suited in this context. Another method to gain knowledge on the map similarity problem is a scatterplot[11] or correlation analysis

[10] T-test based on a sample of 100 gridcells

[11] Scatterplot and correlation coefficient based on 100 random gridcells

(Figure 1-16). The plot and the correlation coefficient (0,999) confirm the first visual impression that both maps are highly significant, thus the Poisson and Monte Carlo method perform in the same way for this dataset. The best choice would be the Poisson test because it is much faster.

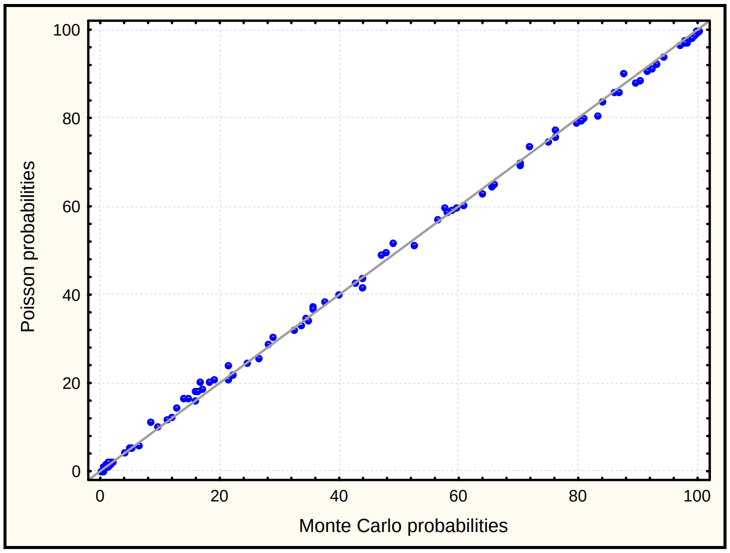


Figure 1-16: scatterplot of Poisson and Monte Carlo probabilities

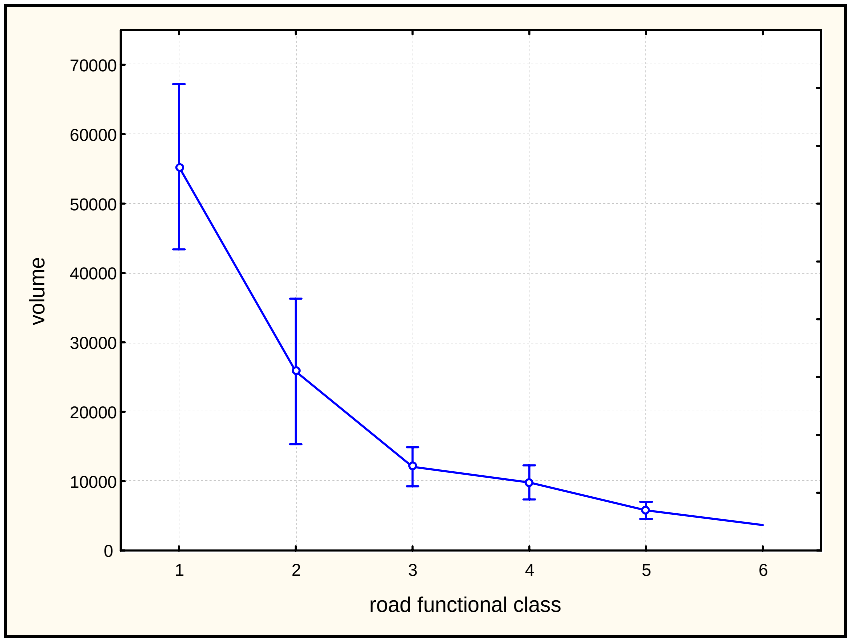
1.4.4.3 Accident *risk* zones: incorporating exposure

A very important factor in the calculation of the black zones is the 'population at risk' or accident exposure. Ideally the population should be some measure of the traffic volume on the road network. So far, we produced black zones with a uniformly distributed population at risk where it is assumed that the traffic volume is constant over the whole study area. This assumption is far from reality for a city like Brussels as there are zones with no traffic at all (e.g. parks) and zones with extreme high traffic volumes (e.g. ringways, main axes). Therefore it is more realistic to use the traffic volume as a measure for the population at risk but this requires at least an estimation of the traffic volume for every road on the network. Such traffic count data were only available for a selection of roads on which traffic counts were performed. Fortunately, we were able to estimate the traffic volume for all roads because the traffic volume[12] is closely linked with the road's function. This relationship is shown in Figure 1-17) and all functional road categories except 3 and 4 have significantly different traffic volumes[13] (alpha < 0.05).

[12] Traffic counts:

- Brussels Capital Region, 2003 and AWV 2002 for highways data (raised to 2003)
- Workday's mean volume over 24 hours, Traffic exiting the city, 108 count locations

[13] It is not possible to check if road category 6 (access streets) differs significantly from the other categories – though it is expected – because this category counts only 1 road.



road function legend:
1 = highways; 2 = urban arterial tunnels; 3 = primary roads; 4 = secondary roads; 5 = important local streets; 6 = access streets

Figure 1-17: Traffic volume vs. functional road class (95% confidence intervals)

With the information from the graph, traffic volume weights were derived and assigned to all roads. It is important to consider the difference between the two – uniform and true – populations at risk datasets. Black zones produced with the *true population at risk* represent zones with a significantly high *accident risk*. On the other hand when we assume a constant or *uniform population at risk*, the black zones will represent zones with a significantly high number of accidents or high *accident frequency.* To conclude, two types of black zones can be delimited depending on the type of population at risk: "absolute black zones" (Figure 1-18) and "risk black zones" (Figure 1-9). The main difference between both maps is that absolute black zones occupy a much larger area than risk black zones but the spatial structure for both types of zones is rather the same. An example that illustrates the difference is the situation on some specific sections of highways, in particular on the R0, E40, A12 and the extension of the E411. Along these major highways we find many absolute black zones but not any risk black zone. These are locations with a high accident frequency but with a low accident risk due to the huge traffic volumes on these roads. However the opposite situation apparently doesn't exist as we find no zones with a low accident frequency and a high accident risk.

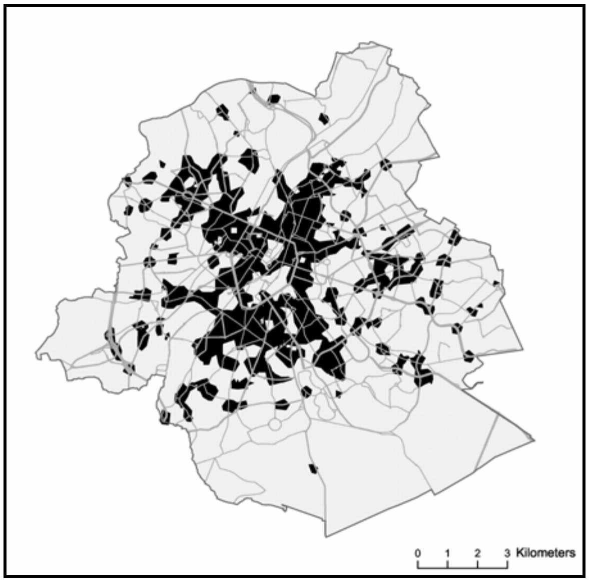


Figure 1-18: absolute black zones, all accidents (1997-99), bandwidth 250 m

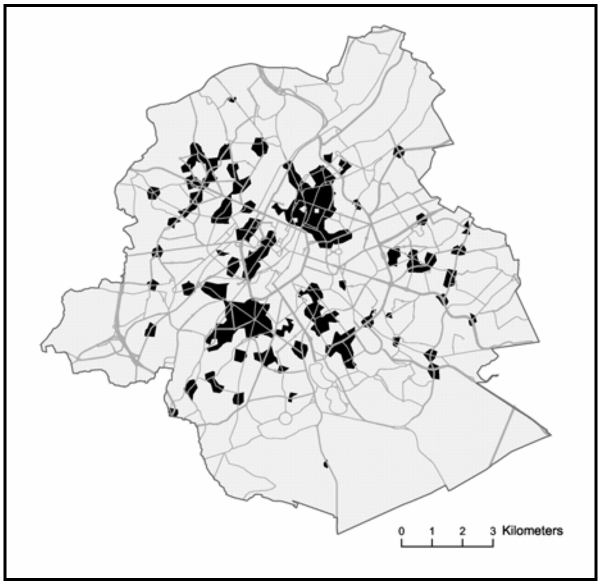


Figure 1-19: risk black zones, all accidents (1997-99), bandwidth 250 m

1.4.5 Discussion of traffic safety in Brussels

1.4.5.1 Temporal stability of black zones Brussels

A small analysis of the temporal stability of the black zones was done to check to what extent the black zones are stable. The 9 year period for which located accidents were available was divided in 3 periods of 3 years for which probability maps and (absolute) black zones were calculated. By means of GIS overlay the proportion of common black zones were calculated (Table 1-3). The results favour the stability hypothesis as the percentage stable black zones oscillates around 50% for two consecutive periods and 30% for all three periods. Figure 1-20 shows the spatial arrangement of the stable zones. The stable zones (during all 9 years) seem to represent the core unsafe traffic zones in Brussels and they consist of a few large zones more or less in the centre and several smaller peripheral zones along highways.

Table 1-3: Relative share of common area of risk black zones

	Common area black zones (%)
91-93 AND 94-96	62
94-96 AND 97-99	46
91-93 AND 97-99	40
91-93 AND 94-96 AND 97-99	30

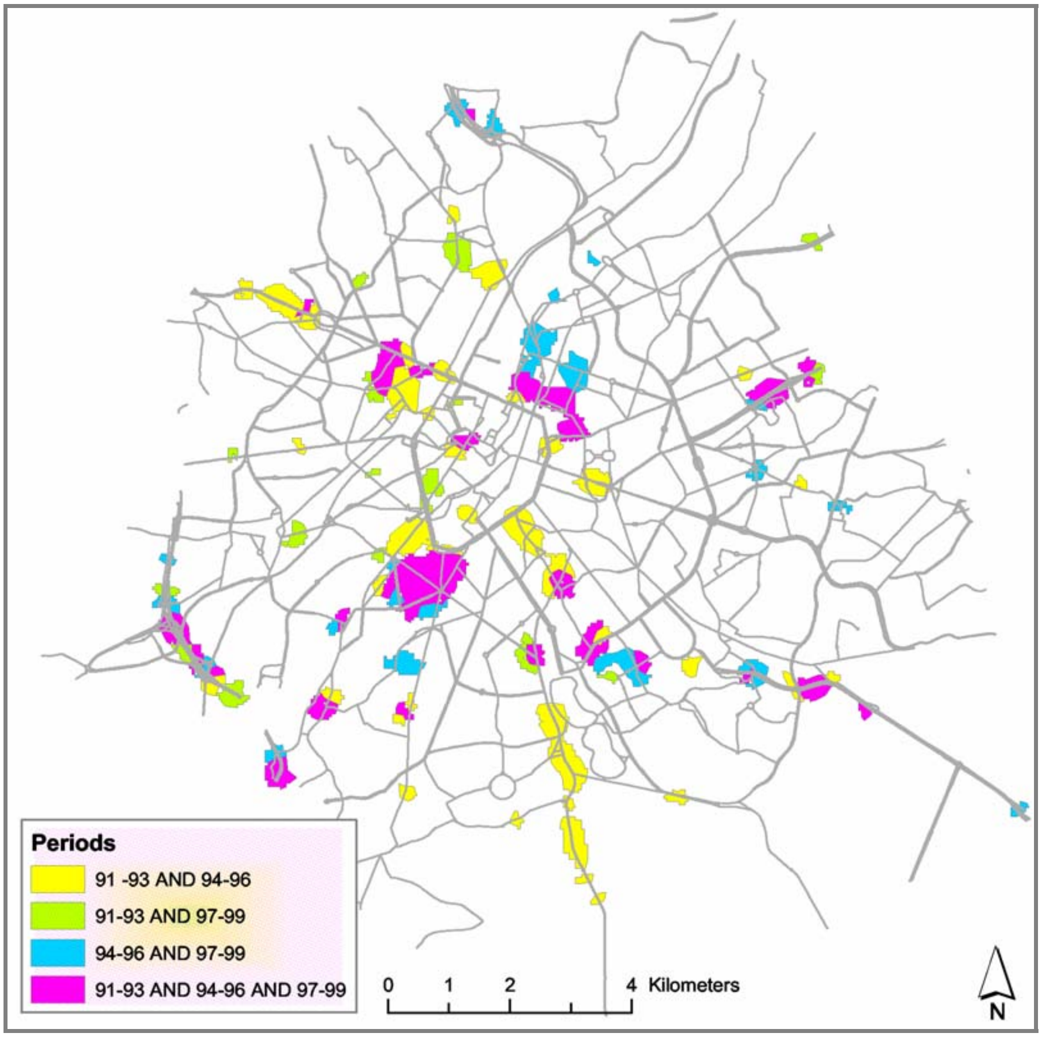


Figure 1-20: Temporal stability of absolute black zones

1.4.5.2 Outline of black zones

A brief presentation of traffic (un)safety in Brussels as measured by risk black zones will be presented. Based on visual inspection we can classify the black zones in 5 classes (Figure 1-21):

- Two large central zones, to the north and south the pentagon
    - In the north: near St. Joost and Schaarbeek
    - In the south: near St. Gillis
- One central zone
    - along the Anspachlaan and M. Lemonnierlaan
- A few small zones associated with major roads or highways
    - E40, R0
    - Vorstlaan & Herrman Debroux viaduct
- Several relatively small zones inside the middle ring
    - Flageyplein
    - Boondaelsestwg
    - Brugmanlaan
    - others

- Several relatively small zones in the outer belt, along important roads
    - M. Groeninckx De Maylaan – L. Mettewielaan – de Smet de Naeyerlaan)
    - Woluwedal and Woluwelaan
    - others

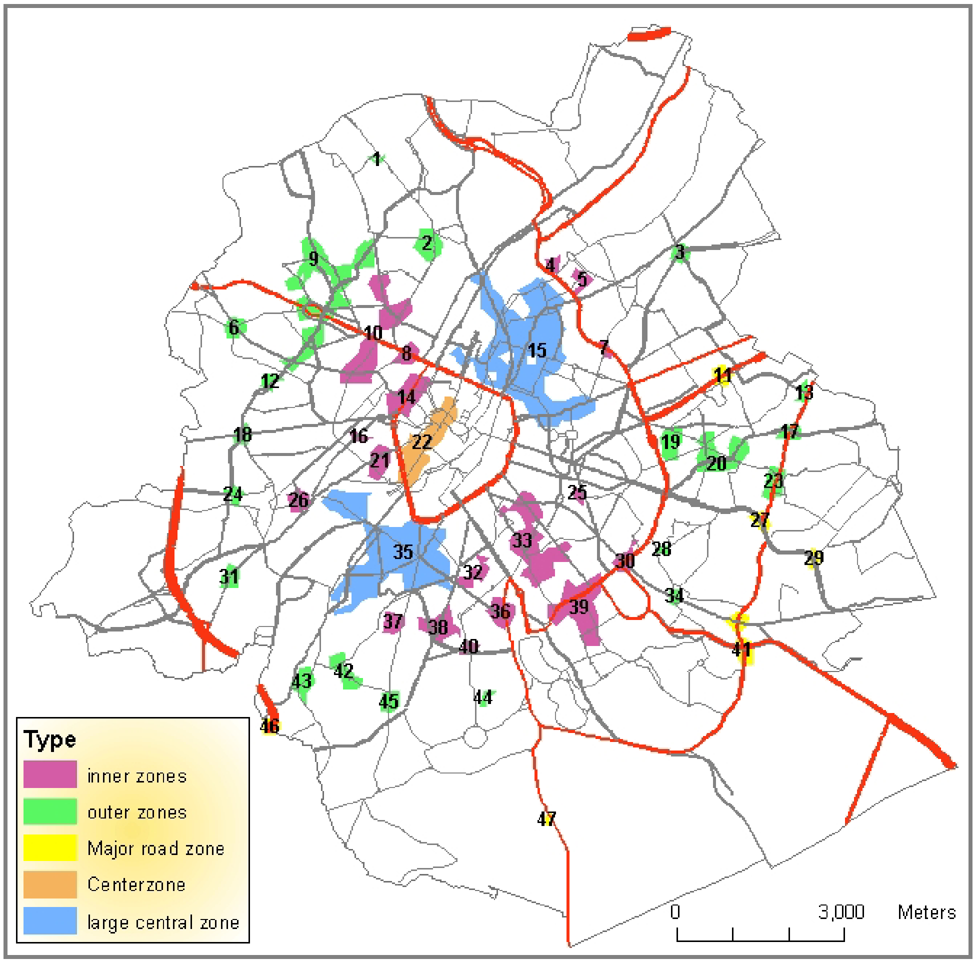


Figure 1-21: Risk black zones by type

1.4.6 Conclusion

This part of the research focussed on delimitation of two dimensional black zones in Brussels and various point pattern analysis techniques were applied in order to detect such clusters. Ripley's K method was first applied to check whether clustering was present at global level and this test confirmed our common sense that accident clusters were present. The test showed that clusters at all scales (at least up to radii of 3,5 km) are present.

Second, we made a kernel density map of the accident locations to enhance the visualization and this map clearly showed concentrations of accidents but it didn't allow a solid delimitation of black zones. To solve this problem, this is to find a statistically significant threshold to delimit significant accident concentrations and thus define black zones, we made probability maps. These are comparable to density maps – at least visually – but they are constituted of probabilities which allow delimitation of significant black zones and hence a black zones map of the accident frequency was created. Besides accident frequency another interesting concept, accident risk or the ratio of accident frequency and accident

exposure, can shed light on traffic safety. Once the spatial distribution of exposure (e.g. traffic volume) is known in the study area, the risk can be mapped the same way with risk black zones as output. Traffic volume was extrapolated for the whole road network of Brussels based on a relation between traffic volume and the function of roads, which was calibrated with real traffic measurements in Brussels.
Finally, the temporal stability of the black zones was checked by overlaying three black zone maps of three consecutive 3-year periods and about 30% is stable over all 9 years and about 50% of the area black zones doesn't move between two periods. Concerning the spatial accident risk pattern in Brussels for the period 1997-99, several zones show up: two large zones to the north and south of the Pentagon, one zone in the city centre, several small zones in outer belts and a few small zones along the major roads or highways.

1.5. Explanatory model for accident frequency

1.5.1 Introduction

The last part of our research is about explaining the accident frequency at neighbourhood level. The basic spatial unit of analysis is the statistical district for which census data is available. The objective is to check whether there is a relation between demographic and socio-economic factors on the one hand and accident frequency on the other hand. We also included simple factors describing the morphological type of built environment and an accident exposure factor which is always a very important predictor of accident frequency.
Special care has been given to the construction of the regression model and the assumptions this model required. Specific assumptions have to be met to perform the regression correct and to achieve reliable results, in particular the discrete nature of the dependent variable (count data) and the spatial autocorrelation problem impose specific assumptions which will be discussed.

1.5.2 Data description

The unit of analysis is the statistical district, the smallest areal unit for which census data is available, from the "Volks- en Woningtelling" in 1991. From the 724 statistical districts in Brussels 699 were valid (no null values or zero inhabitants. e.g. parks). The dependent variable is the accident frequency or the number of accidents within a statistical district during three years (1997-1999). This posed a problem with accidents that are located on streets which act as borders between districts, which is often the case. We used two different methods to solve this ambiguity: first we randomly shifted these border-accidents so they fell completely in a district. The second method redistributed all accidents by means of density map as an intermediate step. The resulting accident frequencies are more smeared out over the statistical districts and the effect on this on the regression conditions, in particular on the spatial autocorrelation problem will be discussed further.

Around 30 socio-economic factors and a measure for the accident exposure[14] in the district are available at the level of the statistical district (E.1 Data: neighbourhood factors). They cover a range of socio-economic dimensions:

- Demographic (Age groups, nationality)
- Housing (comfort level)
- Morphological building type (open, closed, etc.)
- Social (socio-economic position, labour –group)
- Mobility related (car and bicycle availability for the household)

[14] Exposure for a statistical district = cumulative length of roads x traffic volume

Several of these factors are highly correlated making some redundant and useless. Besides, introducing a correlated factor in a regression model gives rise to problems like significant factors becoming insignificant. Luckily the exposure factor is not correlated with any of the other factors. Different solutions are possible (principal components, cluster analysis) for the correlation problem but we used a small subset of factors with low correlations (maximum R = 0.5). This means that the choice for factor X prevents choice for correlated factor Y or vice versa. See Appendix E.2 for sets uncorrelated factors.

1.5.3 Regression methods

1.5.3.1 Regression models for count data: Generalized Linear Models

Statistically it is not justified to model accident frequency (which is count data) with a standard linear model. The assumptions of the "ordinary least squares" (OLS) model are violated when using count data and generalized linear model should be used instead.

Advantages of generalized linear models:

- Data can have other distribution than the normal distribution. (E.g. count data is mostly Poisson distributed)
- Data (dependent variable) can be restricted to ranges (E.g. proportions : 0-1)
- Variance must not be constant for all observations

Components of generalized linear models:

- Linear component
- Stochastic component (distribution)
- Link function
- Variance function

The model has this form:

Accident frequency = $(\text{Exposure})^{a} * e^{(\beta X)}$

with

Factors: Exposure, $X = (x_1, . . ., x_p)$ = p prediction factors[15]
Parameters to estimate: a = exposure exponent; β = p parameters
Distribution = Poisson or Negative Binomial
Link function = Log

Count data can be modelled with a Poisson or a Negative Binomial distribution. The Poisson distribution assumes that the variability of the data is equal to the mean of the data (= Poisson distribution characteristic) but this is often not the case with accident frequency (Sawalha & Sayed, 2001). In such a situation, the data is overdispersed and a Negative Binomial regression should be preferred. Unlike a Poisson regression the variance of the negative binomial distribution can be greater than its mean. The degree of dispersion can be estimated by dividing deviance and Pearson ChiSquare by the degrees of freedom. Values greater then 1 indicate overdispersion.

1.5.3.2 Spatial autocorrelation, spatial regression

A major statistical assumption in classical regression analysis is independency of observations. However when dealing with spatial data the values of variables in neighbouring units are often correlated, having

[15] X is a vector of uncorrelated factors

biased regression results as a consequence. There exist several ways of dealing with spatial autocorrelation ranging from easy to complex approaches (Elffers, 2003).

Approach 1: dependency on own explanatory variables only
"Spatial correlation is seen as the result of misspecification of the model, and a sure sign that we have to hunt for new explanatory variables – displaying some spatial correlation themselves – that could solve the unwanted spatial dependency for us." (Elffers, 2003)

Model

$$\mathbf{y = X\beta + \varepsilon}$$

The reason for spatial autocorrelation has been defined by means of the common values of explanatory variables in neighbouring areas.

Approach 2: dependency on adjacent explanatory variable
"Neighbourhood influence in this case is being channelled through a shared influence of explanatory variables, whose influence spills over the boundaries of areas" (Elffers, 2003).

Model

$$\mathbf{y = (W + I)X\beta + \varepsilon}$$

W = adjacency matrix
I = identity matrix

Approach 3: dependency on adjacent dependent variable (spatial lag model)
This is the case of real spatial autocorrelation and will be clarified with an example from criminological context. Spatial autocorrelation must be understood as in the following: "the feelings of not being safe (dependent variable) in one area is influenced by the feelings not being safe in neighbouring areas" (Elffers, 2003).
This reasoning does not really hold in the context of traffic accidents (LaScala et al. 2000). Such a model assumes that accident rates in one location are related to those in adjacent areas. This can only be the case with large multiple collisions stretching out over different areas.

Spatial lag model

$$\mathbf{y = \rho Wy + X\beta + \varepsilon}$$

y = vector dependent observations,
ρ = spatial autoregressive coefficient,
Wy = spatially lagged dependent variable,
X = matrix of explanatory variables,
β = vector of regression coefficients, and
ε = vector of error terms.

Approach 4: dependency on other external, non-observable explanatory factors (spatial error model)
In this approach it is assumed that other unmeasured factors related to the dependent variable cause the spatial correlation. Inclusion of these supposed factors should remove the spatial correlation (as in approach 1). LaScala et al. (2000) believe that this situation fits better in the context of traffic accidents.

Spatial error model

$$\mathbf{y = X\beta + \lambda W\varepsilon + \xi}$$

$W\varepsilon$ = spatial lag for error terms,
λ = autoregressive coefficient, and
ξ = error term with mean 0 and variance matrix $\sigma^2 I$.

1.5.4 Model results

1.5.4.1 Model Choice

First we attempted a count regression in SAS without a correction for the spatial autocorrelation. The residuals of this regression can then be used to assess the existence and the degree of spatial autocorrelation. In the case of strong spatial autocorrelation, the regression should be repeated with a correction for the autocorrelation. Results in SAS indicate a moderate level of spatial autocorrelation in the residuals. Because there exists other specialised software for spatial regressions we performed the regression with GeoDa, a recently developed application which is free downloadable. Unlike SAS, GeoDa has specific functions to calculate spatial weights as input for the spatial regressions but on the other hand is it not possible to perform count regressions with GeoDa as it supports only OLS regressions. However the count data can be log-transformed to comply with the OLS assumptions.

Thus, two methodologically different regressions have been done: a count regression in SAS using the generalized linear model framework[16] and a spatial OLS regression in GeoDa. Both will be discussed below.

1.5.4.2 Count regression results (SAS)

We modelled accident frequency by means of a count regression which is a member of the family of generalized linear models.

First, we checked whether overdispersion was present in the data by making a Poisson regression and checking the dispersion indicator. There was a serious level of overdispersion present and consequently we fitted a Negative binomial model which fitted the data much better. In this model, the dispersion indicator was much closer to 1 (it dropped from 3,25 to 1,14; cfr. Appendix E.4).

Next, we checked for outliers in the data. SAS' proc GENMOD does not provide exhaustive regression diagnostics as proc REG does for a linear regression. Outliers are data points which have a large influence on the regression parameter estimates. The influence is a combined effect of high leverage and high discrepancy (residual error) and it can be expressed by Cook's distance. The leverage is a measure of the eccentricity of the datapoint in the vectorspace of the predictors. The higher the leverage the more the datapoint is distant from the centre of gravity of all datapoints. However GENMOD does not provide Cook's D nor leverage values. We took leverage values from an ordinary linear regression (with the same predictors) and combined these leverage values with the residuals from the NegBin regression to get Cook's distance values for our regression model. Then we ordered the data points by decreasing Cook's D and iteratively removed the data point with highest Cook's D and recalculated the model's deviance. As long as the new deviance dropped more than 3.84 (ChiSq(0.05,1)) we continued the removal of outliers. A total of four outliers were removed and this resulted in a model that now better fits the data because the dispersion indicator is near 1 (Table 1-4).

Table 1-4: Criteria for assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	688	796.1282	1.1572
Pearson Chi-Square	688	694.8894	1.0100
Log Likelihood		6152.8255	

[16] SAS: GENMOD

The goodness of fit can also be evaluated graphically. One way is a plot of the Pearson residuals against the predicted number of accidents. The residuals are clustered around zero which is OK but there still is a pattern present which is due to the discrete nature of the count values and thus without consequences (Figure 1-22).

Another way is a plot of the average of the squared residuals together with the variance function (for the negative binomial distribution: variance = μ + k * μ²). There seems to be a good correspondence between the variance of the data and the theoretical variance function (Figure 1-1).

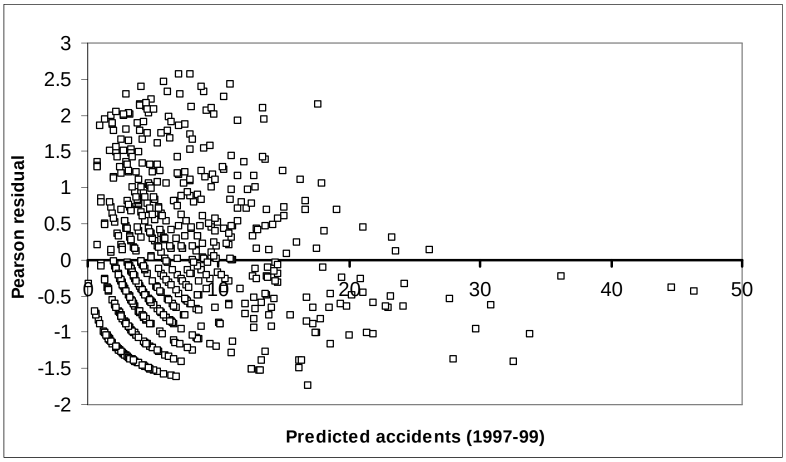


Figure 1-22: Pearson residuals

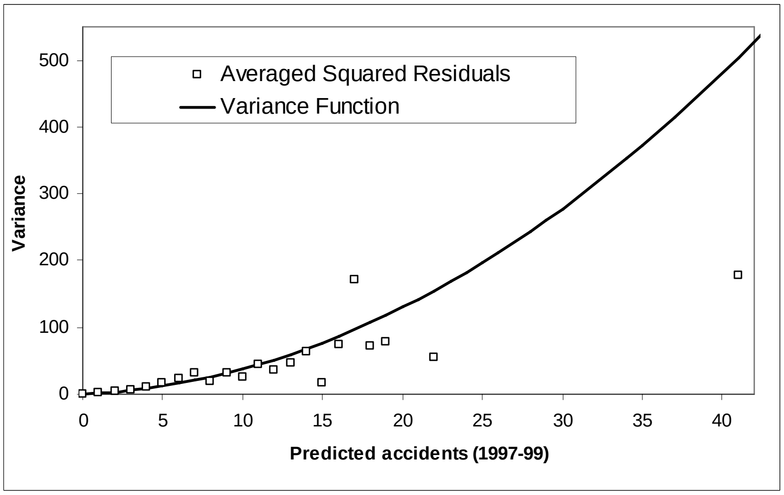


Figure 1-23: Variance of predicted accidents and theoretical variance function

Inspection of the residuals indicated that there is still a certain degree of spatial autocorrelation present (Table 1-5). The degree of autocorrelation for the residuals is however much lower than the one of the observed accident frequency but it can't be ignored and as a consequence model results will not be reliable. A graphic illustration of the spatial autocorrelation effect can be found in the appendix (E.3 Graphic illustration of spatial autocorrelation of Negative Binomial regression residuals). The unreliability of the parameter estimates and probabilities and the fact the parameter estimates are very low (i.e. they are of little importance) make it not meaningful to discuss and interpret the model more thoroughly.

Table 1-5: spatial autocorrelation analysis for the Negative Binomial model

Lag Nr*	Observed Acc. Freq.	Model residuals
1	0.2519	0.1572
2	0.1579	0.1176
3	0.1001	0.0639
4	0.0627	0.0444

* The lag number is the order of adjacency of the spatial units (polygons). A spatial lag of 3 means that three 'rings' of adjacent polygons (adjacent to a central polygon) are used in the calculation.

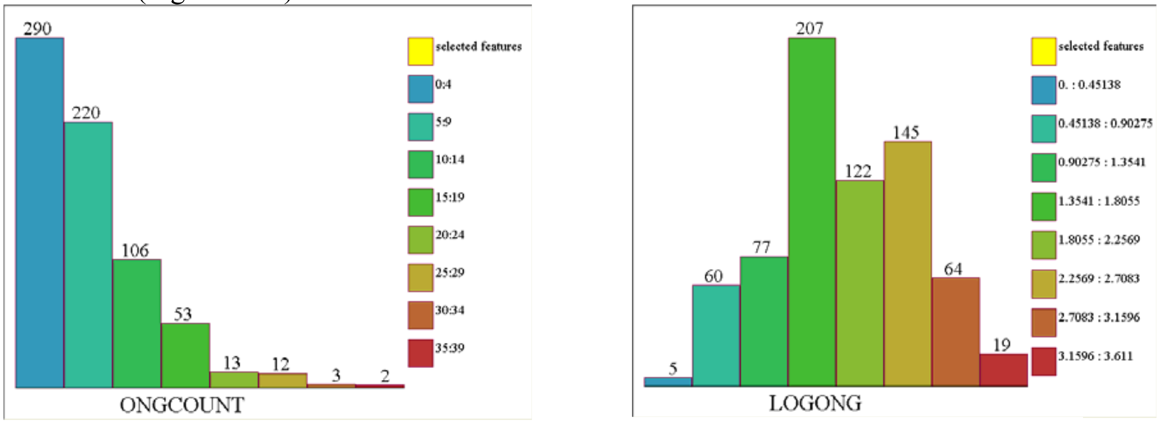
1.5.4.3 Spatial regression results (GeoDa)

Because GeoDa can only model linear relationships and assumes specific conditions (normality of the data etc.) and our data, as has been shown above, follows a negative binomial distribution and should be modeled with a Poisson or Negative Binomial distribution, the data should be transformed to comply with the assumptions of a linear regression:

The nonlinear model $\boxed{\text{accident frequency} = (\text{exposure})^{a} * e^{(\beta X)}}$

can be LOG linearized $\boxed{\mathit{Log}(\text{accident frequency} +1) = a * \mathit{log}(\text{exposure}) + \beta X}$

The log transformation is frequently used for count regressions because it stabilizes the variances, to some extent. The histograms below show that after transformation the distribution seems quite normally distributed (Figure 1-24).



a) original accident frequency

b) log-transformed accident frequency

Figure 1-24: histograms for original and log-transformed accident frequency

The next step is about the building of an appropriate and reliable accident prediction model. The starting point consisted of a *basic OLS model* with only one explanatory variable (exposure, supposedly the most important factor) without spatial correction. As expected, the exposure factor was found very significant and spatial autocorrelation was present in the residuals of this model (Moran Index: 0.52 at lag 1). A first attempt to correct the model is by adding extra spatially correlated explanatory variables (cfr. 1.5.3.2 Spatial autocorrelation, spatial regression, Approach 1: dependency on own explanatory variables only). This way, both the variance explained by the model (R^2) increases and spatial autocorrelation decreases. But applying this approach didn't succeed completely in explaining away autocorrelation because the Moran Index remains high (0.35 at lag 1). Thus a more specialized technique is necessary to deal with the spatial autocorrelation. Based on the OLS regression diagnostics (robust Lagrange multiplier test for spatial lag and spatial error dependence, Anselin et al. 1996) we choose the spatial error model (Table

1-6). The table lists the regression diagnostics for the basic OLS model. Because both the robust LM (for the lag and error model) are significant, the highest LM determines the model, which is the spatial error model ($LM_{error} > LM_{lag}$). As a consequence the spatial error model will be preferred instead of the spatial lag model.

Table 1-6: spatial autocorrelation regression diagnostics for the basic OLS model

TEST	MI/DF	VALUE	PROB
Moran's I (error)	0.52	23.5	< 0.001
Lagrange Multiplier (lag)	1	**291.9**	< 0.001
Robust LM (lag)	1	18.2	< 0.001
Lagrange Multiplier (error)	1	**542.8**	< 0.001
Robust LM (error)	1	269.1	< 0.001
Lagrange Multiplier (SARMA)	2	561.0	< 0.001

Table 1-7 summarizes the characteristics and factor significances of the different models that were fit, starting with the basic OLS model to the full spatial error model. The models' spatial autocorrelation effects are listed in

Table 1-8.

The full OLS model includes 4 extra variables (of which 3 were found significant) and explains accident frequency better than the basic OLS model (R² rise from 40 to 54%) but there is still a lot of spatial correlation (M.I. = 0.35), invalidating this model. Extra variables should not be added because they are correlated with each other. The basic spatial error model is able to explain even more of the variance than the full OLS model (R²=69%) and spatial autocorrelation has disappeared. Thus, the spatial correction of the regression model leads to high increase in explained variance compared with the basic and full OLS model or in other words the spatial effects are far more important than the neighbourhood factors from the full OLS model. This finding is confirmed by the very small increase in R² when extra neighbourhood factors are introduced in the spatial error model (R² increases with only 1%). Although three factors are significant they cannot really increase R². *An important conclusion is that the neighbourhood variables used do not really matter in explaining the occurrence of accidents.* Therefore it doesn't make sense to discuss the factors and their influence more in depth.

Table 1-7: Model results

	Basic OLS	**Full OLS**	**Basic SpError**	**Full SpError**
R² (%)	40	54	69	**70**
LogLikelihood	-561.282	-465.75	-375.753	**-352.744**
Exposure	0.5353*	0.5011*	0.5901*	0.5629*
Poor Neighbourhood	NA	0.2583*	NA	0.0434
ApartmentsStudios	NA	0.0004*	NA	0.0001*
BuiltMix	NA	-0.0032*	NA	-0.0021*
BuiltOpen	NA	0.0004	NA	0.0015*

* significant at 0.05

Table 1-8: Moran indices for accident frequency and regression residuals

Lag Nr	Observed Acc. Freq.	Basic OLS residuals	Full OLS residuals	Basic SpError residuals	Full SpError Residuals
1	0.2905	0.5186	0.3516	-0.0332	-0.0329
2	0.1959	0.3732	0.2230	0.0063	0.0098
3	0.1331	0.2696	0.1460	0.0189	0.0190
4	0.0885	0.1733	0.0883	0.0091	0.0099

We conclude this part with some notes on the heteroscedasticity problem. This is the case when the residual variance is not constant and as a consequence standard errors for the parameter estimates are biased. The estimates themselves are unaffected but their significance is biased. Unfortunately all models, except the full OLS model, exhibit heteroscedasticity according the Breusch-Pagan test (Heteroscedasticity graphs: Cfr. Appendix E.5)

1.5.5 Conclusion

It has been demonstrated that the regression of accident frequency requires some specific regression techniques. The fact that accident frequency is count data requires modelling by means of a generalized linear model which takes the characteristics of count data into account. Besides, this is a spatial regression and as with most spatial data, autocorrelation is present in the data and this effect should be included in the model. We conclude that SAS statistical software is a valuable tool for generalized linear regression but it is very difficult to perform spatial regressions with SAS. Instead it is better to use other software like GeoDa which is an excellent tool for performing ordinary least squares regression with spatial correction.

The results show that accident exposure is very important in explaining accident frequency but the socio-economic factors can't improve the predictive power of the model. Thus, the effect of the neighbourhood, as we measured it, on the accident frequency is in fact inexistent. The fact we found several large neighbourhoods in our black zone analysis might not be the result of causes at neighbourhood level but it might be related to the road infrastructure or traffic flows with similar characteristics in that neighbourhood.

2 LUC

2.1. Research goals

The research tasks of the project "Innovative spatial analysis techniques for traffic safety", assigned to the L.U.C., can be divided into two separate parts. During the first part of the project, the tasks stipulated by the network agreement consisted in performing a first analysis on road infrastructure, land use and traffic accident data, through the application of several data mining techniques. The research goals, associated with this part of the research, are to assess the explanatory variables on their relevance and to make a proper selection of useful variables. These tasks are denoted under heading B2 of the "Technical specifications" of the network agreement. The results of these first research tasks, which are to a great extent already been at the 6th Design and Decision Support Systems in Architecture and Urban Planning Conference in Ellecom (The Netherlands) (Geurts, 2002), are fully discussed in section 2.2 and 2.3.

The second part of the research concentrated on spatial clustering, based on data mining techniques, in order to identify and profile black spots and black zones. The use of data mining techniques, which tries to maximize the similarity within a cluster and the dissimilarity between clusters, should eliminate some disadvantages of classic hot spot and black zone techniques. The goal of this part of the research is not only to detect black zones and black spots, but more importantly, to profile the black spots in order to extract traffic safety policy decisions. These tasks are denoted under heading C2 of the "Technical specifications" of the network agreement. The results of this part of the research have already been published at the 83rd annual meeting of the Transportation Research Board of the national academies in Washington (U.S.A) (Casaer, 2004) and are fully discussed in section 2.4.

2.2. Quality assessment of the Belgian traffic accident data

2.2.1 The characteristics of the data

Our research about the relevance of the different variables in explaining the Belgian traffic situation and more precisely in explaining black spots and black zones rely heavily on the traffic accident data at hand. Therefore, our first goal was to assess the quality of the Belgian traffic accident data. After all, in order to draw a correct policy, one needs to base his policy on correct and valid observations.

The quality assessment is performed on the data for the period 1991-1999, which contained information about traffic accidents with casualties or fatalities involved within the Flemish, Walloon and Brussels regions. These data were acquired by use of the 26 sections of the VOF form which were put at our disposal by the regional administrations[17]. Several agencies, such as the regional ministries of public works, the provinces and several universities[18], have corrected the data when and where needed. This resulted in a dataset of 505880 traffic accidents records(Appendix F.1 and F.2).

[17] Original source: Nationaal Instituut voor Statistiek (NIS)

[18] L.U.C. – Steunpunt Verkeersveiligheid, K.U.L. – Spatial Appplications Division, U.C.L. – Département de Géographie.

First, we verified the content of all the sections and fields for all the traffic accident records. Hereby it is important not to consider all empty fields in the same way. In total, there were only 724 traffic accident records where all appropriate fields were filled in.

However, empty fields can occur for two different reasons. First, the value for the field was unknown or forgotten, implying that the field needs to be treated as a missing value. Or secondly, it could be that the field was non-applicable. However, it is important to make a clear distinction between these two situations, there a non-applicable field has to be treated differently than a missing value when the data is used for data mining and analysis techniques. A missing value needs more attention because, although the proper value is unknown, the real value for the field does exist and has an influence on the many interactions within and the structure of a traffic accident. Therefore, a missing value can not merely be equated to an empty field. It is important that this distinction is made between missing values and non-applicable fields when filling in the data.

2.2.2 Non-applicable fields

However, our research of the data showed that such distinction was not clearly made. Therefore, we could only assume in case of some fields, considering their meaning, that they were rather non-applicable than missing. The following table contains several fields where empty entries rather imply non-applicability than to be missing, although 100% certainty about this conclusion cannot be guaranteed.

Table 2-1: Fields which are non-applicable for a great extent of traffic accidents in the Flanders, Walloon and Brussels region

		% Non-applicable		
FIELDS		**Flanders Region**	**Walloon Region**	**Brussels Region**
13	Local characteristics	94,3	94,2	89,6
21	Varia	94,8 (73,1)	89,5 (66,7)	94 (82,2)
18a	Traffic accident factor – Road/Traffic	88,1	84,0	94,0
18b	Traffic accident factor – Road user	56,0	52,5	64,1
18c	Traffic accident factor – Vehicle	99,4	99,1	99,8
8b	Against obstacle	89,5	83,8	92,7
8	Type collision	44,0	41,0	45,1

Sometimes, there were also empty fields because their values could simply not be translated. For the field 21 (Varia) the possibility exists on the VOF form to write down the specific characteristics of the traffic accident as additional comments. However, these written comments are not available in our database. Therefore, the percentages between brackets represent the real percentage of being non-applicable.

2.2.3 Missing and inaccurate values

Another problem concerning the content of the dataset are the inaccuracies or missing values. Fields with inaccurate or missing values can not be merely considered as empty or non-applicable fields. These fields do have values, although they are not known. Table 2-2: Fields with missing values for traffic accident data in the Flanders, Walloon and Brussels region (in %) lists the fields with the highest percentage of missing values.

Table 2-2: Fields with missing values for traffic accident data in the Flanders, Walloon and Brussels region (in %)

		% Missing Values		
FIELDS		Flanders Region	Walloon Region	Brussels Region
R25	Pedestrian location	100	5,1	7,0
R5/6	House number in case of streetnames	50,3	38,8	28,0
R20	Bicycle path*	58,0	95,0	91,0
R19	Visibility pedestrian*	18,4	18,2	19,0
R17	Dynamics	42,5	43,2	46,0
R24	Country of registration*	12,1	10,0	7,2
R5/6	Kilometre marker on numbered roads	34,9	15,0	83,0
R16	Movement	8,6	8,6	11,8
R15	Sense of movement	11,3	12,0	17,0
R24	Condition	5,5	11,0	7,5
R24	Age	4,8	5,2	10,5

Remarkable about this table is the information about the position of passengers, which is missing for all traffic accident records in Flanders. These data were probably collected and registered, but must have been lost during data processing.

Another problem concerning the quality of the traffic accident data is the amount of accuracy applied when registering and processing the data. It can be concluded that information about the moment of the traffic accident shows quite some inaccuracy due to structural reasons, e.g. the exact moment of each traffic accident is rounded down to the nearest hour.

The most important remarks about inaccuracies and missing values relate to the information about the location of the traffic accident. The values of the *location* fields are to a great extent inadequate and don't allow automatic localisation of all traffic accidents. However, a significant distinction has to be made between the different regions and traffic accidents on road sections are less easily to localize than accidents on crossroads (1.2 Building the accident database).

2.2.4 Inconsistency

Missing values are not only and always indicated by empty fields. Sometimes, it is necessary to treat a field as a missing value, although that field does contain information. This is the case when some information about the traffic accident is registered several times in different fields and these fields show contradictions. These types of inconsistencies are as well present in an implicit way as in an explicit way in our dataset. Additionally, it is important to notice that the verification of the dataset on inconsistencies is an in-exhaustive process, which makes it difficult to assess the exact accuracy of our data. The inconsistencies detected during the data quality assessment shall be shortly discussed (cfr. Appendix F.3) in this section.

The number of fatalities, heavily and lightly wounded casualties and the casualties that passed away within a period of 30 days, are registered in section 23 as well as in section 24, 25 and 26. There are some differences among the totals, but these are of a negligible order. More striking are the differences among the number of pedestrians and (motor)cyclists when one compares section 24, 29/20 and 8. Another discrepancy exists among the number of obstacles registered in section 8a and section 8b.

The data also suggests an inconsistent filling-in of section 13, concerning the presence of public work sites. Great differences exist between the number of traffic accidents with public works in progress when one relies on section 8, section 18 or section 13. Some of these differences can be explained by the fact that section 13 represents the opinion of the police officer, while section 18 represents the opinion of the road user and section 8 is only used when there was a collision with the public works site. However, and this illustrates the inconsistency in section 13, only a part of the traffic accidents related to public works, mentioned in section 18 and 8, were actually registered within section 13.

Furthermore, the tables in appendix F.3 show that there were over 2500 people who where obviously drunk, but didn't have to pass an alcohol test. On inquiry of the federal police it appeared that many road users had to be transported to a hospital for first aid which made it impossible to perform an alcohol test. Therefore, in this case, these inconsistencies refer to missing values.

Other inconsistencies relate to roundabouts and the dynamics of the vehicles. 26%, 31% and 36% of traffic accidents that occurred on roundabouts in respectively the Flanders, the Brussels and the Walloon region are not as such registered in section 4. In a few cases it was even the matter that road users jumped the lights on a roundabout! With regard to the dynamics of a vehicle, the data suggested that in some cases a vehicle could be standing still and be in motion at the same time.

In some cases, there were also impossible values for several other fields. Especially the field *dynamics* contained a high amount of impossible (unknown) values (9%).

2.2.5 Validity

A final obstacle concerning the intrinsic quality of the traffic accident data is whether or not all traffic accidents are reported. The VOF form is only used for traffic accidents with casualties and fatalities. Therefore, traffic accidents with only material damage are not included in the dataset. It should be noted however, that some local authorities do collect these data, but this information is never collected at a national level.

Furthermore it is highly doubtful that all traffic accidents with injuries are registered. A research of "Adviesdienst Verkeer en Vervoer van Rijkswaterstaat" reported that only 60% of all traffic accidents, which concerned injuries who needed to be hospitalized, were eventually registered. In case when the traffic accident only concerned people who needed first aid at the hospital, but who weren't hospitalized, a merely 25% registration level was attained. Other research results (De Somer, 1993; Beaucourt, 1998) align with these finding.

2.2.6 Reliability

Apart from the intrinsic accuracy and consistency of the data, it is also important to have an eye for the reliability of the data. What follows is a non-exhaustive list of the most important criticism concerning the reliability of the data (cfr.Table 2-3). This list is based on our own data research and inquiries of experts. Each expert was asked to evaluate the various sections of the VOF form on its reliability and to underpin their conclusions.

Table 2-3: Main criticism on VOF form concerning reliability of the traffic accident data

Section	Criticism
2	The NIS code is filled in incorrectly in 20% of all traffic accidents.
5/6	In 25% of the cases; the kilometre marker is filled in as 0. Maximal allowed speed is sometimes interpreted as the speed of the vehicle.
8	Only a limited number of the road users involved in a multiple collision is actually registered.
14	It is often the case that the first road is entered twice.
15	Sense of direction is only in 20% of all traffic accidents filled in correctly.
17	In 20% of all the applicable cases, the values *5, 6, 7* or *8* are filled in. These values are unknown to the police.
19	The crossing distance between two protected locations is often interpreted as the distance between the pedestrian and the crossing-place.
20	This section doesn't give a clear picture of the type or sort of the cycle track
24	Only 40% of the involved drivers are submitted to an alcohol test at the spot.
24/25	The field *Consequences* is highly liable to subjectivity.

A final distinction about the reliability of the various sections and fields can be made. The elements of information that are part of the official police report are of a higher level of reliability than the elements of information that are part of the statements of the drivers at the moment of the traffic accident.

2.2.7 Conclusion

As well for the Flanders, the Walloon as the Brussels region, there has been collected a large amount of data during the last decade. In general, it can be stated that the data contains interesting and useful information about traffic accidents in Belgium for the period 1991-1999. However, the relatively high amount of missing values, the fact that no difference is made between non-applicable and missing fields, the questionable validity, the inaccurate location fields and the inconsistencies significantly lower the quality of the data. Except for the accuracy of the localization of the traffic accidents, the traffic accident data for the tree regions are of a comparable quality.

Therefore, all future research based on these data should take notion about the imperfectness of the data. Especially during the data transformation and pre-processing phase and when drawing conclusion based on empirical results forthcoming from these data, the researcher should be aware of the several remarks made in this data quality assessment.

Based on the results of this quality assessment of the Belgian traffic accident data, one can draw following points of interest for improving the data quality in the future. Firstly, each local authority can emphasize other tasks towards their personnel. This has an influence on the accuracy and completeness of the registration event. It could be for example useful to encourage local authorities to put more emphasises on the difference between non-applicable fields and missing values.

Furthermore, because the registration of the traffic accident occurs in different phases and is sometimes liable to subjectivity, the amount of inconsistencies within the data is rather trivial. Although there cannot be done much against the subjective nature of some parts of the registration, one could reduce many inconsistencies by the automation of the entire registration process. The automation would also allow a less rigid form for the registration of the traffic accidents.

One of the greatest inaccuracies within the data concerns the localization of the traffic accident. As is also recommended in the past, the police officers should have maps with indications of the kilometre marks on the numbered roads at their disposal.

A final remark can be made about the many data transformations performed on the traffic accident data. It is questionable if all these transformations are truly necessary there each transformation is susceptible to transformation errors, which only leads to a drop in data quality.

2.3. Relevance assessment of the traffic accident variables

2.3.1 Traffic accident data and association rules

While the previous research (cfr. supra) concentrated on the quality of the data, it was also necessary and part of our task to assess the quality of the information, concealed within the data, in order to describe and analyze traffic accidents. Therefore, at the same time as the data quality assessment, we also investigated the several variables available on their relevance in describing the traffic accident. In order to reveal which variables are relevant, the data mining technique of association rules was used to obtain a descriptive analysis of the accident data.

In contrast with predictive models, the strength of this algorithm lies within the identification of relevant variables that make a strong contribution towards a better understanding of the circumstances in which the accidents have occurred which, in turn, facilitates the definition of different accident types. Hereby, the emphasis will not only lie on the acquired accuracy of the generated patterns, but also on the interpretation of the results, which will be of high importance for improving traffic policies and ensuring traffic safety on the roads.

Association rules is a data mining technique which can be used to efficiently search for interesting information in large amounts of data. More specifically, the association algorithm produces a set of rules describing underlying patterns in the data. Informally, the support of an association rule indicates how frequent that rule occurs in the data. The higher the support of the rule, the more prevalent the rule is. Confidence is a measure of the reliability of an association rule. The higher the confidence of the rule, the more confident we are that the rule really uncovers the underlying relationships in the data.

Generating association rules involves looking for so-called *frequent itemsets* in the data. Indeed, the support of the rule $X \Rightarrow Y$ equals the frequency of the itemset $\{X, Y\}$. Thus by looking for frequent itemsets, we can determine the support of each rule (Mannila 1997). The problem of discovering association rules can therefore be decomposed into two sub-problems:

1. Generating all itemsets that have a support higher than the user-defined minimum support (minsup). These itemsets are called *frequent itemset*s.
2. Use this collection of frequent sets to generate the rules that have confidence higher than the user-defined minimum confidence (minconf).

2.3.2 Empirical study

This study is based on a large data set of traffic accidents obtained from the National Institute of Statistics (NIS) over a six year period (1991-1996) for the region of Brussels (Belgium). In total, 18.639 traffic accident records were available for analysis. Since our main interest in this study lies within the identification and profiling of geographical locations, with as much relevant information as possible, where a high number of accidents occur, this analysis will concentrate on the accidents that can easily be located by the hectometer mark. Selecting these records from the data set resulted in a total of 10.672 traffic accident records.

To explore association relationships between traffic accident attributes, only the traffic accidents that occurred at a high frequency accident location were selected for the analysis. To identify these locations, a criterion of minimum ten accidents per location was used. This resulted in a total of 1.110 traffic accident records that were included in the analysis.

Furthermore, in the present data set, some attributes have a continuous character. Discretization of these continuous attributes is necessary, since generating association rules requires a data set for which all attributes are discrete. Also the attributes with nominal values had to be transformed into attributes with binary attribute values. All these transformations and preprocessing resulted in a data set with 84 attributes, yielding a rich source of information on the different circumstances in which the accidents have occurred: course of the accident (type of collision, road users, injuries, …), traffic conditions (maximum speed, priority regulation, …), environmental conditions (weather, light conditions, time of the accident, …), road conditions (road surface, obstacles, …), human conditions (fatigue, alcohol, …) and geographical conditions (location, functional and physical characteristics, …).

A minimum support of 5 percent was chosen for the application of association rules analysis. A trial and error experiment indicated that setting the minimum support too low, leads to an exponential growth of the number of items in the frequent itemsets. Accordingly, the number of rules that will be generated will cause further research on these results to be impossible due to memory limitations. In contrast, by choosing a support parameter that is too high, the algorithm will only be capable of generating trivial rules. From this analysis, with a minsup = 5 percent and minconf = 30 percent, the algorithm obtained 101.861 frequent itemsets of maximum size 4 for which 313.663 association rules could be generated. These rules are further processed to select the most interesting rules.

A large subset of the generated rules set will be trivial. The purpose of post-processing the association rules set is to identify the subset of interesting (i.e., non-trivial) rules in a generated set of association rules. Two properties of association rules can be used to distinguish trivial from non-trivial rules. A first, more formal method (Brin et al. 1997) to assess the dependence between the two itemsets in the association rule is *interest*.

$$Interest = \frac{s(X \Rightarrow Y)}{s(X) * s(Y)}$$

The more this ratio differs from 1, the stronger the dependence. Table 2-4 illustrates the three possible outcomes for the interest measure and their associated interpretation for the dependence between the items in the antecedent and consequent of the rule.

Table 2-4: Interpretation of interest

Outcome	***Interpretation***
Interest> 1	Positive interdependence effects between *X* and *Y*
Interest = 1	Conditional independence between *X* and Y
Interest < 1	Negative independence effects between *X* and *Y*

A second method to define the interestingness of a rule is looking at the statistical rule significance (Silverstein, Brin and Motwani 1998). The statistical rule significance is determined using the χ^2- test for statistical independence and can be negative, neutral or positive. Table 2-5 gives an illustration for the possible outcomes of the statistical rule significance test (T) and indicates its relation with the interest of the rule.

Table 2-5: Interpretation of Statistical Rule Significance

Outcome	***Interpretation***
T <0	- Itemset X has a negative influence on the occurrences of itemset Y - Interest between 0 and 1 - Valid rule (T = -)
T is neutral	- Interest =1: X and Y are statistically independent → rule gives no extra information - Interest ≠1: rule has failed the χ^2- test → rule is not valid
T >0	- Itemset X has a positive influence on the occurrences of itemset Y - Interest >1 - Valid rule (T = +)

These rules were further post-processed by ranking them on their interest value and removing the rules that give no additional information towards this traffic accident analysis. An example of such a rule is:

Wet ⇒ Rain (sup= 19,91%, conf = 68,21%, T = +, I= 3,43)

2.3.3 Results

This paragraph will give an overview of the most important results from the association analysis. More specifically, five topics highlighting different aspects of traffic accidents will be discussed: collision with a pedestrian (section 3.3.1), collision in parallel (section 3.3.2), sideways collision (section 3.3.3), week/weekend accidents (section 3.3.4) and weather conditions (section 3.3.5). For each topic, the results will refer to the rule numbers (N) of the concerning rule table in which the rules are presented on the basis of a rank ordering of their interest value.

2.3.3.1 Collision with a Pedestrian

Table 2-6 illustrates that in 60,78% of all accidents involving pedestrians, collisions occur on crossroads with traffic lights (7). Moreover, accidents with pedestrians have a higher probability than expected of occurring at daylight (9), during the week (8) and in the afternoon (5).

Additionally, the results show that the pedestrian is often not coming unexpectedly from behind an obstacle through which he would not be visible at the moment of impact (1). In 49,02%, he crosses the road on a zebra crossing with traffic lights for pedestrians (2) and his walking distance between sheltered places will more than expected lie between six and 12 meter (3). This distance could relate to the length of the zebra crossing.

At first sight, these results may look surprising, since under these circumstances the pedestrian should be well visible for the other road users. Only in 13,07% of the accidents, the pedestrian will come from behind an obstacle through which he is not visible for the other road users at the moment of impact.

Moreover, only 4,5% of the collisions with a pedestrian occur while the pedestrian is crossing the street on a road with no zebra crossing, 13,72% while he is walking on a zebra crossing without traffic lights and 16,34% when he crosses the road walking next to a zebra crossing with traffic lights. A possible explanation for these results could be the great number of children that head for school, and therefore will be on the Belgian roads, around these times.

Furthermore, the rules show that collisions with pedestrians mainly occur on roads with just one roadway (6) and one road user is more frequently than expected moving upwards in the street whereas the other road user is moving transversal on this direction (4). The latter rule will probably relate to the walking direction of the pedestrian in relation to the moving direction of the driver since the Belgian Analysis Form for Traffic Accidents states that when the pedestrian is crossing the street while being involved in an accident, the pedestrian is moving in a transverse direction.

Table 2-6: Rules for collision with a pedestrian

N	SUP	CONF	T	I	BODY		HEAD
1	8,65	62,75	+	7,25	[pedestrian]	=>	[visible]
2	6,76	49,02	+	7,25	[pedestrian]	=>	[zebra crossing with traffic lights]
3	5,59	40,52	+	7,25	[pedestrian]	=>	[unsheltered walking distance between 6 and 12 meter]
4	5,50	39,87	+	3,16	[pedestrian]	=>	[one road user upwards, one transverse]
5	5,50	39,87	+	1,29	[pedestrian]	=>	[afternoon]
6	10,99	79,74	+	1,26	[pedestrian]	=>	[road with one roadway]
7	8,38	60,78	+	1,24	[pedestrian]	=>	[crossroad with traffic lights]
8	11,53	83,66	+	1,20	[pedestrian]	=>	[week]
9	10,00	72,55	+	1,19	[pedestrian]	=>	[daylight]
10	10,00	72,55	-	0,82	[pedestrian]	=>	[driving in a straight line]
11	8,11	58,82	-	0,75	[pedestrian]	=>	[constant speed]

Finally, a collision with a pedestrian occurs less often than expected in the presence of a road user that drives at a constant speed (11) or when at least one vehicle is driving in a straight direction (10). This arouses the suspicion that pedestrians will have a higher probability of getting hit by a vehicle when the road user is making a manoeuvre.

In conclusion, collisions with pedestrians will have a higher probability of occurring on crossroads with traffic lights, more specifically when the pedestrian is crossing the street on a zebra crossing with traffic lights, being well visible, at daylight, in the afternoon, during the week and when the road user is making a manoeuvre.

2.3.3.2 Collision in Parallel, Driving in the Same Direction

Table 2-7 shows that when an accident happens as a consequence of not respecting the distance between the different road users, the collision will almost inevitably take place between vehicles driving in the same direction (1). From the definition of the Belgian Analysis Form for Traffic Accidents, this type of accident usually relates to a collision at the back of a vehicle but it can also be a collision between vehicles driving next to each other following the same direction. Additionally, the rule stated above is also valid in the opposite case (2) and in 42,36% of the collisions in parallel, one of the road users will have used his brakes with the intention to stop (3). This type of accident will probably occur mostly in case of a collision at the back of a vehicle.

Furthermore, a collision in parallel will occur less frequently than expected when only two people are involved in the accident (7) and not respecting the distance between different road users will often lead to more than one collision (6).

Finally, a collision in parallel will less often than expected coincide with a road user driving at a constant speed (5) and will have a smaller probability than expected of happening at a crossroad (4).

To summarize, collisions in parallel will often be related with not respecting the distance between road users and with using the breaks with the intention to stop. However, this type of collision will have a smaller probability than expected of occurring on crossroads, at constant speed, with only two persons involved.

Table 2-7: Rules collision in parallel

N	SUP	CONF	T	I	BODY		HEAD
1	5,95	81,48	+	6,28	[distance]	=>	[parallel]
2	5,95	45,83	+	6,28	[parallel]	=>	[distance]
3	5,50	42,36	+	3,27	[parallel]	=>	[brake]
4	11,98	92,36	-	0,95	[parallel]	=>	[crossroad]
5	8,38	64,58	-	0,83	[parallel]	=>	[constant speed]
6	5,135	70,37	-	0,84	[distance]	=>	[one collision]
7	8,56	65,97	-	0,81	[parallel]	=>	[two people involved]

2.3.3.3 Sideways Collision

The rules in Table 2-8 indicate that when a sideways collision occurs, the road user will often not have respected the priority regulation of the crossroad (12). Most of the times he will also drive at a constant speed (15). These sideways collisions where the priority regulation of the crossroad is not respected have a higher probability than expected of happening on crossroads where the road users should give way to the vehicles coming from the right (7). When the priority on the crossroad is regulated by traffic lights, the sideways collision will often occur when a road user makes a left turn (5).

Table 2-8: Rules sideways collision

N	SUP	CONF	T	I	BODY		HEAD
1	10,54	43,82	+	3,02	[no priority]+[crossroad with priority to the right]	=>	[local road]
2	14,00	75,24	+	2,48	[crossroad with traffic lights]+[no priority]	=>	[left turn]
3	7,84	64,44	+	2,12	[crossroad with traffic lights]+[one road user upwards, one opposite]	=>	[left turn]
4	5,59	53,91	+	1,99	[no priority]+[local road]+[crossroad with priority to the right]	=>	[equal road functions]
5	14,41	52,63	+	1,73	[crossroad with traffic lights]+[sideways]	=>	[left turn]
6	21,53	70,92	+	1,45	[left turn]	=>	[crossroad with traffic lights]
7	19,46	49,43	+	1,43	[sideways]+[no priority]	=>	[crossroad with priority to the right]
8	5,77	80,00	+	1,34	[traffic lights]+[night with public lighting]	=>	[sideways]

9	24,05	69,35	+	1,34	[crossroad with priority to the right]	=>	[no priority]
10	20,36	67,06	+	1,29	[left turn]	=>	[no priority]
11	39,37	75,87	+	1,27	[no priority]	=>	[sideways]
12	39,37	66,01	+	1,27	[sideways]	=>	[no priority]
13	17,75	74,34	+	1,25	[one road user upwards, one downwards]	=>	[sideways]
14	21,35	70,33	+	1,18	[left turn]	=>	[sideways]
15	50,63	84,89	+	1,09	[sideways]	=>	[constant speed]
16	6,22	49,29	-	0,83	[one road user upwards, one transverse]	=>	[sideways]
17	5,95	48,53	-	0,80	[traffic lights]+[sideways]	=>	[daylight]
18	18,6	37,87	-	0,73	[crossroad with traffic lights]	=>	[no priority]
19	5,05	38,89	-	0,65	[break]	=>	[sideways]

These results refer to the relation between not respecting the priority regulation of the crossroad and the type of the priority regulation. An accident that occurs on a crossroad where the road users should give way to the vehicles coming from the right, often coincides with a road user that does not respect this priority regulation (9). An accident that occurs on a crossroad with traffic lights will on the contrary less frequently coincide with not respecting this priority regulation (18). In 75,24% of the accidents where this violation does occur with traffic lights, a road user will also have made a left turn (2). Moreover, 70,92% of all accidents that occur when a road user makes a left turn, take place on a crossroad with traffic lights (6).

Unfortunately, there is no information about which road user made the traffic violation, but it could be expected that the road user that turns left will not have respected the priority regulation. Additionally, not giving priority to the right has a higher probability than expected of occurring on crossroads where at least one of the roads is local (1) or where both of the roads have a local character (4). These results could indicate that the local character of a road could lead towards a misplaced feeling of traffic safety, whereas bigger, more important roads could enhance the concentration of the road users.

In general, 70,33% of the accidents where a road user turns left will lead to a sideways collision (14) and often when making a left turn, a priority violation will be the cause of the accident (10). Not respecting the priority regulation of the crossroad will lead in 75,87% of the accidents to a sideways collision (11).

Furthermore, when an accident occurs at night with public lighting and the road user approaches the traffic lights; the accident will often be a sideways collision (8). This type of collision near the traffic lights will less frequently occur at daylight (17). These results will probably relate to visibility that will be smaller at night.

A remarkable result is that when one road user is moving upwards in the street and another road user is moving in the opposite direction, the occurring accident will most of the times be a sideways collision (13). We would rather expect that this road situation would lead to a frontal collision. However, these accidents will occur on crossroads with traffic lights where the road users will drive in opposite directions and at least one of them will make a left turn (3).

Finally, when one of the road users uses his brakes with the intention to stop (19) or when one vehicle is driving upwards in the street and another vehicle is driving transversal on this direction (16) the accident will less frequently be a sideways collision.

In conclusion, there are two types of sideways collisions. The first type takes place at crossroads where road users should give priority to the right. These accidents will most of the times be caused by not respecting this priority regulation. The second type of sideways collisions occurs on crossroads with traffic lights. This type of accident will often be related with a road user making a left turn and will also frequently occur when the road users are moving in an opposite direction.

2.3.3.4 Week/Weekend Accidents

Weekend: from Friday 21 h. - Monday 6 h.
Morning: 6-11h ; afternoon: 12-16h.; evening: 17-23h.; night: 24-5h.

As shown inTable 2-9, most accidents that occur at night, will take place during the weekend (1). Moreover, the accidents that take place in the weekend will more often than expected occur at night with public lighting (2) and will less frequently occur at daylight (11). Similarly, the accidents that happen on Sunday will have a higher probability than expected of occurring at night with public lighting (3), in spite of the fact that on this day a lot of people will also be on the roads in the morning and in the afternoon, making so-called daytrips or family excursions.

However, accidents that happen at night do less frequently than expected coincide with a driver whose physical condition is normal (10). He will have a higher probability of being drunk, under the influence of drugs or just being exhausted or unwell.

In contrast, accidents that occur during the week with one road user driving upwards in the street and another road user driving in the transverse direction, usually take place at daylight (4). In general, accidents that occur at daylight will most of the times take place during the week (7).

As mentioned earlier, a collision with a pedestrian will also often occur during the week (5). Even more, accidents that happen during the week will have a higher probability than expected of occurring in the afternoon (7). The number of accidents that take place in the afternoon is accordingly smaller during the weekend than during the week (9).

Finally, 78,98% of the accidents that occur on crossroads where the crossing street is an important local road, take place during the week (6).

Table 2-9: Rules week/weekend accidents

N	SUP	CONF	T	I	BODY		HEAD
1	7,38	58,57	+	1,92	[night]	⇒	[weekend]
2	14,59	47,93	+	1,45	[weekend]	⇒	[night with public lighting]
3	5,766	45,07	+	1,36	[Sunday]	⇒	[night with public lighting]
4	8,018	80,91	+	1,33	[week]+[one road user upwards, one transverse]	⇒	[daylight]
5	11,53	83,66	+	1,2	[pedestrian]	⇒	[week]
6	11,17	78,98	+	1,14	[crossing important local road]	⇒	[week]
7	46,76	76,89	+	1,11	[daylight]	⇒	[week]
8	23,6	76,61	+	1,1	[afternoon]	⇒	[week]
9	23,6	33,94	+	1,1	[week]	⇒	[afternoon]
10	10,72	85	-	0,92	[night]	⇒	[normal physical condition]
11	14,05	46,15	-	0,76	[weekend]	⇒	[daylight]

To summarize, most accidents that occur during the weekend will take place at night and have a higher probability of occurring with a driver whose physical condition is not normal. Accidents that happen in daylight will more often occur during the week.

2.3.3.5 Weather Conditions

Table 2-10 illustrates that accidents on a wet road surface will have a higher probability than expected of occurring at night with public lighting (2) and a smaller probability of occurring at daylight (10). Accordingly, accidents that happen at night with public lighting will coincide more frequently than expected with a wet road surface (3) and less frequently with a dry road surface (8). Similarly, accidents that take place in the rain will have a higher probability of occurring at night with public lighting (1) and a smaller probability of occurring at daylight (11).

Furthermore, accidents that happen on a wet road surface (4), when it rains (6) or that occur at night with public lighting will less frequently coincide with a driver whose physical condition is normal. Moreover, accidents in the rain have a smaller probability of occurring with a driver of whom the alcohol test will be negative or not required (5).

Finally, when more than two people are lightly injured, the accident will less frequently than expected have occurred on a dry road surface (9).

In conclusion, accidents that happen in the rain or on a wet surface will more frequently occur at night with public lighting. These accidents will also have a higher probability of occurring with a driver whose physical condition is not normal and a smaller probability of coinciding with a driver of whom the alcohol test will be negative or not required (5).

Table 2-10: Rules weather conditions

N	SUP	CONF	T	I	BODY		HEAD
1	9,73	48,87	+	1,48	[rain]	=>	[night with public lighting]
2	13,15	45,06	+	1,36	[wet]	=>	[night with public lighting]
3	13,15	39,78	+	1,36	[night with public lighting]	=>	[wet]
4	25,77	88,27	-	0,96	[wet]	=>	[normal physical condition]
5	18,56	93,21	-	0,96	[rain]	=>	[no alcohol]
6	17,39	87,33	-	0,95	[rain]	=>	[normal physical condition]
7	28,65	86,65	-	0,94	[night with public lighting]	=>	[normal physical condition]
8	19,64	59,4	-	0,84	[night with public lighting]	=>	[dry]
9	5,135	57,58	-	0,82	[>2 lightly injured people]	=>	[dry]
10	14,41	49,38	-	0,81	[wet]	=>	[daylight]
11	8,919	44,8	-	0,74	[rain]	=>	[daylight]

2.3.4 Conclusions

The analysis showed that by generating association rules the identification of accident circumstances that frequently occur together is facilitated, leading to a strong contribution towards a better understanding of the occurrence of traffic accidents. The results indicate that the use of the association algorithm allows discerning different accident types, identifying different relevant accident conditions for each traffic accident type. For example, zebra crossings with traffic lights and pedestrian visibility are important aspects of pedestrian collisions, distance between the road users is an important aspect for collisions in parallel and priority to the right and making a left turn are the most important factors in sideways collisions.

2.4. Model based clustering on traffic accident data

2.4.1 Purpose

One of the main goals of the second research phase, performed by the LUC, was to develop a method for spatial clustering of traffic accidents. Hereby, the main issue was to maximize the similarity within the clusters and the dissimilarity between the different clusters.

Furthermore, when performing data mining techniques on traffic accident data, one has to keep in mind that road accidents – and the consequences of these accidents – are considered the result of a complex interplay between the driver, his vehicle and the road infrastructure. Therefore, to deal with this complexity, we decided to deploy an unsupervised traffic accident examination based on a broad clustering of all available attributes.

A large number of studies have already explored factors associated with accident occurrence and accident involvement. Nevertheless, each time these studies focus on a specific sample of the accident population or choose out some specific contributing attributes. However, these studies also caution for the interaction between the attributes. Therefore we decided to examine the traffic data in this research without any assumption or hypothesis on either the existence of some typical in advanced known group or of relationships between contributing factors.

The data we will deploy is derived from the National Institute for Statistics and represents a collection of the traffic accident records containing multivariate crash analysis features related to the accident type, the accident circumstances, the driver specifications and other potentially influential factors. Furthermore we incorporated a geographic factor, i.e. a black zone attribute, in the modelling process to search for the connectivity between spatial and multivariate crash analysis.

2.4.2 Cluster techniques and applicability

According to Kaufman and Rousseeuw (1990**),** cluster analysis is "the classification of similar objects into groups, where the number of groups, as well as their forms are unknown". Of all available cluster techniques, a model-based or latent class clustering analysis was preferred for this research for several reasons.

First of all, the more prevailing clustering techniques, as there are the partitioning techniques (e.g. K-means) or the hierarchical techniques (e.g. Ward) simply don't qualify well for this type of research for several reasons (cfr. Brijs et al. 2002). The main reason why these techniques are not really qualified is because these techniques are distance based, i.e. the assignment to a cluster is based on a distance measure or a similarity index. These measures or indices are heuristics which are not designed to compare categorical data. Besides, most of these traditional techniques require the number of clusters to be specified in advance, which is incompatible with our unsupervised knowledge discovery approach. The model-based clustering technique does not have these disadvantages.

Since the model-based clustering is based on a statistical approach, i.e. the observations are assumed to be generated from a mixture of underlying probability distributions, a number of statistical tests are available to check the validity of the model. Thus, latent class clustering allows for a statistical treatment of model selection and helps to determine the optimal number of accident clusters.

Besides the better match of this model-based clustering technique with our categorical data and our unsupervised learning approach, this technique also brings along other advantages. Where traditional clustering techniques assign a subject to just one cluster, the assignment of clusters within the model-based clustering is carried out in a probabilistic way. This makes it possible to use Bayes' rule to classify unseen observations into the identified clusters, since their values on the indicator variables can directly be used to compute their individual posterior class-membership probabilities. Recent advances in model-based clustering (Vermunt and Magidson, 2000) also make up another advantage, as they enable the inclusion of variables of mixed scale types (nominal, ordinal, continuous and count variables). Our accident attributes are predominantly nominal, but we also utilize ordinal and count variables**.**

The key idea in model-based clustering, also known as latent class clustering or finite mixture models, is that the observed data (in our case traffic accidents) are assumed to originate from a mixture of density distributions for which the parameters of the distribution and the size and number of the segments are unknown. Therefore, the objective of this model-based clustering is to unmix the distributions and to find the optimal values for the parameters and the number and size of the segments, given the underlying data. Attribute values of accidents belonging to the same class are assumed to come from the same density distribution, whose parameters are unknown and have to be estimated.

To determine the number of segments unsupervisedly, a so-called information criterion is used to evaluate the quality of a cluster solution. Basically, information criteria are goodness of fit measures, which take model parsimony into account. The idea is that the increase of the likelihood of the mixture model on the dataset (which results in a better fit), is penalized by the increase in the number of parameters that was needed to increase the fit. The smaller the criterion, the better the model in comparison with another. So we will augment the number of clusters and therefore the fit until the criterion allows us to. The criteria we employed are the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and the Consistent Akaike Information Criterion (CAIC) :

$\mathrm{BIC}_{\log L} = -2 \log L + (\log N) \times \mathrm{npar}$
$\mathrm{AIC}_{\log L} = -2 \log L + 2 \times \mathrm{npar}$
$\mathrm{CIAC}_{\log L} = -2 \log L + [(\log N) + 1] \times \mathrm{npar}$

where Log$L$ = log-likelihood
npar = number of parameters in the model
N = total number of cases

2.4.3 The traffic accident data

For this study the 1997-1999 accident records of two different regions were used. The first analysis considered the Walloon Brabant administrative region, which is a province located south of Brussels. This region is mainly characterised by urban sprawl, but also by the existence of some former small market towns like Nivelles, Braine-l'Alleud, Wavre or Jodoigne . The Eastern part is still quite rural, the western part more industrial. Limiting the extent of the studied areas enables one to better control for other sources of variations (mobility habits, friction of distance, mobility policies, etc).

The second analysis considered another administrative region, i.e. the Brussels Capital region. This administrative region, consisting of nineteen cities and municipalities among which the Belgian capital, covers an area of 161.4 km² and counts almost one million inhabitants. Furthermore, the Brussels Capital region covers 1,881 kilometers road which accounted for 3.18 billion vehicle kilometres travelled in 2002 (De Groote 2003, National Institute for Statistics, Belgian Federation of the Car and Two-wheeler Industries).

Although the database goes back to 1991 for both regions, it was chosen to apply the cluster algorithm on a more consistent three year period. This 3-year period is long enough to limit the influence of random fluctuations and short enough to embank evolutions in externalities. Furthermore, we focused on the accidents of a certain geographic level, i.e. accidents that happened on numbered roads. In Belgium, the numbered roads make up 11% of the road network but account for about half of the accidents. This pre-processing brought a significant advantage in return. This higher geographical level of accidents relishes - in Belgium - a higher localisation accuracy, which allowed us to add a spatial attribute, i.e. the black zone variable. For the first analysis on the Walloon Brabant region, this variable was defined by Flahaut et al. (2003) as a high number of accidents at the hectometre concerned and a high number of accidents at its neighbouring hectometres. For the second analysis, this variable was obtained from SADL. Furthermore, the first analysis was performed only on accidents where maximum two road users were involved, while for the second analysis only the two road user traffic accidents were retained. In total 1.991 traffic accidents records were available for an extensive cluster analysis on the Walloon Brabant region and 4028 traffic accidents were available for the study of the Brussels Capital region.

To cluster the traffic accidents, the nationally collected data first had to be pre-processed and accommodated. To deal with redundancy, missing values, the large amount of explanatory variables and the skewed character of the data, several variables were transformed. Some variables were discretized into new categories. Some new – often count – variables are an aggregation of different accident form rubrics or variables. Other variables (e.g. age) have been redivisioned into categories. After the transformations during this pre-processing, 32 variables (both categorical, ordinal and count) were retained for the cluster analysis. Table 5-1 in appendix F.4 lists the variables with their definition, data mode and the number of the categories. When marked with '*', the different categories can be found in table Table 5-2 (cfr. Appendix F.4). For the second application of latent class clustering, i.e. the study of the dataset of the Brussels Capital region, the same variables were used. However, some variables were, based on the experience of the first analysis, redefined. More specifically, some count variables were transformed in to categorical variables. The exact definitions for the variables of the second analysis can be found in Table 1 and Table 2 (cfr. Appendix F.4)

2.4.4 Empirical results for the Walloon Brabant region

The unsupervised categorical clustering for this administrative region, i.e. the Walloon Brabant region, resulted in 5 separated accident groups (k=5). The entropy R-squared classification statistic amounted to 93% (McLachlan and Basford, 1988). These statistic indicates how well the model predicts class memberships. Furthermore, some Information Criteria (IC) were needed to determine which model was superior. The IC correct the fit for the parsimony (i.e. number of parameters) of the model. These criteria guard against over fitting. The Bayesian IC, the Akaike IC and the Consistent Akaike IC are the three information criteria we used. When we added the sixth cluster, the three IC stopped declining and increased, indicating that the five cluster model was the optimal one. The five clusters will be discussed shortly.

Cluster 1: Time independent single-vehicle black zone crashes on highways

The first cluster is the largest cluster containing 25,9 % of the 1.991 accidents, but only 21,7 % of the 3.270 involved road users. This cluster (like cluster 2) consists mostly of collisions with off-road obstacles after a loss of control or an unexpected evading. The cluster percentage (cp) amounts 59% for this type of accidents while the average population percentage (pp) is 30 %. Besides these single vehicle crashes the cluster contains many rear-end collisions (cp=25% while pp=18%) and almost no frontal collisions at all. The accidents always took place on road segments (pp=61%), always outside the built-up area (pp=65%), and congruently in the first place on highways (cp=83 % while pp=28%). 38% of the accidents happened within concentrated zones, labelled as black zones (pp= 29%). Furthermore, very typical for the first (and also the second) cluster, is their time-independency. This group of accidents took place both in the week and the weekend, at night or during the day, at rush hours or not at rush hours and no matter which season. First road users drove cars (cp for first road user or cp1= 89% while pp for first road user or pp1=82%). If present, the second road user drove besides cars also trucks (cp2=18% while pp=5%). The cluster presents a younger first road user population. The third age category (18-21 years, pp=7%) and the fourth (22-29 years, pp=15%) are relatively stronger presented with respectively 9 % and 21 %. Only 28 % of the second road users are female (pp2=33%). The cluster shows a typical presence of passengers simultaneously in front and at the back of the vehicle. This is the only cluster where this combination is more frequent than the presence of passengers only at the back.

Cluster 2: Time independent single-vehicle crashes on regional roads outside built-up area

The accidents of the second cluster consist for 45% out collisions with off-road obstacles (pp=30%). The accidents occurred again only on road segments, still 69% outside the built-up area, but this time only on regional roads. Within this cluster only 18% of the accidents took place at black zones (pp=29%). We find almost the same time-independence as in the first clusters, though these accidents occur less in the summer (cp=20% while pp=23%). According to the time-independence, 37% of the accidents happened during the weekend (pp=32%). 83% of the first road users drove cars. This equals the population average. We find average presence of vans, trucks, motorbikes and mopeds. The second age category (15-17 years), the third (18-21 years) and the fourth (22-29 years) are more strongly represented among the first road users. Respectively 2% compared with 1% pp, 9% compared with 7% pp and 19% compared with 15% pp. Among the second road users we notice a cluster percentage of 17% concerning the 50-59 category (pp=11%). Only 25 % of the second road users are female (pp2=33%). Common infractions were: infractions on the right of way, place not in accordance, no distance kept. Furthermore the cluster accounts the highest registration of safety affecting personal conditions (e.g. alcohol, fatigue): 0,28 counts par person compared with a population average of 0,15.

Cluster 3: Lateral collisions on (mixed) crossroads on rainy week days with passengers involved

In this cluster accidents behave more time (or exposure) dependent. It contains only a small amount of single-vehicle crashes (12%). 57 % of the accidents were lateral collisions. The car is strongly present for both the first (89% while pp = 82%) and the second road user (84% while pp = 69%). All the accidents happened on crossroads and mostly on regional ways (cp=91%), but also on highways exit or entries (cp=9%). 30% of the crossroads showed difference on the maximum allowed speed between the roads (pp =20%). Common infractions were ignoring the right of way and ignoring the red light. 23% of the first road users was accelerating (pp=12%). The cluster has a very high passenger presence (45 % while pp = 25%) and is the most rainy cluster. It contains the highest percentage of accidents with precipitation, i.e. 60% (pp =41%).

Cluster 4: Lateral collisions with male two-wheelers involved

Also in this cluster accidents occur time or exposure dependently. Furthermore it exhibits a high afternoon accident frequency and a low weekend frequency (cp=22% and pp=32%). 53 % of the accidents are lateral collisions (pp 31%). The accidents took place on the regional roads, 60% occurred within the built-up area (pp=35%), mostly at crossroads (75% while pp=39%). It's a totally dry (98% while pp=61%) cluster with accidents occurring less in the winter (16% while pp=23%). The first road users of the cluster are only for 54% (pp=82%) made up by cars. In this group, there is a significant presence of motorbikes (16% while pp=4%), mopeds (15% while pp=4%) and bicycles (7% while pp=2%). Only 18% of them are women (pp=26%). The second road users are made up by the cars (86 % while pp=69%). Infractions on the right of way and the passing of a vehicle are the most common ones.

Cluster 5: Elderly (female) colliding with the vulnerable road users within the built-up area while accelerating

Cluster 5 is an even more time or exposure dependent accident group then cluster 3 or 4. Accidents strongly occur around rush hours and 78% of the accidents occurred during the week (pp=68%). The accidents happened on the regional ways, 75% of them within the built-up area, both on crossroads (48%) and road segments. 36% of the accidents happened within black zones (pp=29%). The first road users drove cars (88%) while the second road users consisted almost entirely out of vulnerable road users : 18% motorbikes, 21% mopeds, 13% bicycles, 33% pedestrians. The first road users appeared to be rather older drivers: 40-49 category (pp1=18%) accounted 23%, 50-59 category (pp1=9%) accounted 10%, the 60-65 category (pp1= 3 %) made up 6% and the 65+ category (pp1=5%) made up 9%. Furthermore, 34% of the first road users were female (pp1=26%). This cluster almost has no passengers involved. 80% of these road users drove around without passengers (pp=70%). The second road users strongly represented the younger categories: 12% were under the age of 15 (pp2=2%), 16% was 15 or 16 years old (pp2=4%) and was aged in between 18-21 (pp2=9%). Besides the lateral collisions (45%) the cluster is mostly made up by pedestrian collisions (33%). Furthermore there are no single vehicle crashes present and 40% of the first road users is characterised by an acceleration dynamic.

2.4.5 Empirical results for the Brussels Capital region

Originally, this model started with 31 variables (i.e. 17 variables at traffic accident level and 7 variables at road user level for each road user, cf. Appendix F.4). However, for some variables, the local independence assumption was violated. Therefore, five direct effects between these variables had to be integrated into the final model, which lead towards a better fit (cf. Table 3 in Appendix F.5).

The necessity to incorporate these direct effects imply that it was not possible to explain away the correlation among these variables by use of an "unknown" variable, i.e. traffic accident type. However, when examining the five direct effects, they all seem to be rather trivial.

Furthermore, three variables could be eliminated from the model, because they did not had any statistically significant effect on the model, i.e. the road users' conditions, the first road user's gender and the second road user's dynamics. Among others, this implies that for further research about the road users' conditions[19], a distinction between different traffic accident types should not necessarily be made.

[19] for the Brussels Capital region

The same applies for the first road user's gender and the second road user's dynamics. However, why in both cases this only applies for only one of the two road users, involved in the accident, is not clear. To be able to answer that question, a deeper knowledge about the intrinsic aspects of the traffic accident's registration process is indispensable.

Ultimately, this resulted in a final model with 28 indicators, and five direct effects, identifying seven different traffic accident types (cf. Table 4 in Appendix F.5). In the following, we will shortly discuss these traffic accident types and make some comparisons among them.

Cluster 1: Traffic accident with a pedestrian involved

This cluster discriminates itself from the other clusters there it exists for 99.21% out of traffic accidents with a pedestrian involved. Furthermore, for 99.99% of all traffic accidents within this cluster, the first road user was a pedestrian. Remarkably, this type of traffic accident accounts for at least 28.64% of all traffic accidents in the Brussels Capital region.

Additionally, when comparing the in-cluster distributions with the overall population distributions, some interesting facts are discovered. Obviously, the estimated probabilities show that there was no passenger in the vehicle present for 99.9% of the traffic accidents with a pedestrian involved. Furthermore, the results show that traffic accidents with pedestrians involved have a higher tendency to occur on a crossroad with traffic lights (35%) than the average traffic accident (19%)[20]. Also, it seems that for 73% of all traffic accidents with pedestrians involved, the pedestrian ignored a red light. Again, this probability is much higher than the probability that the first road user ignores a red light in an average traffic accident (48%).

Finally, it seems that for 24% of all traffic accidents with pedestrians involved, the pedestrian was not visible to the other road user.

Cluster 2: Light traffic accident with passengers involved

This cluster contains 19% of the studied population and distinguishes itself from other clusters on the number of passengers involved and their position within the vehicles. Based on the variable `passenger position', the conclusion can be drawn that for 94.3% of all traffic accidents in this cluster, there was at least one passenger involved in one of both cars. Except for the following cluster, the other clusters primarily exist out of traffic accidents without any passenger involved.

With regard to the severity of the traffic accidents, it seems that the traffic accidents within this cluster have rather light consequences. For 98.37% and 89.13% of all the traffic accidents within this cluster, the first, respectively the second, road user is uninjured.

Finally, when comparing with the average traffic accident, it seems that this type of traffic accidents concerns a lateral collision in 68% of all the cases, while this is only the case for 48% when considering all traffic accidents.

Cluster 3: Severe traffic accident with passengers involved

This cluster only contains 4.34% of the entire studied population traffic accidents. Like the previous cluster, this cluster distinguishes itself from the other clusters due to a high amount of traffic accidents

[20] For ease of reading, probabilities of the entire sample of traffic accidents are referred to as the probability of the average accident.

with passengers involved. However, the presence of passengers is less strong for this cluster than for the previous cluster, i.e. there are more traffic accidents without any passenger involved (26.16% versus 5.70%).

The main difference with the previous cluster is the severity of the traffic accidents. While for the previous cluster the consequences for the road user were mainly of a light nature, the opposite holds for this cluster. After all, the variable `detrimcount', which measures the severity of the traffic accident, is, with a value of 4.16 significantly higher for this traffic accident type than for the other traffic accident types.

Furthermore, another difference between the light traffic accidents with passengers involved and the severe traffic accidents with passengers involved is the distribution of the accidents over the seasons. Figure 1 in Appendix F.6 shows that the light traffic accidents primarily follow the distribution of the average traffic accidents with an equally distributed number of traffic accidents over the four seasons. However, according to figure Figure 1, it seems that the severe traffic accidents follow a different distribution with a clear peak during the fall and an above average number of traffic accidents during the winter.

Ultimately, the results also showed that this traffic accident type has a higher tendency to occur during the weekends than the average traffic accident (39% versus 24%).

Cluster 4: Traffic accident with two-wheelers (motorcycle, motorbike or bicycle) involved

This cluster contains 11.78% of all traffic accidents in the data set and distinguishes itself on the fact that for 96.09% of all the traffic accidents of this type, the second road user was a two-wheeler. The exact distribution is: a motorcycle in 31.46%, a motorbike in 29.50% and a bicycle in 35.33% of the cases.

Furthermore, when comparing the average traffic accident with traffic accidents of this type, two different conclusions can be drawn. Firstly, the conditional probability distributions of this cluster suggest that this type of traffic accident is most likely (74.55%) to concern a lateral collision. This implies that a two-wheeler has a 55.6% higher risk of getting hit from aside than the average road user (only 47.9% of the average traffic accidents concern a lateral collision).

Secondly, the results indicate that two-wheelers are weak road users. 94.63% of all two-wheelers involved in a traffic accident were lightly injured or worse, which is seriously higher than for the average traffic accident, where only 68.33% was injured or worse.

Cluster 5: Traffic accident on a road segment outside the built-up area against high speed

This cluster contains only 1,77% of the entire studied population and therefore has to be handled with the necessary caution when drawing any conclusions. Because of the small cluster size, the distributions can not be considered very stable. As far as the conclusions can be considered valid, the following unique properties of this cluster can be identified.

Firstly, it shows that this cluster mainly consists out of traffic accidents outside the built-up area (93.83%), while the other clusters mainly exist out of traffic accidents inside the built-up area. Furthermore, all traffic accidents within this cluster happened on locations where the maximum speed limit is 120 kilometres per hour. Furthermore, another striking difference with the other traffic accident types is that only 1.18% of all traffic accidents within this cluster occurred on a crossroad.

Considering this specific profile, the small cluster size can now be justified. After all, the Brussels Capital region consists almost completely out of built-up area and has very few kilometres road where a speed limit of 120 kilometres per hour prevails. Therefore, this cluster holds the necessary validity to consider traffic accidents on a road segment outside the built-up area as a distinct traffic accident type in the Brussels Capital region.

Furthermore, this traffic accident type seems to be more severe than the average traffic accident. The first and the second road user for this type of traffic accident have respectively a 92% and 42% higher risk to become lightly injured than the road users in an average traffic accident.

Finally, the estimated distributions also give an interesting insight in what might be important factors that cause this type of traffic accident. It shows that for 33.73% and 22.59% of all traffic accidents on a road segment outside the built-up area against high speed, the first road user respectively lost control of his vehicle or did not keep his distance. These percentages are much higher than for the average traffic accident, where only 4.57% and 3.48% of the first road users respectively lost control of his vehicle or did not keep his distance.

Cluster 6: Traffic accidents strongly related with failing to give right of way

This cluster contains 14.94% of all the studied traffic accidents, but is harder to discriminate in terms of traffic accident properties than the other clusters. Compared to the other clusters, there is one variable that really catches the eye, i.e. the behaviour of the first road user. For 91.44% of the traffic accidents within this cluster, the first road user failed to give right of way. However, the distinction with the other clusters is limited, there this is also the case for 41.67%, 38.38% and 33.36% of respectively the light and severe traffic accidents with passengers involved and the road accidents with two-wheelers. Therefore traffic accidents of this cluster can not be univocally identified as traffic accidents caused by failing to give right of way.

Furthermore, it seems that this type of traffic accident often concerns a lateral collision (81.27%) compared to the average traffic accident (47.90%). This confirms the fact that these collision are primarily caused by the first road user who fails to give right of way.

Finally, it seems that traffic accidents, where the first road user fails to give right of way, have a more severe impact on the second road user than in an average traffic accident. This shows from the fact that people who are involved in this type of traffic accident as the second road users, have a 40% higher risk of getting lightly injured than second road users in an average traffic accident.

Cluster 7: The dustbin traffic accident type

It is very hard to differentiate this cluster from the other clusters in terms of attributes. One could consider this cluster as a sort of dustbin for all the traffic accidents which do not belong to any of the other clusters. Actually, one could define this traffic accident type as the traffic accident with no pedestrian, passenger or two-wheeler involved, which occurred inside the built-up area and where the first road user did not fail to give right of way.

2.4.6 Conclusions

Traffic accident segmentation is indispensable when dealing with heterogeneous traffic accident data. When one does not make a distinction among the different traffic accident types present within the studied data set, the effects of certain factors may be obscured leading towards biased conclusions.

The segmentation technique used for this paper, i.e. latent class clustering analysis, has proved itself very useful as a descriptive analysis technique. This technique, which is well-known within the research field of market segmentation, contains several interesting advantages, e.g. its statistical basis and the flexibility to incorporate all kinds of probability distributions. Furthermore, because latent class clustering assumes local independency among the variables, one can already achieve a good understanding of the contributions of a specific factor to certain traffic accident types by just looking at the final model. However, the main strength of this technique is in its descriptive nature, and is not very well suited for use as explanatory technique.

2.4.6.1 The Walloon Brabant region

As for the first analyses on the traffic accident data for the Walloon Brabant region for the period 1997 until 1999, due to the rather small amount of accidents involved, the clusters are not fully homogeneous. But while examining the enormous amount of cluster specific attribute value distributions – originating from the 32 variables containing several categories – several interesting features and attribute relationships were discovered. However, the list of topics we will discuss is not exhaustively.

Black zones, Clusters and Accident Severity

The first and the fifth cluster contain much higher percentages of black zone accidents (respectively 38% and 36% while pp=29%). As both clusters differ totally, this fact points at the existence of typical classes within the black zones. A future focus on black zone accidents will have to consider on the one hand the highway zones with high concentrations of single-vehicle crashes and on the other hand it will have to take the typical accidents with vulnerable road users on the regional roads within the built-up area into account. Further research and mapping of the accidents and the subpopulations will make a more thorough differentiation possible. Moreover, a geographical comparison of the accident groups and the whole of black zone accidents would be interesting to examine the calculation of accident concentration indices.

When considering the severity of the clusters, highway accidents are known to be more severe than accidents on other roads (confirmed by the European Statistical Report (*15*) on road accidents). Though when we examine our accident subpopulations and we compare the fatal injuries of the first cluster (highway) with the second cluster (regional road), we noticed an average of 0,029 deaths for each accident of the first cluster and 0,041 within the second cluster. So the most severe accidents do not seem to happen on highways but find themselves at the regional roads of cluster 2. Although the name of the zones could hint otherwise, the presence of black zones is clearly not associated with the severity of the clusters. The most lethal cluster showed the lowest black zone concentration (only 18%). This means that these accidents happen dispersed over the vast amount of regional roads within the study area, which makes the cluster a very hard target for policy makers. Nowadays investments in infrastructure works to deal with the black zones in our country will have little effect on this lethal accident group.

The safety effect of passengers.

When studying the influence of passengers, Vollrath (2002) pointed out that passengers increase total safety of the vehicle. Their social control on the driver's behaviour generally seemed to prevail on the distraction they can bring about. Our data and more particularly the registered infractions of the passenger cluster (cluster 3) accord with these findings. A loss of control (13%) and certain infractions (e.g. wrongly passing (2%), disregarding a safe distance between cars (3%)) occurred less frequently in this cluster (respective pp = 35%, 3%, 4%). On the other hand, ignoring a red light or giving no right of way – possibly caused by distraction – seemed to occur more frequently (respectively 4 times more and 2 times more).

The former research also pointed out that the effect was reversed when young male drivers were involved. We think there is not only an interaction of the safety effect with age and gender but also with passenger formula and/or weather conditions. Having passengers both in front and at the back of the car simultaneously is a typical feature present in the first cluster and more specific within this clusters weekend accidents (22 of the 35 accidents with these typical passenger formula took place in the weekend). So this type of passenger formula appears to be a dangerous one in some circumstances. When considering the combination of passenger presence and weather conditions, cluster 3 exhibits a certain attribute-relationship between passengers presence, precipitation and lateral collisions.

According to the findings of Vollrath et al. (2002), the positive effect of passengers is decreased at crossroads. This cluster, which only consists out of crossroad accidents, brings the hypothesis that the positive effect of passengers could be even more decreased at crossroads when it rains, causing lateral collisions.

Age, Gender and Type of Accident

The relevance of our cluster will be made clear through this topic by checking if some existing findings can be confirmed within our population and its clusters.

In the light of the rapid increase of the number of older drivers in our developed countries, reflections on the oldest age categories gain more and more importance (see Maycock (*1997*)). The results of the Claret et al (*2003*) research in Spain suggest that the risk of causing a collision between vehicles with four or more wheels is directly dependent on the driver's age. They concluded that for both sexes, the risk increased significantly with age. Regarding sex differences, they presented among young drivers higher risk ratios for men than for women. When we accounted for the type of crash, as they suggested necessary for further research, our cluster model points out that these young male drivers are particularly involved in the first two single-vehicle crash clusters and rear end collisions on road segments. The (rather female) older drivers of cluster 5 are involved in lateral & pedestrian collisions within the built-up area. We had a closer look and crossed the age category with the type of accident. This confirmed the importance of the collision type when considering the influence of drivers age. Up to 60 % of the younger drivers categories were involved in single-vehicle crashes, while these only apply to 20% of the 65+ age category. These differences are even stronger within the population of female road users. While the middle-age road users are the ones strongly present within the rear end collisions, the elderly (65+) are more involved in lateral (40% of them) and frontal (19% of them) collisions. So to obtain a thorough view it appeared again necessary to account for a broader context e.g. the collision type. When examining the oldest age category within our cluster approach we note a high appearance in the third (30% of them) and the fifth cluster (25%). Claret (2003) and Hakamies-Blomqvist (1998) already suggested in the past that the aging feature is especially influential in complex situations when many stimuli must be processed. The profile of our 3rd and 5th cluster confirms the suggestions made: cluster 3 accidents happen mostly during rainfall at crossroads with mixed maximum speed limits. Cluster 5 also represents accidents who took place at moments of multitasking within the built-up area. Furthermore, in the context of the fifth cluster the findings that the male gender is associated with particularly high risk of death among pedestrians are confirmed. A crossing of consequences and gender showed that 10% of the male pedestrians died and only 5% of the female.

Our data does not confirm the Levine et al. (*1999*) findings, concerning the interaction between age and gender: younger women would have an overall 50% increased risk for fatal injuries compared with men. The youngest fatally injured female (first) road user in our data belongs to the age category 30-39. The youngest fatally injured second road user belonged to the sixth cluster 50-59. A larger amount of records will be needed for a more thorough research.

Li et al (*1998*) pointed out that female drivers were less likely to die in a car accident than male drivers, because women were involved in less serious accidents than men. The differences in fatal injury between man and women would be explained by behavioural differences between the sexes. Our data can confirm this : while 26% of the first road users were female, only 18% of the deceased first road users appeared to be female. But these results can differ strongly when examining the different clusters. None of the first road users of the fifth cluster, colliding with the vulnerable second road users within the built-up area died. 34% of this first road users were female (pp=26%). So this accident group helps confirming the above findings. These findings however are not confirmed in the first cluster: a proportional part of the fatally injured is female (28% while pp=29%). Female drivers here do not appear to behave differently. The different traffic subpopulations and accident clusters clearly have to be taken into account.

2.4.6.2 The Brussels Capital region

Furthermore, for the empirical results coming forth from the analyses of the traffic accident data for the Brussels Capital region, it is also possible to draw several interesting conclusions. However, it is not remarkable that these conclusions differ from the conclusions for the Walloon Brabant region there it concerns a different demographic and geographical region.

First of all, the results show that one should make a distinction of at least seven different traffic accident types in the Brussels Capital region, indicating a great deal of heterogeneity among the data. Furthermore, it seems that more than one out of four traffic accidents in the Brussels Capital region involves a pedestrian. Being the most frequent traffic accident type, one could legitimately claim that traffic accidents with pedestrians involved, demand a great deal of attention in the near future in order to increase traffic safety in the Brussels Capital region. Additionally, it seems that this type of traffic accident often occurs at a crossroad with traffic lights where the pedestrian ignores a red light. Finally, the results show that there were almost never any passengers present in the vehicle for this type of traffic accident, indicating that passengers might have a protective effect for this type of traffic accident. This finding is in accord with the research of Vollrath et al. (2002).

As for traffic accidents with passengers involved, two distinctive traffic accident types can be identified. It seems that most traffic accidents with passengers involved, primarily have non-severe consequences. Only 4.34% of all traffic accidents in the Brussels Capital region consists out of severe traffic accidents with passengers involved. However, for this type of traffic accident, an interesting finding shows up.

It seems that severe traffic accidents with passengers involved have a remarkable different distributions over the four seasons, compared to the other traffic accidents. These traffic accidents primarily tend to occur during fall or wintertime. However, the probability distributions of the visibility or environmental variable are not of such kind to explain these peaks during fall or wintertime. Therefore, further research on this topic is necessary to reveal the true relationship between the season and the severity of traffic accidents with passengers involved.

Furthermore, traffic accidents with two-wheelers involved, which account for 11.78% of all traffic accidents in the Brussels Capital region, are most likely to concern a lateral collision. Our findings also confirm that two-wheelers must be considered to be weak road users.

Additionally, it seems that the main causes for road traffic accidents on a road segment outside the built-up area against high speed are the loss of control of the vehicle or failing to keep enough distance. Finally, the final model also indicates that traffic accidents, when caused by a first road user who fails to give right of way, have a more severe impact on the health of the second road user than the average traffic accident.

However, as with all results arising from clustering techniques, one has to question the stability of these clusters. Therefore, further research is advised for both the Walloon Brabant region and the Brussels Capital region to see whether or not these traffic accident types are valid over the long run. Nevertheless, if one considers these traffic accidents to be valid, it is clear that the above conclusions form some interesting starting points for further research.

Furthermore, aside from the information these finding provide about the complexity of traffic accidents in the Brussels Capital region, they also indicate that it is important to acknowledge the heterogeneity of traffic accidents. Therefore, prior to any explanatory research on traffic accident data, it is indispensable to apply a traffic accident segmentation technique in order to acquire a deep understanding of the complexity of the traffic accident data in advance. In order to expose the heterogeneity, the latent class clustering technique proved itself very useful.

2.5. Understanding accidents in black zones using frequent item sets

2.5.1 The technique of frequent item sets

Finally, during the last part of the research, we tried to develop and apply a data mining technique in order to profile black spots. Therefore, the aim of this research phase was to analyse the characteristics of the accidents occurring in black zones compared to those scattered all over the road network.

Therefore, in this research, after having defined the location and the length of the black zones, data mining was applied for understanding the characteristics of the accidents associated to those spatial concentrations. In particular, the technique of frequent sets is used to identify accident circumstances that frequently occur together inside black zones. Furthermore, these patterns are compared with accident characteristics occurring outside those black zones. This allows investigating the differences between accident patterns inside and outside black zones, and hence to understand why spatial concentrations are observed.

The data mining algorithm, underlying the frequent item set technique, i.e. association analysis corresponds to the one used as basis for the relevance assessment of the traffic accident variables. Analogous with the previous research, this research also generates a large set of frequent item sets or rules from which a large subset can be considered trivial. Therefore, the interestingness measure of Brin et al. (1997) is used to assess the dependence between the items in the item set (cfr. Section 2.3:Relevance assessment of the traffic accident variables).

However, besides ranking the item sets on this interestingness measure, another measure was used to limit the accident patterns to only the discriminating or useful ones (Anand et al. (1997), Geurts et al. (2003)).

Definition: *Interest (I)*

$$I = \frac{S_b - S_n}{\max\{S_b, S_n\}}$$

This interestingness measure is based on the deviation in support values of the frequent item sets discovered for the accidents that occurred within a black zone from the accidents that occurred outside a black zone. The nominator S_b -S_n measures the difference in support for the accident characteristics in the black zones (S_b) and non black zones (S_n). The expression max $\{S_b, S_n\}$ is called the normalizing factor as it normalizes the interestingness measure onto the scale [-1, 1]. Since in this research we are mainly interested in profiling the black zones, we will pay special attention to the item sets with a positive interest value, i.e. approximating '1'.

2.5.2 The data set

Analogous with the first analyses of the study discussed in section 2.4, this research was also performed on traffic accident data of the Walloon Brabant administrative region, for the period from 1997 until 1999. These traffic accident data are obtained from the Belgian "Analysis Form for Traffic Accidents" that needs to be filled out by a police officer for each road accident that occurs on a public road and that involves casualties. Hence, this analysis is like all the other studies limited to accidents with casualties on numbered roads.

Furthermore, these data indicate that for the region of Walloon Brabant, 1.861 injury accidents occurred between 1997 en 1999. In these accidents, 81 persons were deadly injured, 333 seriously and 2.374 lightly injured.

Additionally, an initial analysis of these data indicated that traffic accident data are highly skewed. This means that some of the attributes will have an almost constant value for each of the accidents in the database. For example, 73% of the accidents in the dataset occurred under normal weather conditions. As was explained earlier, this will have no effect on the validity of the results since the association algorithm produces the lift value that corrects the importance of each rule by taking the frequency of the attributes in the dataset into account.

The definition of the black zones is here defined by a preceding study (Flahaut et al., 2003). Shortly said, the hectometre (100 meters) is considered as the smallest spatial unit for which road accident data are spatially available (stonemarkers on the numbered roads). We want to know if accidents are concentrated in space, if hectometres with large accidents records are scattered or clustered together. We therefore used the concept of local spatial autocorrelation. We know that spatial independence is an arrangement of accidents such that there are no spatial relationships between them. The intuitive concept is that the location of an accident is unrelated to the location of any other accident. The opposite condition - spatial autocorrelation – is an arrangement of accidents where the location of the hectometres is related to each other, that is they are not statistically independent of one another. In other words, spatial autocorrelation is a spatial arrangement where spatial independence has been violated (Levine, 2002). When accidents are clustered together, we refer to this arrangement as positive spatial autocorrelation. Conversely, an arrangement where accidents are dispersed is referred to as negative spatial autocorrelation. Global autocorrelation gives then a rough idea of the general spatial arrangements of accidents.

The Moran's *I* statistic is one of the oldest indicators of spatial autocorrelation (Moran, 1948). We here used a local index developed as a local indicator of spatial autocorrelation (LISA) by Anselin in 1995. It takes high values for hectometres located close to each others and having large numbers of accidents. Closeness is here measured in terms of distance measured on the road network. Sensitivity analyses were performed to the way distance is measured (Flahaut and Thomas, 2002); the technique of local Moran *I* has also been compared to kernel methods for defining road accidents black zones (Flahaut et al., 2003). Moreover, stability of the spatial structure put forward with Local Moran I has also been analysed over time and space: the locations of the black zones remain comparable from one year to the other (Eckhardt, Flahaut and Thomas, in press). These sensitivity analyses confirm the performance of the technique as well as a strong spatial structure of the road accidents in the Brabant Walloon area.

In the 1997-1999 period, 476 kilometres of black zones have been defined in the Walloon Brabant. These black zones concentrate 26,1 % of the total number of accidents on numbered roads. Selecting the accidents that occurred inside the black zones results in a total of 553 road accidents. The second dataset, containing the accidents that took place outside a black zone involves 1.287 road accidents. (note: for 21 accidents, the belonging or not to a black zone could not be defined).

2.5.3 Conclusions

2.5.3.1 Frequent Item Sets and Accident Analysis

In this paper, the association algorithm was used on a data set of road accidents to profile black zones in terms of accident related data and location characteristics. More specifically, frequent item sets are generated to identify accident circumstances that frequently occur together in order to find out which factors explain the occurrence of the accidents in black zones. As explained in the introduction, the use of this technique coincides with the explorative character of this research since it describes the co-occurrence of accident circumstances and gives direction to more profound research on the causes of these accident patterns and explanation. These patterns represent interesting interactions in accident factors which accordingly can be used to test in statistical models. Furthermore, the use of frequent item sets not only allows to give a descriptive analysis of accident patterns inside black zones, it also creates the possibility to find the accident characteristics that are frequent for all accidents but that occur more frequently inside than outside black zones.

2.5.3.2 Accidents Patterns in Black Zones

The most important result of this research is that road accidents concentrations in black zones correspond to specific frequent items. Taking a left turn is an important accident factor as well inside as outside black zones. However, in black zones, these accidents frequently take place on intersections with traffic lights while outside black zones, this accident type frequently occurs on intersections with traffic signs which could be explained by the traffic on these accident locations. Based on the results of an in-depth analysis of left-turn collisions by Larsen and Klines (2002), we could conclude that better signalised intersections could be a short term solution for this type of accidents. In the long term, one should consider the use of roundabouts to increase traffic safety in these black zones.

A second important accident circumstance as well inside as outside a black zone is the rainy weather conditions. This is partly due to the slippery roads (Brodsky and Hakkert, 1988) but also to the risk taking behaviour of drivers (Edwards, 1996). However, inside black zones this factor frequently coincides with aquaplaning, which is not the case outside these black zones. These results suggest that black zones and non black zones are characterised by different infrastructure specifications, explaining the occurrence of the clustering of accidents in black zones.

Furthermore, a collision with a pedestrian involving young road users inside the built up area is a typical accident pattern that frequently occurs inside a black zone. This confirms former papers showing that pedestrian injury collisions often occur when and where large numbers of pedestrians travel within complex roadway systems with high traffic flows. Education and environmental prevention efforts should hence focus on aspects of traffic flow, local neighbourhood as well as raising community awareness about the risks associated with them.

Additionally, loss of control over the steering wheel and the resulting collision with a crash barrier is a frequently occurring accident pattern in black zones. These run-off-roadway accidents often occur on freeways and are related to an inadequacy of the speed and/or behaviour of the user to the driving circumstances. The problem is then to identify cost-effective countermeasures that improve highway designs by reducing the probability of vehicles leaving the roadway and the severity of accidents when the do (roadside features).

In conclusion, the findings of this paper are rather suggestive but limited in that they are based on one data set. They show the usefulness of the frequent item sets in analysing the combination of rules associated with road accidents occurrences in black zones. This exploratory technique enables to prepare multivariate explanatory statistical analyses. Our results show that a special traffic policy towards accidents in black zones and accidents outside these zones should be considered. Indeed, these spatial concentrations of accidents are characterized by specific accident circumstances, which require different countermeasures to reduce their number such as improvements in terms of road design, signalisation, and local environment. Accordingly, infrastructure and land-use can enhance traffic safety but is not an answer to all problems. Finally, one should also mention that there is no unique combination of characteristics associated to road accident occurrences: it is a complex phenomenon for which only some aspects are reported here.

3 UCL

3.1. The spatio-temporal distribution of road traffic accidents in a periurban environment. The case of Brussels

The full paper is located in **appendix G.**

The aim of this paper is to analyze the spatio-temporal variations in the spatial concentrations of traffic accidents in a periurban environment over a nine-year period (1991-1999). Concentrations of accidents, or black zones, have been identified on the basis of local indices of spatial autocorrelation, whose value in accident studies is firmly established.

The analyses have been conducted on numbered roads that have been divided into 100 metre sections in a Belgian province (Walloon Brabant) which is predominantly periurban. This province extends southwards from Brussels and contains 460.4 km of numbered roads. Between 1991 and 1999, 6,905 personal injury accidents were recorded on these roads. Of these, 1,305 could not be localized with precision, which reduced the number of accidents which can be analyzed to 5,600. These 5,600 accidents are distributed over 244.4 km of road (i.e. 53% of the network). The black zones contain 29.5 % of the accidents which occurred between 1991 and 1999 and account for 20% of the total length of numbered roads.

It is important to discover whether the black zones have a marked and stable spatial structure over time. Before attempting to construct an explanatory model for accident concentrations — which is a project for the future — we propose a preliminary exploratory analysis which aims to test the following hypothesis: *the geographic distribution of black zones varies as a function of the selected period of time.* If this hypothesis is rejected it is highly likely that there is a strong environmental and spatial link, that accidents are always concentrated at the same places and that the results of the explanatory statistical model are independent of the period of time that is chosen.

For each of the pairs of three-year analysis periods which we wish to compare (1991 to 1993 and 1994 to 1996 on the one hand and 1994 to 1996 and 1997 to 1999 on the other), we have therefore drawn up contingency tables by comparing the number of accidents located in a black zone in no period, in one period or in both periods (Table 1). We have then compared these periods using a range of similarity indices.

The smallest spatial unit of road length for which accident data are available on Belgian numbered roads is one hundred metres. We have examined the spatial change in the binary variable of belonging to a black zone (1 if the one hundred metre zone belongs to a black zone, otherwise 0) during three successive time periods and have used a variety of indices and statistical similarity measurements to do this. The indices measure the statistical similarity between the periods by combining, in a variety of ways, the number of hundred metre sections belonging to no black zone (double absence), the number belonging to a black zone in the first period, but not in the second and vice-versa (difference), and lastly the number of hundred metre sections belonging to a black zone in both periods (double presence).

One of the problems we encountered is related to the disproportionality generated by the large number of hundred metre sections which do not belong to a black zone. However, the apparently straightforward problem of comparing maps at different times is linked with the difficulty of measuring only black hundred metre sections during two periods of time: the presence of a hundred metre section in a black zone at a given period is of more interest for analysis than its absence.

Because of this, several indices had to be rejected because they gave inconclusive results. The indices which were finally selected are those which concentrate on the hundred metre sections that are observed to belong to a black zone for two successive periods (Table 2) and which provide conclusive results.

To begin with, we calculated these indices for all of the black zones. We then measured their stability for subsets of black zones, which were formed on the basis of physical and environmental characteristics such as type of road (motorways and ring roads on the one hand, trunk roads on the other), traffic intensity, the physical characteristics of roads (roads with a central reservation, roads with two or more lanes but without a central reservation) and the environment (dense urban, sparse urban, wooded rural, open country).

The indices highlight stability among the black zones, in particular on motorways and a degree of instability which it may or may not be possible to explain, on trunk roads. The major roads in the studied region consist of radial roads (North-South) towards the Brussels region. The transverse routes (East-West) have fewer accidents and less traffic.

The characteristics of the environment and infrastructure can explain a certain amount of stability in the location of accident zones. In general, the average level of stability between two successive periods is 40%. If we consider environmental factors, the highest stability (between 50 and 60%) is a result of the following characteristics:

– on motorways and ring roads (1994-1996/1996-1999);

– for traffic in excess of 17,300 vehicles (1994-1996/1997-1999);

– for roads with a central reservation (1994-1996/1997-1999) particularly when the traffic level exceeds 24,000 vehicles;

– for roads without central reservations and with traffic of more than 9,400 vehicles (1991-1993/1994-1996);

– in a dense urban environment (1991-1993/1994-1996);

– in open country.

These results confirm the trends that we have observed on the maps (Fig. 2). The major roads form an influential spatial structure for the two last periods (1994-1996 and 1997-1999). In dense urban areas and on secondary roads, the similarities — and hence the stability — are more marked between the periods 1991-1993 and 1994-1996: the road improvements of recent years appear to have had safety benefits.

A proportion of the hundred metre sections on the network which belong to the black zones (40%) is therefore stable and constitutes a *hard core* for which it has not yet been possible to improve road safety through improvements to either geometric design or road signing. Another proportion is more mobile, no doubt as a result of the random nature of accident occurrence. We can therefore reject the hypothesis that *the geographical distribution of black zones varies as a function of the period time selected.* This temporal stability of black zones provides confirmation for the idea that the selection of a year is of little importance in geographical studies which deal with road traffic accidents, on condition that the environmental parameters remain stable. It seems highly likely that there is a strong environmental and spatial link and that at all times accidents tend to be concentrated in the same places.

3.2. Spatial nested scales for road accidents in the periphery of Brussels

The full paper is located in **appendix H**.

Abstract

This paper examines, by means of a multilevel model (MLM), how far the characteristics of the geographical environment influence the occurrence of road accidents at two levels of spatial aggregation. The results are compared to those obtained from a more classical logistic regression. The analysis is performed on data from the southern periphery of Brussels (Belgium). The main findings are: (1) that MLM is a potentially useful technique for modelling road accidents, but that hierarchical levels are not easy to define for spatial data and so MLM are less useful than other regression techniques; (2) that the characteristics of the environment and the road itself significantly influence the occurrence of road accidents, and changes in these characteristics are quite important elements in the explanation, leading to the suggestion that road users do not adapt their behaviour sufficiently to changes in road conditions. Hence, concentrations of road accidents often correspond to places where improvements could be made in terms of road design, signalling and land-use planning.

4 Overall conclusion

Three very different approaches were explored to gain a better understanding of road accident occurrence: remote sensing and spatial analysis, data mining and multilevel analysis. Although these researches have produced new techniques and conclusions on traffic safety, the interaction between them was not as intensive as expected. One of the reasons is the fact that the three research groups were focussed on different study areas and hence interaction was not possible at all. Also, due to several problems in the beginning of the project concerning the remote sensing analysis, it was not feasible to automatically derive new traffic indicators, which could have been a source of input for the subsequent research. However during the research many paths were explored and data and results were produced. The major conclusions are now presented:

Belgian traffic accident data

Building the attribute and spatial accident database (KUL)

Both the attribute and spatial database were successfully constructed for the three regions in Belgium: Flanders and Walloon regions and Brussels. The accident attributes were imported into a relational database from raw text files. Next, for the Flanders and Walloon regions, accidents on numbered roads were located while for Brussels accidents on all roads were located. Due to several missing values and inconsistencies not all accidents on numbered roads could be located (about 80%) but this problem is gradually being corrected by the regional authorities. The situation in Brussels is worse, only 60% of all accidents could be located because. Misspelled street names and lacking house numbers are the main causes of the rather low share of located accidents. These rather low percentages are due to the important inaccuracies within the data concerning the localization of the accident. As is also recommended in the past, the police officers should have maps with indications of the kilometre marks on the numbered roads at their disposal.

Quality assessment of Belgian traffic data (LUC)

In general, it can be stated that the accident data contains interesting and useful information about traffic accidents in Belgium for the period 1991-1999. However, the relatively high amount of missing values, the fact that no difference is made between non-applicable and missing fields, the questionable validity, the inaccurate location fields and the inconsistencies significantly lower the quality of the data. Except for the accuracy of the localization of the traffic accidents, the traffic accident data for the tree regions are of a comparable quality.

Based on the results of this quality assessment of the Belgian traffic accident data, one can draw following points of interest for improving the data quality in the future. Firstly, each local authority can emphasize other tasks towards their personnel. This has an influence on the accuracy and completeness of the registration event. It could be for example useful to encourage local authorities to put more emphasises on the difference between non-applicable fields and missing values.

Furthermore, because the registration of the traffic accident occurs in different phases and is sometimes liable to subjectivity, one could reduce many inconsistencies by the automation of the entire registration process.

Exploration of the accident attribute data (LUC)

Relevance assessment of Belgian traffic data

It was also relevant to assess the quality of the information, concealed within the data, in order to describe and analyze traffic accidents. In order to reveal which variables are relevant, the data mining technique of *association rules* was used to obtain a descriptive analysis of the accident data. This study is based on a data set of traffic accidents on high frequency locations (spots with a minimum accident frequency of 10) covering a six year period (1991-1996) for the region of Brussels (Belgium). After preprocessing the data set consisted of accident records with 84 attributes, yielding a rich source of information on the different circumstances in which the accidents have occurred: course of the accident, traffic conditions, environmental conditions, road conditions, human conditions and geographical conditions.

These results indicate that the use of the association algorithm allows discerning different accident types, identifying different relevant accident conditions for each traffic accident type. For example, zebra crossings with traffic lights and pedestrian visibility are important aspects of pedestrian collisions, distance between the road users is an important aspect for collisions in parallel and priority to the right and making a left turn are the most important factors in sideways collisions.

Model based clustering of accidents

One of the main goals of the LUC research was to develop a method for clustering of traffic accidents. To deal with the complexity, an unsupervised traffic accident examination based on a broad clustering of all available attributes was chosen. Of all available cluster techniques, a model-based or latent class clustering analysis was preferred. The clustering was done on both the city of Brussels as on the peri-urban region of Walloon Brabant, for the period of 1997-1999. For both regions, an extra spatial attribute 'belongs to a black zone' was available. Accidents in Walloon Brabant were structured into 5 clusters while the situation in Brussels was optimal with 7 clusters. Several conclusions can be drawn after inspection of the accident characteristics of the clusters.

This cluster analysis of accidents indicates that it is important to acknowledge the heterogeneity of traffic accidents. Therefore, prior to any explanatory research on traffic accident data, it is indispensable to apply a traffic accident segmentation technique in order to acquire a deep understanding of the complexity of the traffic accident data in advance.

Association rules for profiling black zones

During the last part of the research, we tried to develop and apply a data mining technique in order to profile black spots. Therefore, in this research, after having defined the location and the length of the black zones, data mining was applied for understanding the characteristics of the accidents associated to those spatial concentrations. In particular, the technique of association rules, which corresponds to the one used as basis for the relevance assessment of the traffic accident variables, is used to identify accident circumstances that frequently occur together inside black zones.

Analogous with the model based clustering analysis, this research was also performed on traffic accident data of the Walloon Brabant administrative region, for the period from 1997 until 1999 and the same definition of black zones was used.

The most important result of this research is that road accidents concentrations in black zones correspond to specific association rules. Here are a few examples:

- Taking a left turn on intersections with traffic lights is a typical pattern for accidents inside black zones while outside black zones, the left turn manoeuvre on intersections with traffic signs is typical.
- Rainy weather conditions are important but inside black zones, they are frequently associated with aquaplaning what suggest that black zones and non black zones are characterised by different infrastructure specifications.

Remote sensing (KUL)

For several technical reasons and because of the impossibility of monitoring traffic during the course of day with Ikonos images, the idea of extracting traffic safety indicators and traffic volume out of the image was abandoned and replace with the construction of a land use map. In order to make a land use classification, some pre-processing had to be done. After geometric rectification, Adaptive Image Fusion was used to fuse the bands into an input image which was an optimal input for the land use classification. Land use classification requires a more complex approach than just land cover classification, which can fairly easy done by means of multispectral classification. In order to classify land use (e.g. residential area, industrial area) an external layer with land parcels was overlaid on the land cover image and several parameters were calculated for each parcel. Statistical analysis of this information allowed classification of the parcels into land use classes. Although the methodology seemed quite promising the overall accuracy is low (about 75%) with the specific urban land use classes being the least well classified. However, this does not mean that the technique is wrong because much depends on the quality of the image and the initial land cover map.

Spatial Analysis

Two Dimensional black zones in Brussels, a highly urbanized area (KUL)

The known technique of kernel density maps, which provide a good visualisation of the structure of a point pattern, was enhanced to delimit statistically significant two-dimensional black zones. To solve this problem, a significant threshold to delimit significant accident concentrations was needed. Therefore, instead of calculating densities, we calculated probabilities which allow delimitation of significant black zones. Besides accident frequency another interesting concept, accident risk or the ratio of accident frequency and accident exposure, can shed light on traffic safety. Once the spatial distribution of exposure (e.g. traffic volume) is known in the study area, the risk can be mapped the same way with risk black zones as output. Traffic volume was extrapolated for the whole road network of Brussels based on a relation between traffic volume and the function of roads, which was calibrated with real traffic measurements in Brussels.

Modelling of accident frequency at neighbourhood level in Brussels (KUL)

An attempt was made to predict neighbourhood accident frequency as a function of accident exposure and socio-economic factors; the idea behind this test was to measure the effect of the socio-economic neighbourhood characteristics on accident frequency. The regression of accident frequency requires specific regression techniques because accident frequency is a count variable and it is a spatial phenomenon requiring spatial regression techniques. The results show that accident exposure is very important in explaining accident frequency but the socio-economic factors can't improve the predictive power of the model. Thus, the effect of the neighbourhood, as we measured it, on the accident frequency is in fact inexistent. The fact we found several large neighbourhoods in our black zone analysis might not be the result of causes at neighbourhood level but it might be related to the road infrastructure or traffic flows with similar characteristics in that neighbourhood.

The spatio-temporal distribution of road traffic accidents in a periurban environment: the case of Brussels (UCL)

The aim of this analysis is to analyze spatio-temporal variations in the spatial concentrations of road accident in a periurban environment. Analyses are conducted for the numbered roads in a Belgian province (Walloon Brabant) over a nine year period (1991-1999). Initially, we identify the concentrations of road accidents (or black zones) by means of spatial autocorrelation. Their locations during three three-year study periods are then compared by means of contingency tables and similarity indices. The research reveals a considerable similarity between the locations of black zones and suggests the existence of a powerful spatial structure. The findings suggest that an understanding of space is necessary in accident studies in order to identify more clearly its causal role and in order to specify local accident reduction measures (road improvements, local signing to modify driver behaviour, etc.)

Multilevel modelling of road accidents in a peri-urban region: Walloon Brabant (UCL)

This research examines, by means of a multilevel model (MLM), how far the characteristics of the geographical environment influence the occurrence of road accidents at two levels of spatial aggregation. The results are compared to those obtained from a more classical logistic regression. The analysis is performed on data from the southern periphery of Brussels (Belgium). The main findings are: (1) that MLM is a potentially useful technique for modelling road accidents, but that hierarchical levels are not easy to define for spatial data and so MLM are less useful than other regression techniques; (2) that the characteristics of the environment and the road itself significantly influence the occurrence of road accidents, and changes in these characteristics are quite important elements in the explanation, leading to the suggestion that road users do not adapt their behaviour sufficiently to changes in road conditions. Hence, concentrations of road accidents often correspond to places where improvements could be made in terms of road design, signalling and land-use planning.

5 Bibliography

5.1. SADL KUL R&D

Anselin, L., Bera, A.K., Florax, R., Yoon, M.J., 1996. Simple diagnostic tests for spatial dependence In: Region Science and Urban Economics. 26, 77–104.

Anselin, Luc, and Anil Bera. 1998. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics." Chapter 7 (pp. 237-289) in Aman Ullah and David Giles (eds.) *Handbook of Applied Economic Statistics* (New York: Marcel Dekker).

Aplin P., Atkinson P.M., Curran P.J. (1999) Fine spatial resolution satellite sensors for the next decade. In: International journal of remote sensing Vol. 18, pp. 3873-3881

Bailey T, Gatrell AC (1995) Interactive Spatial Data Analysis. Longman Scientific & Technical, Essex, England

Baltsavias, Pateraki, Zhang (2001), "Radiometric and geometric evaluation of Ikonos Geo images and their use for 3D building modelling", Proc. Joint ISPRS Workshop "High Resolution Mapping from Space 2001", Hannover

Barnsley M.J. & Barr S.L. (1996) Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. In: Photogrammetric Engineering and Remote Sensing, Vol. 62(8), pp. 949-958

Barnsley, M.J. & Barr S.L. (1997) A graph-based structural pattern recognition system to infer land use from fine spatial resolution land cover data. In: Computers, environment and urban systems, Vol. 21(3/4), pp.20—225

Bauer T. & Steinnocher K. (2001) Per-parcel land use classification in urban areas applying a rule-based technique. In: GeoBIT Vol. 6, pp. 24-27

Besag J. & Newell J. (1991) "The detection of clusters in rare diseases", Journal of the Royal Statistical Society, A 154, Part 1, 143-55

Blaschke T. & Strobl J. (2001) What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. In: GeoBIT Vol. 6, pp. 12-17

Blaschke T., Lang, S., Lorup, E., Strobl, J., Zeil, P. (2000) Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. In: Cremers, A., Greve, K. (2000) (eds.) Environmental information for planning, politics and the public, Metropolis Verlag, Marburg, Volume II

Cressie N. (1993) "Statistics for spatial data", John Wiley and sons

Dial, Gibson, Poulsen (2001) "Ikonos satellite imagery and its use in automated road extraction" in: Baltsavias, Gruen, Van Gool (Eds.), Automated extraction of mand made objects from Aerial and space images. Balkema publishers, Lisse pp. 357-367

Dufays T., Flahaut B., Steenberghen T., Thomas I. (2004) "Intra-urban location and clustering of road accidents using GIS: a Belgian example, International Journal of Geographical Information Systems, Vol 18, pp. 169-181

Flahaut, B., Mouchart, M, San Martin E, Thomas I (2003) "The local spatial autocorrelation and the kernel method for identifying black zones" Accident analysis & prevention, Vol. 35 pp.991-1004

Fotheringham S.A., Zhan F.B. (1996) "A comparison of three exploratory methods for cluster detection in spatial point patterns" Geographical Analysis 28:200–218

Fraser, Baltsavias, Gruen (2002) "Processing Ikonos imagery for submetre 3D positioning and building extraction", Journal of Photogrammetry & Remote Sensing pp.177-194

Fraser, C.S., Baltsavias, E., Gruen, A. (2002) Processing of Ikonos imagery for submetre 3D positioning and building extraction. In: Photogrammetry and remote sensing, Vol. 56, pp. 177-194

Jacobs, D. & Swyngedouw, M. (2000) 'Een nieuwe blik op achtergestelde buurten in het Brussels Hoofdstedelijk Gewest', In: Tijdschrift voor Sociologie, 21 (3): 197-228.

Hoffmann, A., Smith, G., Hese, S., Lehman, F. (2000) Die Klassifizierung hochauflösender Daten: ein per-parcel-Ansatz mit Daten des digitalen Kamerasystems HRSC-A. In: Vorträge der 19. Jahrestagung der DGPF, Band 8

Kayitakire, F., Farcy, C., Defourny, P. (2002) Ikonos-2 imagery potential for forest stands mapping. Presented at ForestSAT symposium, Heriot Watt University, Edinburgh

Kelsall J.E., Diggle P.J. (1995) "non-parametric estimation of spatial variation in relative risk", Statistics in Medicine, Vol.14 p.2335-2342

Kulldorff. M., Nagarwalla N. (1995) "Spatial disease clusters: detection and inference". Statistics in Medicine 14:799–810

Levine N., Kim K., Nitz L.H. (1995) "Spatial analysis of Honolulu motor vehicle crashes: II. Spatial patterns

Levine N., Kim K., Nitz L.H. (1995) "Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators

Openshaw, S., Charlton, M., Wymer, C., and Craft, A., (1987) 'A mark I geographical analysis machine for the automated analysis of point data sets', International Journal of Geographical Information Systems, 1, p335-358

Rogerson P.A. (2001) "A statistical method for the detection of geographic clustering", Geographical analysis, Vol.33 No 3 p.215-227

Rushton G., Lolonis P. (1996) "exploratory spatial analysis of birth defect rates in an urban population", Statistics in medicine, Vol.15 p.717-726

Smith "Disease cluster detection methods: the impact of choice of shape on the power of statistical tests", www.coblestoneconcepts.com/ucgis2summer/smith/SMITH.HTM

Smith, G.M. & Fuller, R.M. (2001) An integrated approach to land cover classification: an example in the Island of Jersey. In: International journal of Remote Sensing, Vol. 22, pp. 3123-3142

Steinnocher K., (1999) Adaptive fusion of multisource raster data applying filter techniques. In: Remote Sensing of Environment, Vol. 32, pp. 107-117

Talbot et al. (2000) "Evaluation of spatial filters to create smoothed maps of health data", Statistics in medicine Vol.19 p.2399-2408

Tango T. (2000) A test for spatial disease clustering adjusted for multiple testing. Statistics in Medicine 19:191–204

5.2. LUC

Anand, S. S., Bell, D.A., Hughes J.G., Patrick A., (1997) "Tackling the cross sales problem using data mining." Proceedings of the 1st International Conference On Knowledge Discovery and Data Mining.

Anselin, L., (1995) "Local indicators of spatial association-LISA." Geographical Analysis 27 (2), 93-115

Beaucourt L. en Van Aken, P. en Beel, G., (1998) "Zelfmoord of verkeersongeval?" in opdracht van de Vlaamse minister van Financiën, Begroting en Gezondheidsbeleid.

Brijs T., Swinnen G., and Vanhoof K. (2002) "Retail Market Basket Analysis: A Quantitative Modelling Approach" Phd dissertation

Brin, S., Motwani, R. and C. Silverstein (1997) "Beyond market baskets: generalizing association rules to correlations" Proceedings of the ACM SIGMOD Conference on Management of Data, Tucson, Arizona, USA, May 13-15, 1997, pp. 265-276.

Brodsky, H., Hakker A.S., (1988) "Risk of a road accident in rainy weather." Accident Analysis and Prevention, 20, 161-176.

Claret P.L. et al., (2003) "Age en sex differences in the risk of causing vehicle collisions in Spain, 1990 to 1999." Accident Analyses and Prevention 35, 261-272.

De Groote, Patrick, and Vicky Truwant, (2003) "Demografie & Samenleving." Leuven: Universitaire Pers Leuven.

De Somer, A., (1993) "Registratie van fietsongevallen in spoedgevallendiensten van een stedelijke regio" Tijdschrift voor Geneeskunde, 49/22,1993

Eckhardt N., Flahaut B., Thomas I., (2003) "Stationarité du système spatial en matière d'insécurité routière." Recherche Transport et Sécurité (accepted for publication)

Edwards J., (1996) "Weather-related road accidents in England and Wales: a spatial analysis." Accident Analysis and Prevention 4 (3) 201-212.

Flahaut, B., Mouchart, M., San Martin, E., Thomas I., (2003) "The local spatial autocorrelation and the kernel method for identifying black zones. A comparative approach." Accident Analysis and Prevention 35 (6), 991-1004.

Flahaut B., Thomas I., (2002) "Identifier les zones noires d'un réseau routier par l'autocorrélation spatiale locale." Revue Internationale de Géomatique 12 (2), 245-261.

Geurts K., Wets G., Brijs T. and Vanhoof K., (2002) "The Use of Rule Based Knowledge Discovery Techniques to Profile Black Spots." Paper presented at the the 6th Design and Decision Support Systems in Architecture and Urban Planning Conference, Ellecom, The Netherlands, July 7-10.

Geurts, K., Wets G., Brijs T., Vanhoof K., (2003) "Profiling high frequency accident locations using association rules." Proceedings Transportation Research Board (CD-ROM), Washington D.C, USA, January 12-16.

Hakamies-Blomqvist, L., (1998) "Older Drivers' Accident Risk : Conceptual And Methodological Issues" Accident Analysis and Prevention. 30, 293-297

Kaufman, L. and Rousseeuw, P.J., (1990) "Finding Groups in Data : An Introduction to Cluster Analysis"

Larsen L., Klines P., (2002) "Multidisciplinary in-depth investigations of head-on and left-turn road collisions." Accident Analysis and Prevention 34, 367-380

Levine, E., Bédard, M., Molloy, D.W. Basilevsky, A., (1999) "Determinants of driver fatality risk in front impact fixed object collisions." Mature Medicine Canada 2, 239-242

Levine N., (2002) "CrimeStat II: A Spatial Statistics Program for the Analysis of Crime Incident Locations (version 2.0)." Ned Levine & Associates: Houston, TX/National Institue of Justice: Washington, DC.

Li, G., Baker, S.P., Langlois, S.A., Kelen, G.D. (1998) "Are Female Drivers Safer? An Application Of The Decomposition Method", Epidemiology 9, 379-384

Mannila, H., (1997) "Methods and problems in data mining" Proceedings of the International Conference on Database Theory, Delphi, Greece, January 8-10, 1997, pp. 41-45.

Maycock, G., (1997) "The Safety of Older Car-drivers in the European Union." European Road Safety Federation, ERSF, AA Foundation for Road Safety Research, Basingstoke, UK.

McLachlan, G., and Basford, K. E., (1988) "Mixture Models." Marcel Dekker, INC, New York Basel,

Moran, P., (1948) "The interpretation of statistical maps." Journal of the Royal Statistical Society 10b, 243-251.

Silverstein, C., Brin, S. and R. Motwani, (1998) "Beyond market baskets: generalizing association rules to dependence rules" Data Mining and Knowledge Discovery 2(1), pp. 39-68.

Thomas I., (1992) "Verkeersveiligheid per wegsegment in België" Politeia 2,2, pp.21-23.

Vermunt, J.K., and Magidson, J., (2000) "Latent Gold 2.0 User's Guide" Belmont, MA : Statistical Innovations, Inc.

Vollrath M. et al., (2002) "How the presence of passengers influences the risk of a collision with another vehicle." Accident Analyses and Prevention 34, 649-654.

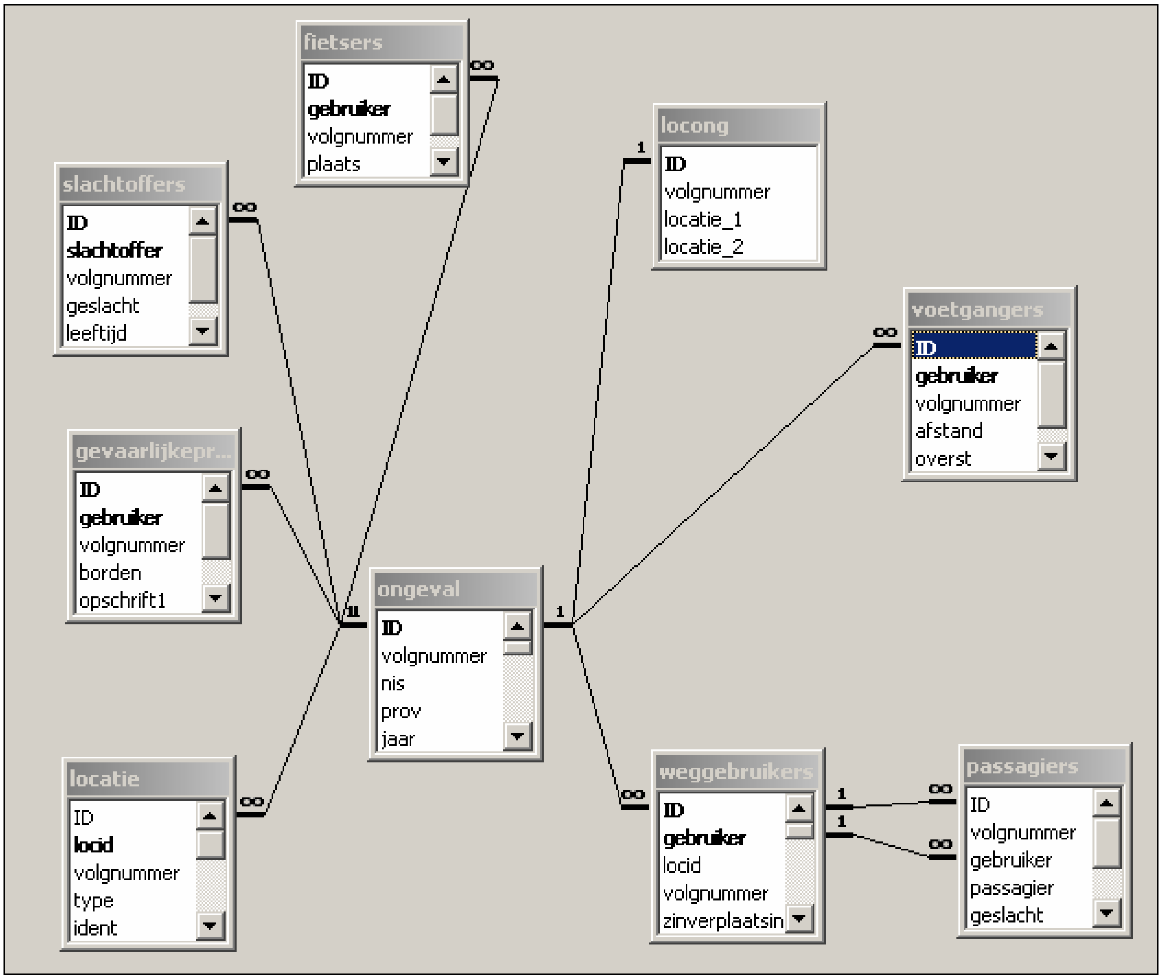
5.3. UCL

Cfr. Papers in appendices.

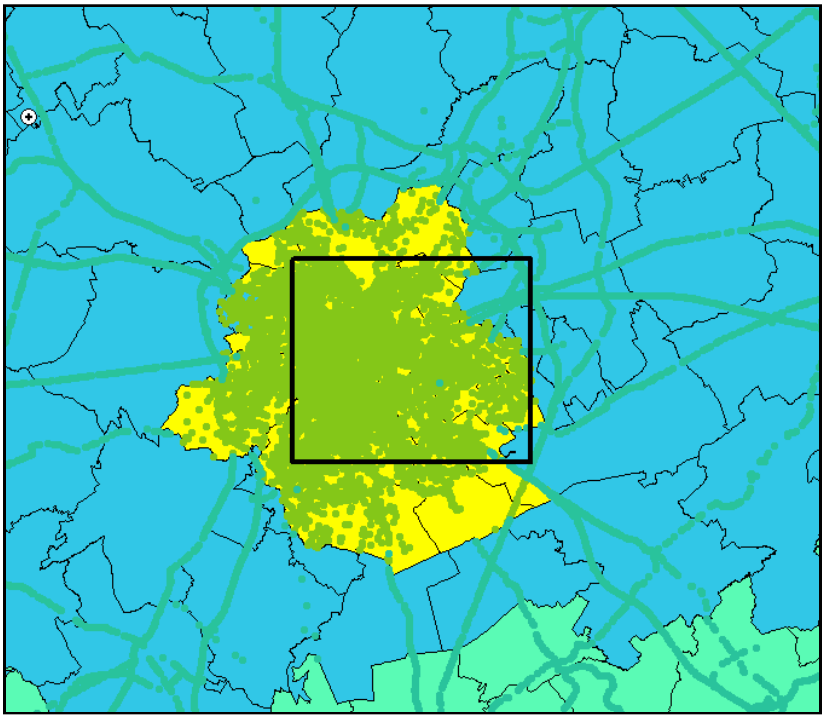
Part III: Appendix

A. Accident location

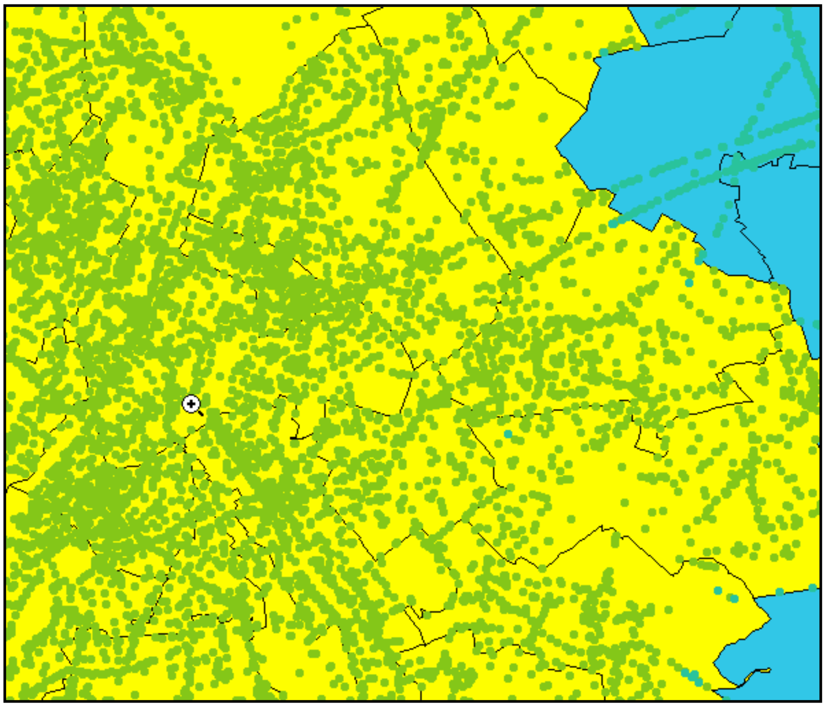
A.1 Relational database



A.2 Accidents map



Accident locations on numbered roads (Flanders, Walloon) and on all roads (Brussels, centre)



Accident locations on all roads in Brussels

B. Ikonos characteristics

B.1 Ikonos products

Accuracy of Ikonos products (www.spaceimaging.com)

Name	CE90	RMS	US NMAS
Geo	15m[1]	NA	NA
Reference	25m	11.8	1:50,000
Pro	10m	4.8	1:12,000
Precision	4m	1.9m	1:4,800
PrecisionPlus	2m	0.9m	1:2,400

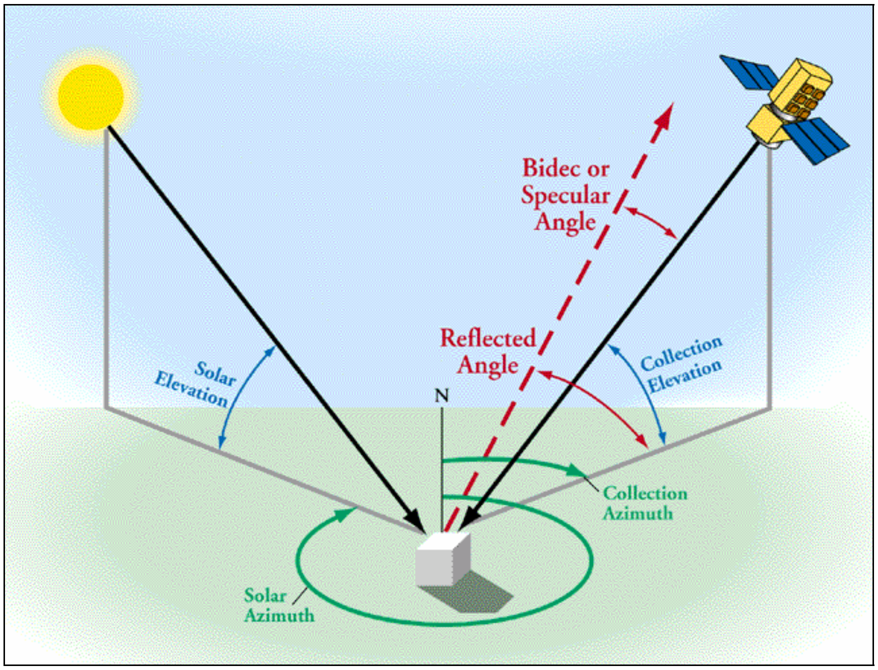
[1] Not including effects of terrain

The accuracy for the Geo image is 15m CE90 (it used to be 50 meters in the past) but what is important is that this accuracy measure does not include effects of terrain displacements due to the relief. In other words the Geo product is not orthorectified and this produces extra errors, especially in mountainous terrain. Only basic preprocessing has been done: resampling of the pixels to a uniform resolution of 1 x1 meter and projection to a map coordinate system (mostly UTM).

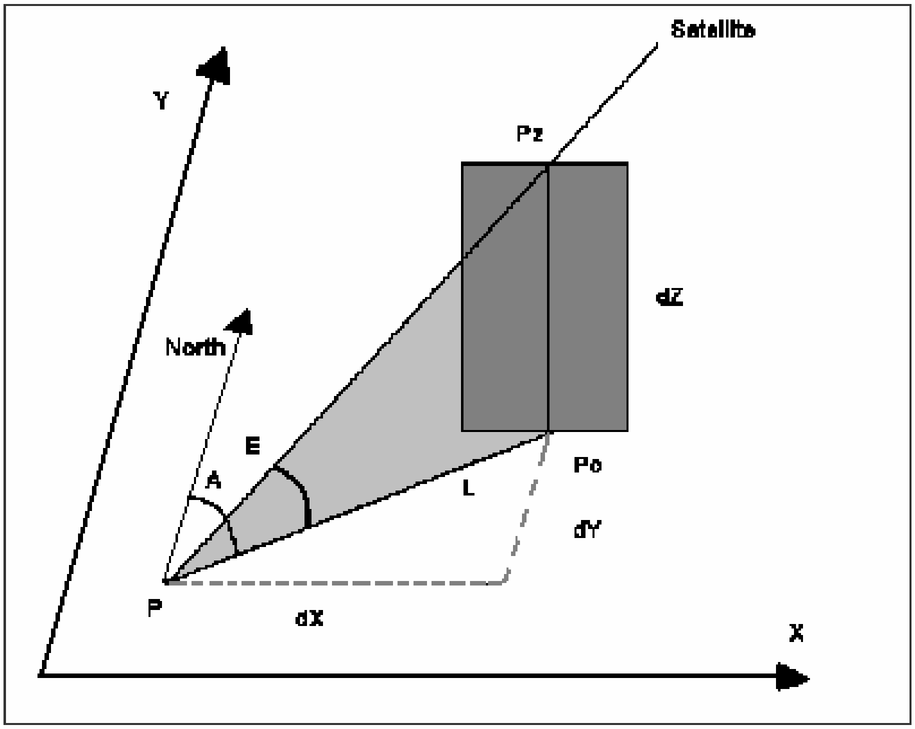
A panchromatic Ikonos Geo image costs about 28 USD / km² while the PrecisionPlus reaches 120 USD: km² (prices: march 2003, Space Imaging Eurasia) so it is worth trying to orthorectify the Geo images ourselves.

B.2 Orthorectification method (Baltsavias et al. 2001)

Elevation and azimuth



Relief-corrected affine transformation



A = Sat. Azimuth
E = Sat. Elevation

- Step 1: eliminating terrain displacements

Because of the satellite’s oblique view angle the point Pz with height dZ to the horizontal reference plane will be projected to P instead of its real position $P_{0.}$ For simplicity we assume this reference plane to be at height 0 because in this way dZ equals Z. Knowing the satellite’s position (azimuth and elevation) and the point’s height we can also project P_0 to P, just as the satellite does:

The radial displacement (along the azimuth) = L = dZ / Tan (E) and along the X and Y axes:

- o dX = - L * Sin (A) = - dZ * Sin (A) / Tan (E) (1)
- o dY = - L * Cos (A) = -dZ * cos (A) / Tan (E) (2)

- Step 2: Defining the transformation

We use an affine transformation to establish a link between both the Ikonos and the reference coordinate systems. Note that the transformation links the reprojected coordinates (from Step 1) to the Ikonoscoordinates.

- o x = a1+ a2 * X’ + a3 * Y’ (3)
- o y = b1+ b2 * X’ + b3 * Y’ (4)

where

X’ = X + dX

Y’ = Y + dY

(X’, Y’) = a GCP in the reference system, reprojected to the plane at reference height

(x,y) = the corresponding GCP in the Ikonos image

- Both steps combined

Substitution of X' and Y' (Eq. 1, 2) results in the following equations, called the "relief corrected affine transformation"

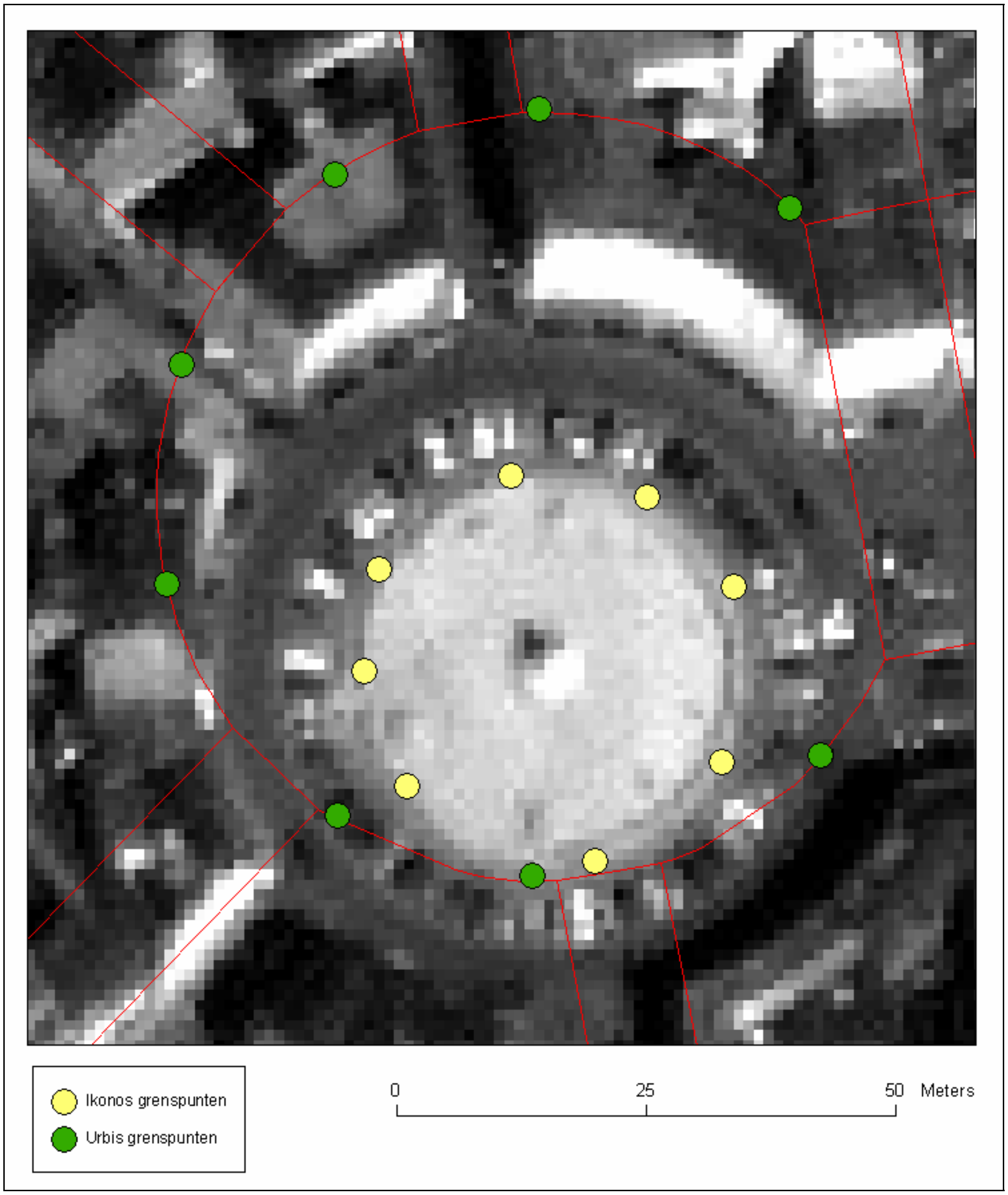
- $x = a1 + a2 * X + a3 * Y – a4 * Z$
- $y = b1 + b2 * X + b3 * Y – b4 * Z$

where

$a4 = [a2 * \mathrm{Sin} (A) + a3 * \mathrm{Cos} (A)] / \mathrm{Tan} (E)$
$b4 = [b2 * \mathrm{Sin} (A) + b3 * \mathrm{Cos} (A)] / \mathrm{Tan} (E)$

This transformation permits to project every point (X,Y,Z) to its corresponding point in the Ikonos image. At that specific location in the image a grey value is determined by means of a resampling technique. Note that orthorectifying an image as described above works reversely to our copmmon sense: instead of directly projecting the Ikonos pixels to their exact locations, we reproject the exact and correct pixel location to a location somewhere on the Ikonos image and copy that (resampled) pixelvalue into the original exact location.

B.3 Precise location of ground control points



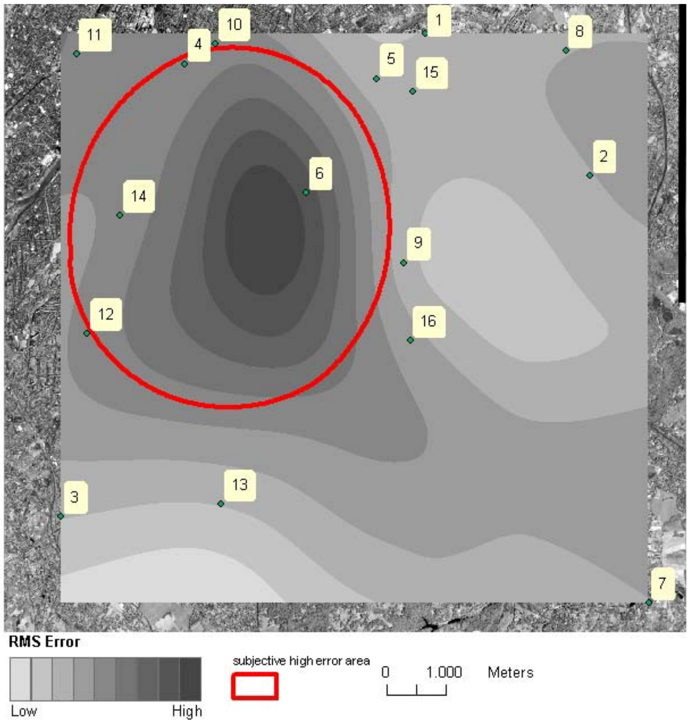
yellow: points along the border of roundabout in image
green: points along the roundabout in reference map

→ Both sets of points compose an ellipse of which the central point can be calculated. This point is located with higher precision then each of the composing points.

B.4 Orthorectification results

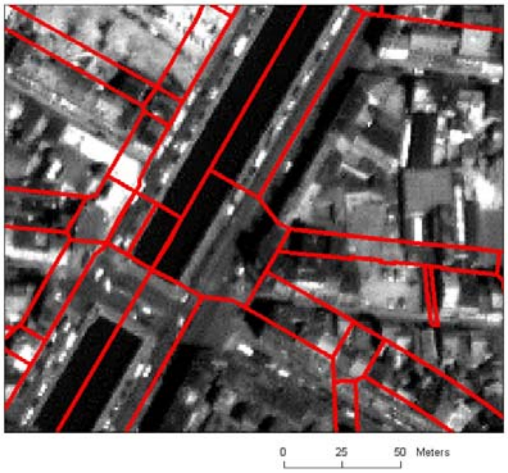
NR	Res.X	Res.Y	RMS
1	-0,4	-0,3	0,5
2	-2,0	-1,9	2,8
3	0,0	-0,5	0,5
4	**-2,1**	**-5,6**	**5,9**
5	1,0	0,6	1,1
6	**6,4**	**11,1**	**12,9**
7	-1,6	-1,7	2,4
8	0,0	-2,5	2,5
9	1,3	-0,9	1,5
10	**1,2**	**5,8**	**5,9**
11	**-2,4**	**-6,0**	**6,4**
12	**-1,5**	**-5,0**	**5,3**
13	-0,1	0,0	0,1
14	**0,3**	**5,5**	**5,5**
15	-1,6	-1,1	1,9
16	1,8	2,4	3,0

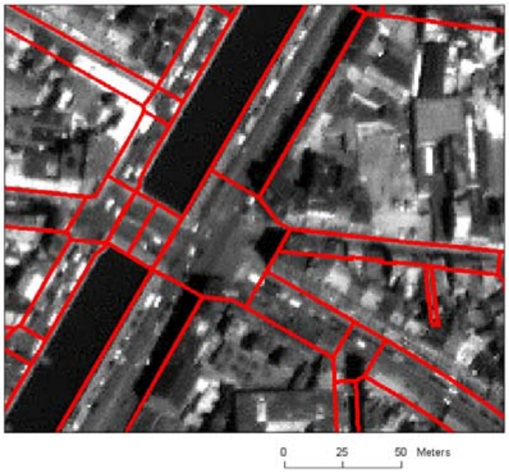
residuals and RMS errors (meters)



interpolated RMS errors (grey) and subjectively large error area (red ellipse)

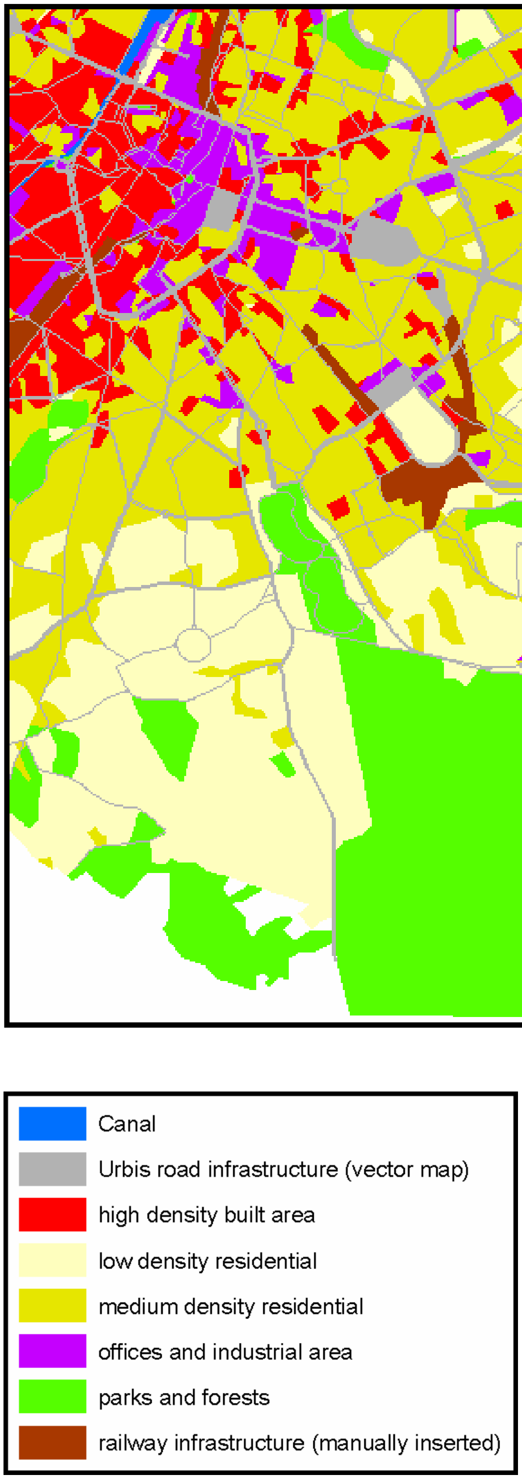
B.5 Rubber sheeting results





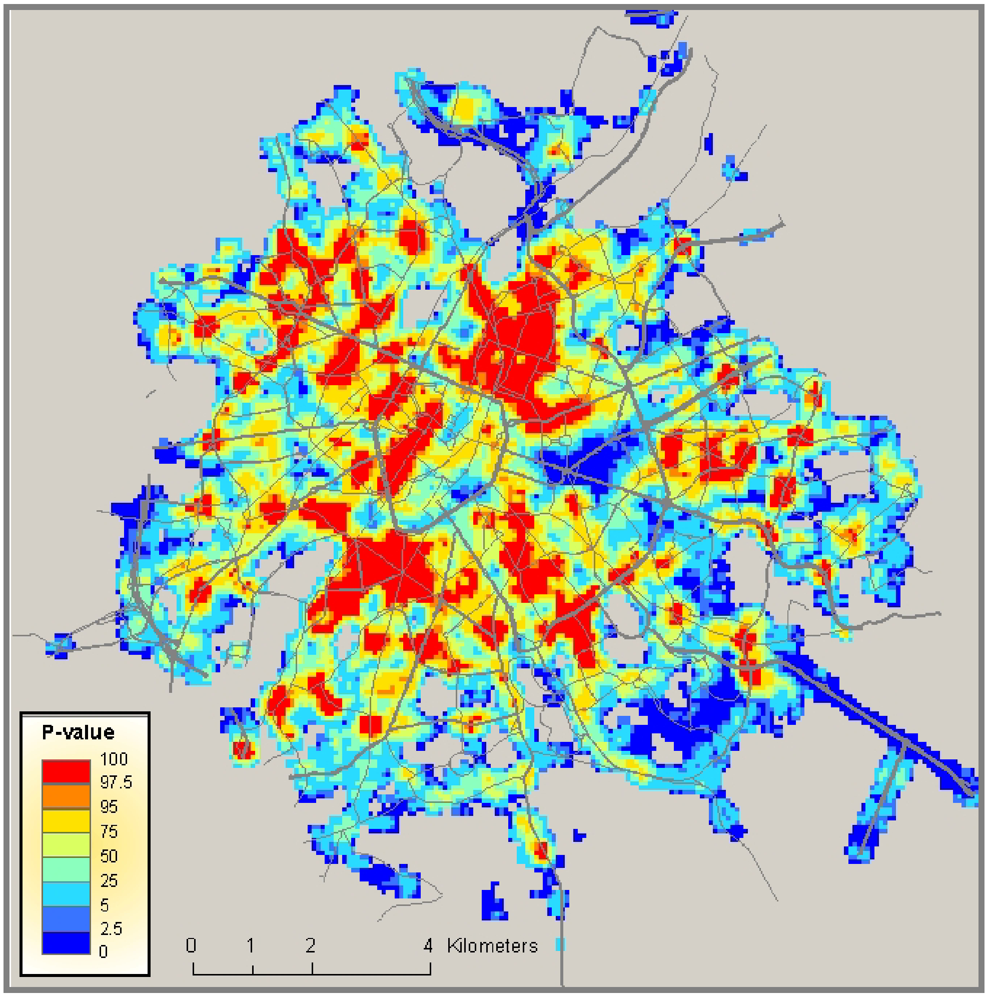
C. Land use

Final land use map of Brussels


Canal
Urbis road infrastructure (vector map)
high density built area
low density residential
medium density residential
offices and industrial area
parks and forests
railway infrastructure (manually inserted)

D. Black zones

D.1 Risk Significance map



E. Accident modelling

E.1 Data: neighbourhood factors

AgeMin20	Population aged below 20
Age20-64	Population between 20 and 64 years
AgePlus65	Population older then 64 years
Europeans	Total Europeans
TurksAfricans	Total Turks and Africans
Foreigners	Total Foreigners
BuiltOpen	Morphological building type: number of open buildings
BuiltMix	Morphological building type: number of semi-open buildings
BuiltClosed	Morphological building type: number of closed buildings
ApartmentsStudios	Number of Apartments & Studios
WorkExecutives	Labour group: Executives
WorklEmployers	Labour group: Employers
WorklIndependent	Labour group: independents (shopkeepers etc.)
WorklOther	Labour group: other work
WorkUnemployed	Labour group: unemployed population
SchoolPop	School population
WorkPop	Labour population
ComfortWithout	Housing comfort level: no basic comfort
ComfortSmall	Housing comfort level: basic comfort (running water, toilet, bath)
ComfortMedium	Housing comfort level: average comfort (basic comfort + central heating)
ComfortHigh	Housing comfort level: high comfort
CarNone	Mobility means of the household: no car
CarMore	Mobility means of the household: one or more cars
BikeNone	Mobility means of the household: no bicycle
BikeMore	Mobility means of the household: one or more bicycles
TotalPop	Total population
TotalDwellings	Total dwellings
PoorNeighbourhood	District is classified as a poor neighbourhood*

* Poor neigbourhoods according to the study of Jacobs & Swyngedouw (2000). To a large extent this factor’s spatial pattern is similar to the pattern of the factor TurksAfricans. In the perspective of explaining and interpreting neighbourhood traffic safety, it is better to use this concept.

E.2 Correlations among the factors

Below are three sets of uncorrelated factors

Uncorrelated set 1:

- Exposure
- SchoolPop
- LabourPop
- TurksAfricans
- BuiltOpen
- BuiltMix
- BuiltClosed
- WorkEecutives
- ComfortWithout

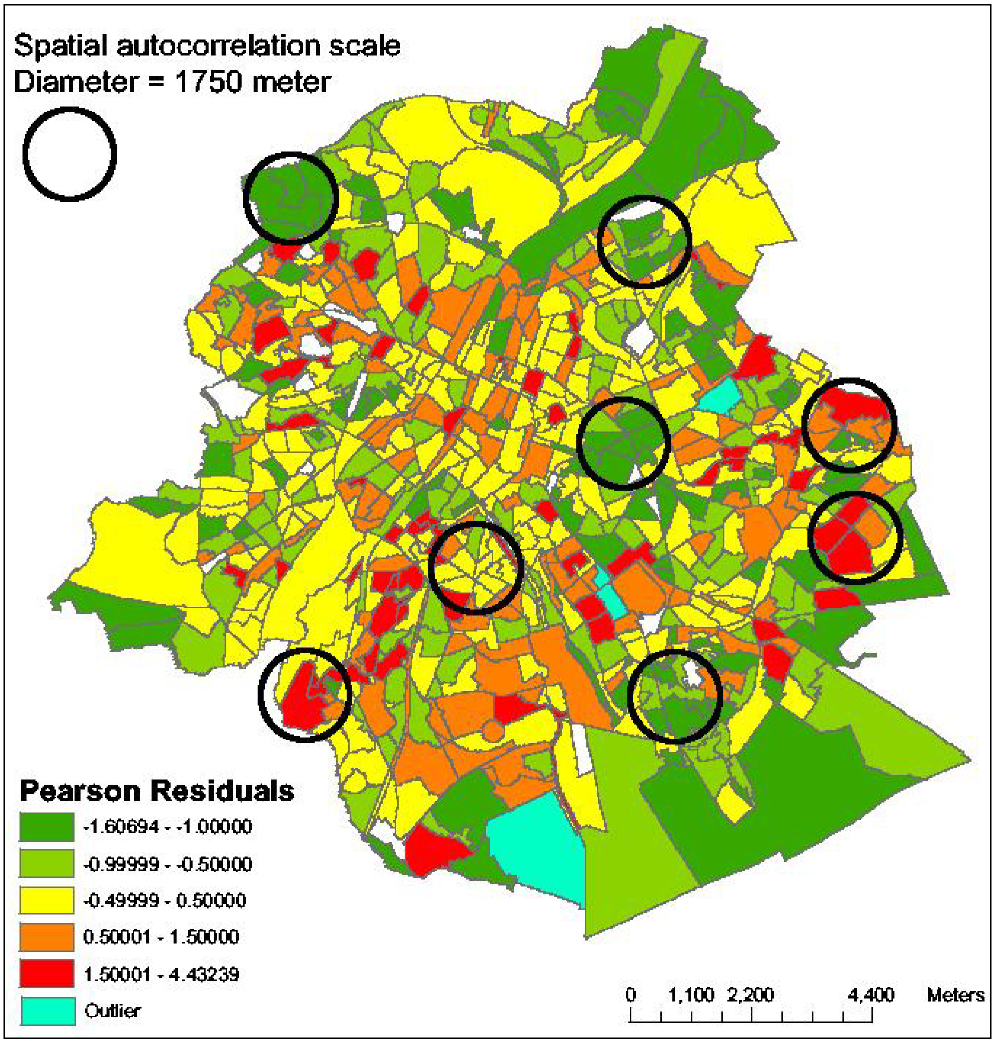
Uncorrelated set 2:

- Exposure
- SchoolPop
- WorkPop
- BuiltOpen
- BuiltMix
- BuiltClosed
- WorkEmployer
- WorkUnemployed

Uncorrelated set 3:

- Exposure
- SchoolPop
- LabourPop
- BuiltOpen
- BuiltMix
- BuiltClosed
- ComfortLarge
- ComfortSmall

E.3 Graphic illustration of spatial autocorrelation of Negative Binomial regression residuals



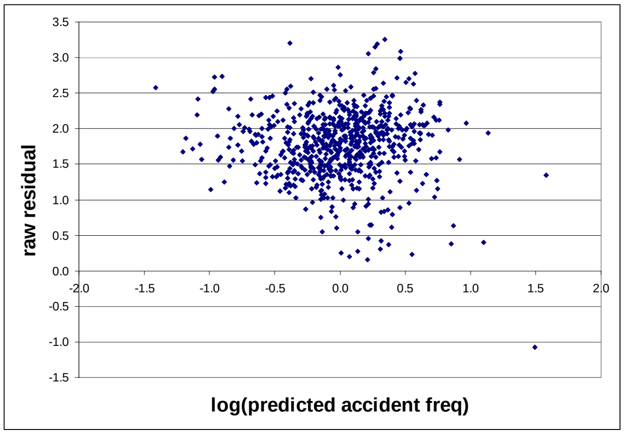
E.4 Goodness-of-fit results for Poisson and negative binomial models

		Poisson		Negative binomial	
Criterion	*DF*	*Value*	*Value/DF*	*Value*	*Value/DF*
Deviance	692	2158,24	3,12	798,63	1,15
Pearson Chi-Square	692	2247,53	3,25	788,25	1,14

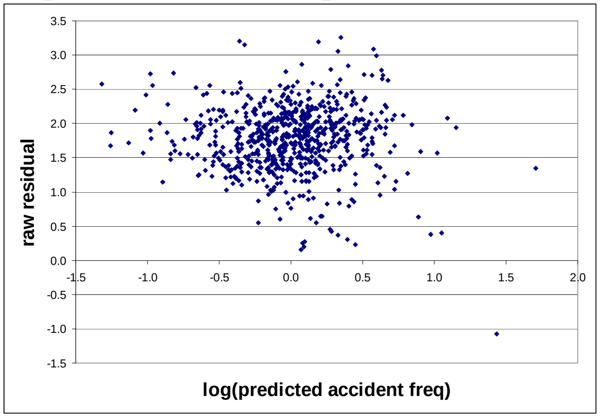
E.5 Heteroscedasticity graphics

A wedge shaped point cloud indicates heteroscedasticity)

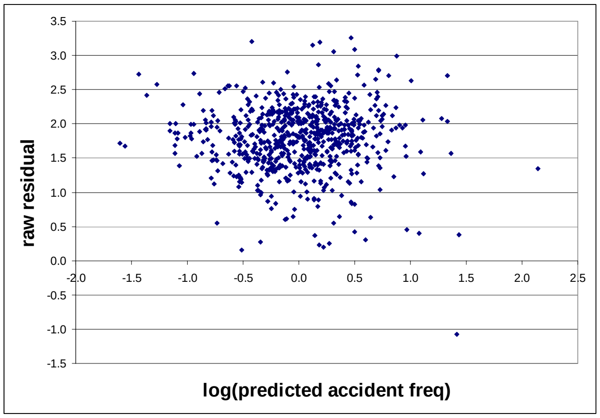
Graphic: residuals for basic spatial error model: high heteroscedasticity



Graphic: residuals for full spatial error model: moderate heteroscedasticity



Graphic: residuals for full OLS model: NO heteroscedasticity



F. Traffic accident data for Belgium from 1991 – 1999

F.1 Number of traffic accidents, injuries and casualties

Period 1991-1999	Belgium	Flanders region	Walloon region	Brussels region
Traffic accidents	505880	340184	138285	27411
Casualties and injuries	969379	660086	252573	56720

F.2 Recorded fields per traffic accident record

TABLE “ONGEVAL”

FIELD	**SECTION NUMBER**	**DESCRIPTION**	**VALUES ON FORM**	**VALUES IN TABLE**
id		New unique ID for each traffic accident		
volgnummer	-	Unique traffic accident ID	Num	Num
prov	-			Num
jaar	-			Num
r_p	1	Accident reported by police (P) or state police (G)?	Free field	G, P
eenheid	1	Code of the (state) police	Num	Num
pvnr	1	Number of the official police report	Num	Num
nis	2	NIS code of the city or town	Num	Num
tijdstip	3	Moment of the accident (day, month, year, hour)	Num	Date/Time, general date
kruispunt	4	1 = on a crossroad 2 = not on a crossroad	1, 2	1, 2
in_bbkom	12	1 = inside built-up area 2 = outside built-up area	1, 2	1, 2 missing (642)
kprgl1	7	Traffic control in the center of the crossroad (ONLY IN CASE OF A CROSSROAD)	1, ..., 5	1, ..., 6 missing (197581)
kprgl2	7	Ditto kprgl1	1, ..., 5	0, 4, 5 missing (338461)
licht	10	Illumination	1, ..., 4 9 = unkown	1, ..., 4 missing (1077)
plakar1	13	Other local characteristics	1, ..., 5 9 = unknown	1, ..., 5 missing (321109)
plakar2	13	Ditto plakar1	1, ..., 5 9 = unkown	2, 3, 4, 5 missing (339922)

staatw1	11	Condition of the road	1, ..., 5 9 = unknown	1, ..., 5 missing (5073)
staatw2	11	Ditto staatw1	1, ..., 5 9 = unkown	3, 4, 5 missing (312004)
factorenw1	18	Accidentfactors: Road/traffic characteristics	1,…8	1, ..., 8 missing (302872)
factorenw2	18	Additional answer	1,….8	2, ..., 8 missing (236823) spatie (66692) □□□□"□" (33266)
varia1	21	Varia (factors which played an influence on the accident or the severity of the accident	01, 02, ..., 09, 10, ..., 14 (14 = comment field)	01, 02, ..., 09, 10, ..., 14 1 (7) missing (254691)
varia2	21	Ditto varia1	Ditto varia1	02, 03, 04, 06, ..., 14 missing (334502)
varia3	21	Ditto varia1	Ditto varia1	06, 07, 08, 10, ..., 14 missing (339763)
varia4	21	Ditto varia1	Ditto varia1	08, 10, 12, 14 missing (340170)
weer1	9	Weather circumstances	1, ..., 7 9 = unknown	1, ..., 7 missing (4860)
weer2	9	Additional answer	Ditto weer1	3, ..., 7 missing (337803)
aantweggebruikers	24	Number of road users		Num
aantpass	25	Number of dead or injured passengers		Num
aantslachtoffers	26	Casualties, indirectly involved in the accident		Num
totaalbetrokken	23	Total number of involved drivers and pedestrians.	Num (vrij veld)	Num
totaaldoden	23	Total number of casualties	Num (vrij veld)	Num
totaalligew	23	Total number of light injuries	Num (vrij veld)	Num
totaalzwgew	23	Total number of heavy injuries	Num (vrij veld)	Num
dodenlg	23	Deceased light injuries	Num (vrij veld)	0, 1
dodenzg	23	Deceased heavy injuries	Num (vrij veld)	0, 1, 2

TABLE "WEGGEBRUIKERS"

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLES
volgnummer	-	Unique traffic accident ID	Num	Num
gebruiker	Letters A,B,..	Distinct users for each accident		Num (1 – 27)
locid				Num Missing (15764)
zinverplaatsing	15	Sence of direction (for user A and B)	1, ..., 4 9 = unknown	1, ..., 4 space (37542) missing (33708)
aard	24	Driver and pedestrian (casualties, injuries and uninjured people) (see list CODES FOR DRIVERS AND PEDESTRIANS)	01, ..., 24 99 = unknown	01, ..., 24 missing (899)
aantpass	24	Number of pedestrians (uninjured people also considered)	Num (free field)	Num (0 – 90) Missing (5225)
alcohol	24	Alcohol test	1, ..., 4	1, ..., 4
geslacht	24	Sex (M / V)	1 = M 2 = V	1, 2 missing (9329)
gevolgen	24	Consequences (deceased, heavily injured, lightly injured, uninjured, deceased heavily injured, deceased lightly injured)	1, ..., 4	0, 1, ..., 6 missing (8568)
land	24	Country of registration	Free field	Char missing (161414) (several codes for same country!)
leeftijd	24	Age (rounded to lowest integer)	Num (free field)	Num (0 – 98) missing (31653)
nrplaat	24	License plate (only in case of Belgian license plate); blank for bycicle or moped.	Free field	Char/Num Missings (458641) Impossible values??
toest1	24	Concition (1 or 2 codes)	1, ..., 4	1, ..., 4 missing (36299)
toest2	24	ditto toest1	ditto toest1	3, 4 missing (659064)
typeaanrijding	8 A	Type of collision (for each driver)	1, ..., 8 9 = unknown	1, ..., 8 missing (288434)
tegenhindernis	8 B	Road users and obstacle involved with each collision (for each driver) See also form: CODE OBSTACLE	50, ..., 67 99 = unknown	50, ..., 67 missing (590789)

tegenweggebr	8	Identifying other roadusers involved in the collision		0, 1, ..., 7 missing (308754)
beweging	16	Motion / insight of the roaduser (for each roaduser)	1, ..., 16 99 = unknown	01, ..., 16 space (44603) missing (32240)
dynamica	17	Dynamics	1, ..., 4 9 = unknown	1, ..., 8 space (37542) missing (181780)
factorengebr1	18	Accident factors roaduser	1, ..., 9	1, ..., 8 space (37542) missing (336146)
factorengebr2	18	Ditto factorengebr1	1, ..., 9	2, ..., 8 space (37542) missing (603286)
factorenv1	18	Accident factors vehicle and/or trailer	1, ..., 4	1, ..., 4 space (37542) missing (618734)
factorenv2	18	Ditto factorenv1	1, ..., 4	2, 3, 4 missing (435173) space (163665) □□□□”□” (61167)

TABLE "PASSAGIERS"

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
gebruiker	-	Unique user ID	Num	Num (1 – 24)
passagier		Several passengers possible for each user		Num (1 – 47)
geslacht	25	Sex (M / V)	1 = M 2 = V	1, 2 missing (252)
leeftijd	25	Age (rounded to lowest integer)	Num (vrij veld)	Num (0 – 98) Missing (8849)
gevolgen	25	Consequences	1, 2, 3	1, 2, 3, 5, 6
plaats	25	Position in the vehicle	1, 2 9 = Unknown	Space (80671) Missing (32301)

TABLE "VOETGANGERS"

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
gebruiker	-	Unique user ID	Num	Num (1 – 6)
afstand	19	Crossing distance between protected location (in the event of the pedestrian crossing the road)	Free field (Num)	1, ..., 19 missing (14699)

overst	19	In the event of the pedestrian crossing the road: Pedestrian is positioned on the road behind an obstacle or vehicle which makes him invisible for the driver	1, 2 9 = unknown	1, 2 space (6600) missing (1522)
plaats	19	Position of the pedestrian	10, 11, 20, 30, 31, 40, ..., 46, 50 99 = unknown	10, 11, 20, 30, 31, 40, ..., 46, 50 Missing (311)

TABLE "FIETSERS"

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
gebruiker	-	Unique user ID	Num	Num (1 – 6)
plaats	20	Position of two-wheeler	1, 2, 3	1, 2, 3 space (6)
fietspad	20	In case the two-wheeler rides on a bicycle track or leaves a bicycle track	1, 2, 3	1, 2, 3 space (64570)

TABLE "SLACHTOFFERS"

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
slachtoffer		Multiple casualties possible for each accident		Num (1 – 8)
geslacht	26	Sex (other deceased or injured casualties)	Free field!!	1, 2
leeftijd	26	Age	Num (free field)	Num (0 – 94) Missing (14)
gevolgen	26	Consequences	Free field	1, 2, 3

TABLE “LOCATIE”

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
locid	-	Unique location ID	-	Num
type	5	Type of numbered road	1 = highway 2 = state highway or province road	0, 1, 2, 3, 4, 7, 8 Char (2 cases)
ident	5	Identification of numbered road	Letter (A, B, N, R, P, T) + number	Letter (A, B, N, R, P, T) + number Missing (270891)
wegtype	-	Roadtype (part of ident)	-	A, B, N, R, P, T Missing (270891)
wegnummer	-	Roadnumber (part of ident)	-	Num Missing (270891)
wegindex	-	Roadindex (part of ident)	-	Num Missing (270891)
kmp	5	Kilometre marker	Free field(num)	Num Missing (220913)
nrgebouw	5	House number	Free field	Char + Num $ (1) Missing (368613) Also other impossible values
snelheid	5	Maximum allowed speed	Free field	Num (0 – 195 !!) Missing (30938)
soort	5	Sort of road	1 = Road with 1 roadway 2 = Road with several roadways, seperated by roadside	1, 2 Missing (948)
straatnaam	5	Streetname (for non-numbered roads)	Free field	Char ”□” (20939) Space (43233) Missing (147734)
straattype	5	Designation of non-numbered road	Char (see table on form)	Char (codes) 03, 12, 13, 23, 41 Missing (211906)

TABLE “LOCONG”

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
Locatie_1	-	Location 1 ID	Num	Num
Locatie_2	-	Location 2 ID (only for crossroads)	Num	Num Missing (197573)

TABLE "GEVAARLIJKE PROD"

FIELD	SECTION NUMBER	DESCRIPTION	VALUES	VALUES IN TABLE
volgnummer	-	Unique traffic accident ID	Num	Num
gebruiker	-	Unique user ID	Num	Num (1 – 5)
borden	22	Orange plates	1 = blank 2 = provided with following numbers	1, 2
opschrift1	22	Inscription (In case *borden* = 2)	Free field (7 locations)	Num Spaces (40)
opschrift2	22	Ditto opschrift1	Ditto opschrift1	Num Spaces (182)
staatlading	22	Condition of the vehicle load	1, 2, 3	1, 2, 3 "1□", "2□", "3□" Spaties (2)

F.3 Inconsistencies between data from several sections on the VOF form

Discrepancies between the totals from section 23 and the summation over sections 24, 25 and 26

	Total (23) – sum (24+25+26)				
Region	Casualty	Lightly injured	Heavily injured	Deceased lightly injured	Deceased heavily injured
Flanders region	+ 1	- 9	- 1	0	- 1
Brussels region	+ 1	0	0	0	- 1
Walloon region	0	- 11	- 1	0	0

Discrepancies between pedestrians and cyclists totals for section 24, section 19/20 and section 8

	R24 sort	R19/20 specific	R8 Collision
Pedestrians			
Flanders region	21804	19803 (92,5%)	20160 (94,2%)
Brussels region	6497	6067 (93,4%)	6060 (93,3%)
Walloon region	14218	12706 (89,4%)	13259 (93,2%)
(Motor)cyclists			
Flanders region	122514	111408 (90,9%)	Not specifically mentioned in R8
Brussels region	2769	2719 (98,2%)	
Walloon region	25324	24387 (96,3%)	

Discrepancies between totals for the *Road users against obstacles* field and the *Obstacles* field

	R8a Road user against obstacle	R8b Type of obstacle
Flanders region	69491	69297 (99,72%)
Brussels region	4143	4128 (99,63%)
Walloon region	40935	40837 (99,76%)

Discrepancies between totals for the field *Works in progress* in section 13, 18 and 8

	R13 Local characteristics. – Works in progress	*R18 Traffic accident factors - Works in progress*	*R8 Obstacle – Works in progress*
Flanders region	*4441*	*3278*	*967*
Brussels region	*337*	*170*	*46*
Walloon region	*2450*	*1542*	*379*

Inconsistencies within section 24 among the field *Obviously drunk* and the results of the alcohol test

	Obviously drunk	Pos/Neg	Refusal	Missing
Flanders region	21486	17575/269	1045	2597 (12,1%)
Brussels region	810	651/8	50	101 (12,5%)
Walloon region	10902	8838/84	552	1428 (13,0%)

Intrinsic inconsistencies between section 16 and 17

	Standing still(R17) and evasive manoeuvres (R16)	*Standing still (R17) and loss of control (R16)*
Flanders region	*143*	*55*
Brussels region	*12*	*9*

Walloon region	*89*	*81*

F.4 Variables used for latent class clustering technique

Table 5-1: Description of the variables

Label	Definition	Mode	Categories
Id	Unique identification of accident	Continuous	/
Weekend	Time specification	Categorical :Nominal	No /Yes
Hour	Time specification	Categorical : Nominal	0 - 23
Season	Time specification	Categorical : Nominal	4 categories*
AccType	Describing direction of impact between road users, collision with obstacle, collision with pedestrian	Categorical :Nominal	8 categories*
Passenger position	Describing the position of the passenger in the vehicle	Categorical :Nominal	4 categories*
CrossRoadchar	Location specifications on priority regulation	Categorical :Nominal	4 categories*
Built-up Area	Area specification	Categorical :Nominal	Yes/No
Type	Road specification	Categorical :Nominal	Highway / Regional way
Soort	Road specification : Separation of lanes	Categorical :Nominal	Separated or not
Loc snelheid	Max allowed speed on road	Categorical :Ordinal	30 /50/ 60/ 90/ 120
Loc delta	Difference in Max allowed speed	Categorical : Nominal	Yes/ No
Black zone	Area specification	Categorical :Nominal	Yes /No
VisibilityAggr	Accounts for factors influencing the visibility	Count	/
MomentAggr	Accounts for moment specific factors	Count	/
StructuralAggr	Accounts for infrastructural factors	Count	/
PersonalAggr	Accounts for personal factors	Count	/
Detrimcounts	Weighted sum of human detriment in the accident	Count	/
First Road User			
Sort1	Type of first road user	Categorical :Nominal	9 categories*
Gend1	Gender of first road user	Categorical :Nominal	Male/Female
Age1	Age of first road user	Categorical: Ordinal	8 categories*
NumbPass1	Number of passengers	Count	/
Consequences1	Consequences for first road user	Categorical :Nominal	5 categories*
Behavior1 (motion or infraction)	Situational factor describing Motion & Positioning & Infraction	Categorical :Nominal	10 categories*
Dynamics1	Driving dynamics of first road user	Categorical :Nominal	5 categories*
Second Road User			
Sort2	Type of second road user (if there is one, otherwise category '0')	Categorical :Nominal	9 categories*
Gend2	Gender of second road user	Categorical :Nominal	Male/Female
Age2	Age of second road user	Categorical: Ordinal	8 categories*
NumbPass2	Number of passengers	Count	/
Consequences2	Consequences for second road	Categorical :Nominal	5 categories*

	user		
Behavior2	Situational factor describing Motion & Positioning & Infraction	Categorical :Nominal	10 categories*
Dynamics2	Driving dynamics of second road user	Categorical :Nominal	5 categories*

Table 5-2: Description of different categories

Label	**Categories**									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
AccType		(Multiple)	Frontal	Rear end	Lateral	Pedestrian	Obstacle on the road	Obstacle off-road	1 road user, no obstacle	Unknown
Sort		Car	Van or Small Truck	Truck	Bus or Coach	Motorbike	Moped	Bicycle	Pedestrian	Other (10 : unknown)
Gender		Male	Female							
Age	0-15	16-17	18-21	22- 29	30-39	40-49	50-59	60-65	65+	
Consequences		Fatal	Heavily injured	Lightly injured	Uninjured	Fatal (30 days)				
Dynamics		Constant Speed	Brings to a stop	Accelerates	Stands still	Unknown				
Behavior (motion or Infraction)		Ignores red light	Gives no right of way	Crosses a continuous line	Passes wrongly	Evades unexpectedly	Place not in accordance	Loss of control	Keeps no distance	Continuous normal direction (10 : road side)
Weather		Normal	Rainfall	Fog & Smoke	Wind	Snow- & hailstorm				
Crossroad		Crossroad with Traffic light	Crossroad with priority of main road	Other crossroad (Agent, right of way, flashing light)	No crossroad, but road segment					
Season		Winter	Spring	Summer	Autumn					
Passenger Position	No passenger	At the front seat	At the back seat	At both front and back seat	Not known where					
Type of road		Highway	Regional way	Regional way, in built up area						
Soort		1 roadway	2 roadways Separated roadway	Combination						

Table 1: Indicators of traffic accidents at traffic accident level (analysis on Brussels Capital region)

Variable	**Values**
Weekend	Monday 02h00 - Friday 21h00; Friday 21h00 - Monday 02h00
Hour	7h-9h; 10h-12h; 13h-15h; 16-18h; 19h-21h; 22h-6h
Season	Winter; Spring; Summer; Fall
Accident type	Frontal collision; Collision from behind; Lateral collision; Collision with a pedestrian; Collision with an obstacle on the road; Collision with an obstacle next to the road; One road user, no obstacle
Passenger Position	At least one of the road users had passengers as well in front of the vehicle as in the back; There were passengers involved, but no car had passengers in front and in the back of the vehicle; There were no passengers involved
Crossroad	Crossroad with traffic lights; Crossroad with a priority road; Other type of crossroad; No crossroad
Built-up Area	Yes; No
Road Type	Highway; National, regional or provincial road; Local road
Road Sort	Single roadway; Divided highway
Speedlimit	30km/h; 50km/h; 60km/h; 90km/h; 120km/h
Speedlimit difference	Yes; No
Blackzone	Yes; No
Visibility	No visibility problems; Twilight; Rain; Hidden passengers; Twilight and rain; Twilight and hidden passenger; Rain and passenger; Twilight and rain and hidden passenger; Other reason
Road structure	Roundabout; Bridge or viaduct; Tunnel; Grade crossing; Sharp bend; School, recreation centre or bus stop; Signalisation; Steep descent

Environment	No special characteristics; Wet or snowy road surface; Wet or snowy road surface and other special characteristics; Other special characteristics
Detrimcounts	= # fatalities x 5 + # light and seriously injured persons who died within 30 days x 4 + # seriously injured persons x 3 + # slightly injured persons
Road user condition	The situation was completely normal (= negative alcohol test and normal condition of the road user) AND there were no protective measures missing (e.g. safety belt, helmet,. . .); The situation was not normal BUT no protective measures were missing; The situation is completely normal BUT some protective measures were missing; The situation is not normal AND some protective measures were missing; We can not tell with certainty whether or not the condition was normal.

Table 2: Indicators of traffic accidents at road user level (analysis on Brussels Capital region)

Variables	**Values**
Vehicle sort	Car; Medium sized vehicle; Large truck; Large bus; Motorcycle; Motorbike; Bicycle; Pedestrian; Other
Gender (road user)	Male; Female
Age (road user)	0-15 years; 16-17 years; 18-21 years; 22-29 years; 30-39 years; 40-49 years; 50-59 years; 60-65 years; 65+ years
Number of passengers	0 passengers; 1 passengers; 2 passengers; 3 or more passengers
Consequences (roaduser)	Deceased; seriously injured; slightly injured; not Injured
Behavior (road user)	Ignores a red light; Fails to give right of way; Crosses a full white line; Passes incorrectly; Makes an evasive manoeuvre; Is not positioned on the road according the regulations; Loss of control; Not keeping enough distance; Road user fell; No abnormal behavior
Dynamics	Road user travels at constant speed; Road user brakes in order to stop; Road user accelerates or starts; Road user is not moving

F.5 Results model based clustering on traffic accident data from the Brussels Capital region

Table 3: Direct effects incorporated in two-road user model

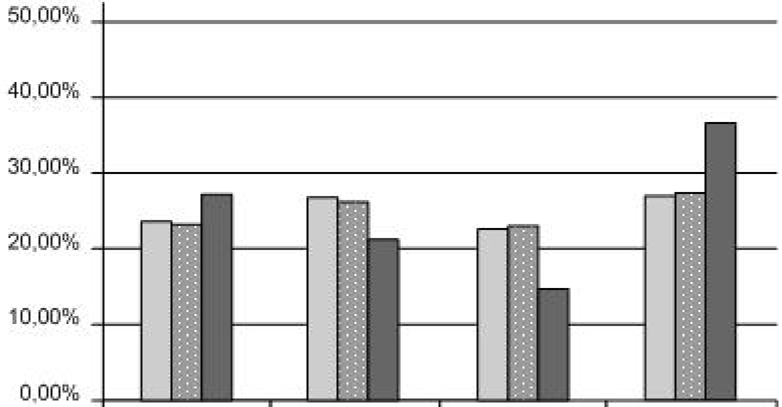
Direct effect	**p-waarde**
Road type & Crossroad	9.4 e^{-180}
Road sort & Road type	1.9 e^{-77}
Speedlimit difference & speedlimit	5.3 e^{-59}
Environment & visibility	8.3 e^{-12}
Consequences (Road user 2) & Consequences (Road user 1)	3.1 e^{-105}

Table 4: Traffic accident types

Traffic accident type		**Cluster size**
Cluster 1	Traffic accident with pedestrian involved	28.64%
Cluster 2	Light traffic accident with passengers involved	19.00%
Cluster 3	Severe traffic accident with passengers involved	4.34%
Cluster 4	Traffic accident with two-wheelers involved	11.78%
Cluster 5	Traffic accident on a road segment outside the built-up area against high speed	1.77%
Cluster 6	Traffic accidents strongly related with failing to give right of way	14.94%
Cluster 7	The dustbin traffic accident type	19.52%

F.6 Probability distribution of season indicator for traffic accidents with passengers involved

Figure 1: Probability distribution of season indicator for traffic accidents with passengers involved



	Winter	Spring	Summer	Fall
Entire population	23,6%	26,8%	22,6%	27,0%
Light traffic accidents w.p.i.*	23,24%	26,28%	23,00%	27,47%
Severe traffic accidents w.p.i.*	27,21%	21,42%	14,71%	36,65%

*w.p.i. = with passengers involved

G. Spatio-temporalité des accidents de la route en périphérie urbaine. L'exemple de Bruxelles

The spatio-temporal distribution of road traffic accidents in a periurban environment
The case of Brussels

Nathalie Eckhardt[1], Benoît Flahaut [1, 2], Isabelle Thomas[1, 2]

[1] Université catholique de Louvain, Département de Géographie, 1348 Ottignies-Louvain-la-Neuve, Belgique.
[2] Fonds national de la recherche scientifique, rue d'Egmont n° 5, 1000 Bruxelles, Belgique

Article reçu le 6 décembre 2002; accepté le 4 septembre 2003

courrier: eckhardt@geog.ucl.ac.be

Abstract The aim of this paper is to analyze spatio-temporal variations in the spatial concentrations of road accident in a periurban environment. Analyses are conducted for the numbered roads in a Belgian province (Walloon Brabant) over a nine year period (1991-1999). Initially, we identify the concentrations of road accidents (or black zones) by means of spatial autocorrelation. Their locations during three three-year study periods are then compared by means of contingency tables and similarity indices. Our research reveals a considerable similarity between the locations of black zones and suggests the existence of a powerful spatial structure. Our findings suggest that an understanding of space is necessary in accident studies in order to identify more clearly its causal role and in order to specify local accident reduction measures (road improvements, local signing to modify driver behaviour, etc.).

Key words: Road accidents; Belgium; Black zones; Similarity indices; Space and time

1. Introduction

L'objectif de l'étude que nous présentons dans cet article est d'analyser les variations spatiales et temporelles des zones noires en matière d'accidents de la route.

Notre projet de recherche s'inscrit dans le cadre du deuxième plan d'appui scientifique à une politique de développement durable (PADD II) financé par les services fédéraux belges des affaires scientifiques, techniques et culturelles (SSTC).

L'identification des zones noires se fonde sur des indices locaux d'autocorrélation spatiale dont l'utilité en accidentologie n'est plus à démontrer (Flahaut et al., 2003), (Flahaut et Thomas, 2002). Si l'opérationnalité de la méthode est incontestable, il est important d'examiner si les zones noires ont une structure spatiale forte et stable dans le temps. Nous proposons pour ce faire une analyse exploratoire préalable qui vise à tester l'hypothèse suivante : *la distribution géographique des zones noires varie en fonction de la période de temps choisie*. Si cette hypothèse est rejetée, nous pouvons penser qu'il est fort probable qu'un lien environnemental et spatial fort existe, que les accidents se concentrent toujours aux mêmes endroits et que les résultats du modèle statistique explicatif seront indépendants de la période de temps choisie. De tels résultats sont des préalables indispensables à tout modèle explicatif des concentrations d'accidents de la route et à la compréhension de leur structure spatiale, le but étant de les réduire, ce qui constitue une priorité en matière de politique fédérale.

La plus petite unité spatiale pour laquelle les données d'accidents sont disponibles en Belgique sur des routes numérotées est l'hectomètre de route. Nous examinons l'évolution spatiale de la variable binaire d'appartenance à une zone noire (1 si l'hectomètre appartient à une zone noire, 0 autrement) à travers trois périodes de temps successives, grâce à différents indices et mesures statistiques de similarité (voir par exemple (Gatrell, 1983), (Kent & Coker, 1996), (Legendre & Legendre, 1998)). Les indices choisis mesurent la similarité statistique entre les périodes, en tenant compte de la disproportion induite par le grand nombre d'hectomètres n'appartenant pas aux zones noires.

Dans un premier temps, ces indices sont calculés pour l'ensemble des zones noires. Dans un second temps, nous avons mesuré la stabilité pour des sous-ensembles de zones noires que nous distinguons d'après leurs caractéristiques physiques et environnementales, telles le type de route (les autoroutes et périphériques d'une part, les nationales d'autre part), l'intensité du trafic, les caractéristiques physiques des infrastructures (avec ou sans berme centrale [21]) et l'environnement (urbain dense, urbain lâche, rural boisé, milieu rural ouvert).

Après avoir défini le concept de zone noire sur la base d'indices d'autocorrélation spatiale, nous présentons la zone géographique sur laquelle nous avons suivi l'évolution spatiale de ces zones noires et la base de données construite à cet effet. Puis nous détaillons la méthode qui nous a permis de comparer les zones noires sur les périodes de temps successives (hypothèses sous-jacentes, choix des indices, etc.). Les résultats des analyses sont présentés dans la troisième partie. Ils permettent de conclure à la stabilité temporelle des zones noires, principalement dans certaines conditions d'infrastructure, de trafic et d'environnement.

[21] En Belgique on emploie le terme berme centrale pour désigner un terre-plein central.

2. Méthodologie

2.1. Définition des zones noires

Chaque accident est localisé suivant la borne hectométrique la plus proche, soit avec une précision de 100 m. Un nombre (*n*) d'accidents est ainsi affecté à chaque hectomètre de route. Une agrégation spatiale des hectomètres permet d'identifier les zones noires, qui sont définies comme des tronçons de route constitués de plusieurs hectomètres contigus et caractérisés ensemble par un nombre élevé d'accidents (et non par des mesures de risque d'accident). Cette agrégation spatiale est effectuée par des mesures locales d'autocorrélation spatiale (indices de Moran) suivant une méthodologie proposée et discutée dans (Flahaut et al., 2003) et (Flahaut et Thomas, 2002). Cette méthode présente l'avantage de tenir compte de la nature de l'espace (mesures de proximité), d'être statistiquement fondée et de fournir non seulement un indice de dangerosité, mais aussi la longueur des segments dangereux. La zone noire est moins volatile et moins mobile que le point noir, dont les inconvénients ne sont plus à démontrer (voir par exemple (Silcock & Smyth, 1985), (Nguyen, 1991), (Joly et al., 1992), (Hauer, 1996) et (Vandersmissen et al., 1996)). Les zones noires varient en longueur et en intensité.

Le calcul des indices de dangerosité permet l'identification de zones noires de longueur variant de 100 à 700 m. La définition de la longueur d'une zone noire est un aspect important à prendre en compte. I. Thomas (1996) a montré que le choix de la longueur des sections de route a une forte influence sur les mesures statistiques liées au nombre et à la densité des accidents. L'indice local de Moran constitue à la fois un indicateur de dangerosité (selon la valeur de l'indice) et un indicateur de la longueur des zones dangereuses (selon le nombre de voisins pris en compte dans le calcul de l'indice).

2.2. Zone d'étude et données

Nous allons étudier l'évolution dans le temps de la structure spatiale des zones noires à l'échelle d'une province belge : le Brabant wallon (Fig. 1).

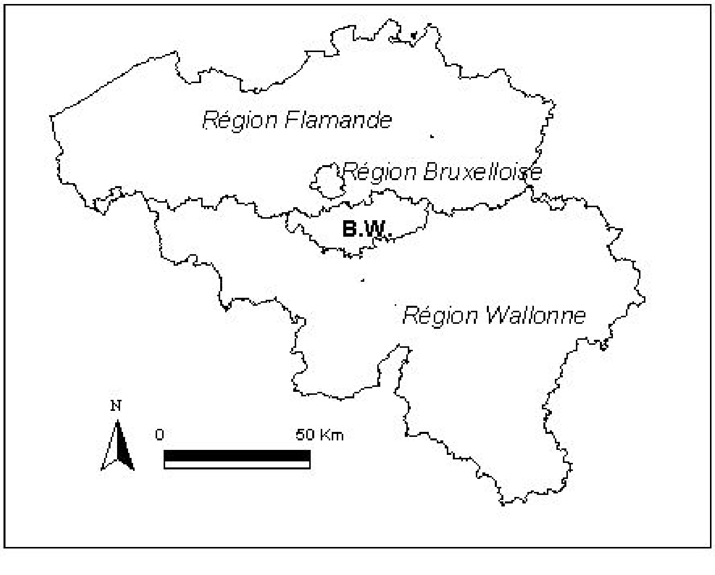


Fig. 1 Situation du Brabant wallon au sein de la Belgique

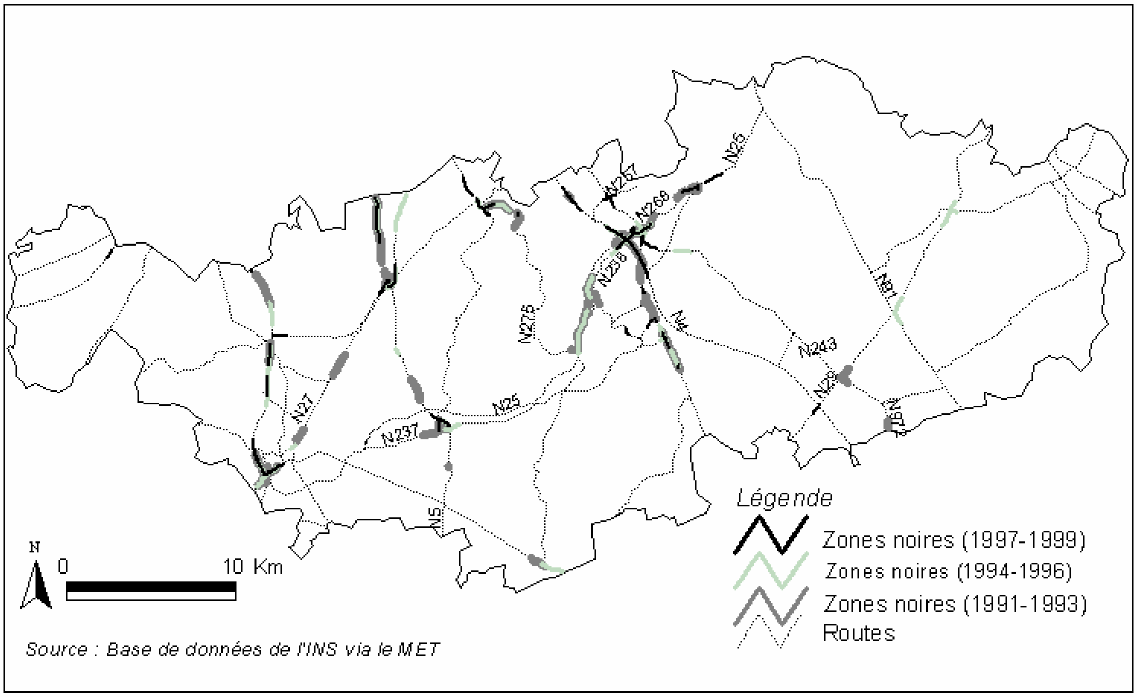
Cette province, qui s'étend au sud de Bruxelles sur une superficie de 1 090 km^2, comprend vingt-sept communes et comptait 347 423 habitants lors du dernier recensement de 1999. Bien qu'à caractère principalement périurbain (Thomas et al., 2000), elle présente cependant des disparités géographiques : à côté du résidentiel périurbain, un résidentiel ancien (anciens noyaux), quelques petites villes (Wavre, Nivelles), ainsi qu'un tissu industriel (à l'ouest) et un tissu à dominante agricole (à l'est). La zone d'étude est donc fort variée en termes d'occupation du sol. Conformément à l'organisation spatiale du trafic dans les zones périurbaines (voir par exemple (van der Laan et al., 1998)), les axes importants de la province sont des axes radiaux (nord-sud) vers la Région bruxelloise. Les axes transversaux (est-ouest) supportent un trafic moins important et sont moins touchés par les accidents (Fig. 2).

En Belgique, seuls les accidents avec lésions corporelles sont recensés et font l'objet de statistiques annuelles (Institut national de statistique (INS) via le ministère wallon des Équipements et des Transports (MET)). Les accidents sont localisés à l'hectomètre près sur les routes numérotées et par adresse postale sur les routes non numérotées. Nous avons limité notre analyse aux hectomètres de routes numérotées. Le Brabant wallon en compte 4 604. Entre 1991 et 1999, les routes du Brabant wallon ont comptabilisé 6 905 accidents avec lésions corporelles, dont 1 305 n'ont pu être localisés avec précision, ce qui ramène à 5 600 le nombre d'accidents analysables. Ces 5 600 accidents sont répartis sur 2 444 hectomètres de route (soit 53 % du réseau). Les zones noires regroupent 29,5 % des accidents qui se sont produits entre 1991 et 1999 et sont concentrées sur 20,4 % du total des hectomètres de routes numérotées.

Nous avons distingué les zones noires sur les trois périodes suivantes : 1991-1993, 1994-1996, 1997-1999. Une période de trois ans peut en effet être utilisée en accidentologie, car elle est estimée suffisamment longue pour limiter la part d'aléatoire dans l'occurrence des accidents. Les tests statistiques sont conduits séparément pour chaque groupe de deux périodes.

Notons que ces comparaisons peuvent être légèrement oblitérées par le fait que les enregistrements opérés par les forces de police sont plus précis et plus performants à la fin des années 1990 qu'au début. Ajoutons à cela que les corrections effectuées pour redresser les données sont également plus fiables et plus complètes ces dernières années, notamment pour le repositionnement d'une partie des accidents initialement non localisés.

Pour les trois périodes étudiées, les zones noires sont identifiées (Eckhardt et Thomas, 2004) et concernent respectivement 513, 466 et 476 hectomètres, soit chaque fois environ 10 % des 4 604 hectomètres du réseau total. Elles ont été successivement le théâtre de 9 %, 9 % et 12 % des accidents. Leur longueur varie dans le temps comme dans l'espace (Fig. 2), ainsi que leur niveau de dangerosité.



Source : Base de données de l'INS via le MET

Fig. 2 Zones noires en Brabant wallon pour les trois périodes 1991-1993, 1994-1996 et 1997-1999
Données de l'Institut national de statistiques (INS) transmises par le ministère de l'Équipement et des transports (MET).

2.3. Comparaison des zones noires dans le temps

Pour chaque période étudiée, chaque hectomètre du réseau routier est caractérisé par une variable binaire (0 ou 1) d'appartenance à une zone noire. Nous allons comparer les périodes entre elles par des indices de similarité, principalement fondés sur des tableaux de contingence que nous présentons dans cette section.

Soulignons tout d'abord une difficulté méthodologique qui nous a amenés à choisir différents indicateurs de similarité statistique pour répondre à l'inadéquation des tests classiques du χ^2. Celle-ci est due à la disproportion causée par le grand nombre d'hectomètres (80 %) n'appartenant à aucune zone noire pour aucune des trois périodes, qui complique la comparaison spatio-temporelle des cartes. Les indices devront également tenir compte de cette disproportion.

Tous les indices calculés se fondent sur un tableau de contingence où deux périodes de temps sont croisées (tableau 1).

Tableau 1 Tableau de contingence par croisement des périodes successives 1 et 2

Période 2	**5.3.1.1.1.1 Période 1**			
		0	1	total
	0	a	b	$a+b$
	1	c	d	$c+d$
	total	$a+c$	$b+d$	*5.3.1.1.1.1.1.1* N

a est le nombre d'hectomètres qui n'appartiennent à aucune zone noire au cours d'aucune période 1 ni 2 (double absence).

b est le nombre d'hectomètres qui appartiennent à une zone noire au cours de la période 1, mais pas de la période 2 (différence).

c est le nombre d'hectomètres qui appartiennent à une zone noire au cours de la période 2, mais pas de la période 1 (différence).

d est le nombre d'hectomètres qui appartiennent à une zone noire au cours des deux périodes (double présence).

N est le nombre total d'hectomètres ($N = a + b + c + d$).

De nombreux indices sont disponibles dans la littérature, mais aucun ne répond parfaitement à notre objectif et, de ce fait, plusieurs indices ont été abandonnés. Comme nous l'avons déjà dit, une des difficultés rencontrées est le nombre élevé de doubles absences. Or, le problème apparemment simple de comparaison de cartes dans le temps est lié à la difficulté de mesurer uniquement les doubles présences des hectomètres : la présence d'un hectomètre pour une période donnée est analytiquement plus intéressante dans notre cas que son absence. Ainsi, les indices choisis se focalisent sur les doubles présences, c'est-à-dire sur les hectomètres appartenant à une zone noire pour deux périodes successives, ce qui laisse espérer des résultats forts. Les indices de similarité *S* que nous avons choisi de tester (Everitt, 1977), (Fleiss, 1981), (Kent & Coker, 1996) (Legendre & Legendre, 1998), (Stokes et al., 2000) sont donnés dans le tableau 2.

Ces indices permettent de mesurer une similarité statistique entre les périodes de temps prises deux à deux : respectivement entre 1991-1993 et 1994-1996 et entre 1994-1996 et 1997-1999. La similarité maximale est égale à 1, 0 étant la valeur minimale. Un indice proche de 1 signifie donc une grande stabilité temporelle, tandis qu'un indice proche de 0 sera le signe d'une très faible stabilité temporelle.

Tableau 2 Les indices de similarité testés

Intitulé de *S*	Formulation de *S*
indice de simple concordance	$(a+d)/N$
indice de Jaccard	$d/(d+b+c)$
indice de Soerensen	$2d/(2d+b+c)$
variante de Jaccard	$3d/(3d+b+c)$
indice de Sokal & Sneath	$d/(d+2b+2c)$
indice de Kulczynski	$d/(b+c)$
variante de Kulczynski	$1/2\left[d/(d+b)+d/(d+c)\right]$
indice de Ochiai	$d/\sqrt{(d+b)\cdot(d+c)}$
indice de Faith	$(a/2+d)/N$

L'indice de simple concordance est l'indice de similarité le plus simple puisqu'il compte les doubles présences (*d*) et les doubles absences (*a*). Les autres indices excluent les doubles absences pour le calcul de la similarité, sauf l'indice de Faith qui les divise par deux. L'indice de similarité de Jaccard mesure le rapport des doubles présences sur la somme de *b, c* et *d*. L'indice de Soerensen (une variante de l'indice de Jaccard) multiplie par deux le poids des doubles présences. Une deuxième variante de Jaccard multiplie ce poids par trois. Les indices de simple concordance, de Jaccard et de Soerensen sont les plus utilisés, notamment en biologie végétale pour les inventaires botaniques (Kent & Coker, 1996), (Legendre & Legendre, 1998).

Les indices de similarité de Sokal & Sneath et de Kulczynski opposent tous les deux les doubles présences aux différences (*b* et *c*), mais avec une pondération différente. Une variante de Kulczynski oppose les doubles présences aux totaux marginaux $(d+b)$ et $(d+c)$. Enfin, l'indice de Ochiai utilise comme mesure de similarité le rapport entre les doubles présences et la moyenne géométrique des totaux marginaux des présences.

D'autres indices existent comme les mesures d'association, positive ou négative entre deux périodes. Dans ce cas, nous considérons la valeur observée pour les doubles présences (1-1) et nous calculons la valeur attendue par $(d+b)\cdot(d+c)/N$. Si la valeur observée est plus grande que celle attendue, la nature de la relation entre deux périodes est positive, c'est-à-dire que les hectomètres appartenant à une zone noire pour une période ont plus de risque d'appartenir également à une zone noire pour l'autre période et vice-versa.

Une dernière méthode utilisée concerne les odds ratios (*OR*), qui mesurent le risque ou la chance d'observer un événement au cours d'une période, si un événement ou non-événement est présent au cours de l'autre période. Ainsi, la probabilité d'observer 1 au cours de la période 2 si l'on a observé 1 au cours de la période 1 est donnée par $d/(d+b)$ et la probabilité d'observer 1 au cours de la période 2 si l'on a observé 0 au cours de la période 1 est donnée par $c/(a+c)$. Le ratio de ces deux probabilités $r = [d/(d+b)]/[c/(a+c)]$ signifie que le risque d'observer 1 au cours de la seconde période, sachant qu'on a observé 1 au cours de la première, est *r* fois plus grand (ou plus petit) que celui d'observer 1 au cours de la seconde période, sachant qu'on a observé 0 au cours de la première. Les odds ratios varient entre 0 et l'infini (Everitt, 1977), (Fleiss, 1981).

3. Résultats

3.1. Analyse globale de la stabilité spatio-temporelle

Les analyses sont réalisées en deux étapes : d'abord en prenant en compte la totalité du réseau (4 604 hectomètres), puis seulement les hectomètres du réseau qui appartiennent à une zone noire au cours d'au moins une des trois périodes de temps (941 hectomètres, soit 20,4 %). Le tableau 3 donne les tableaux de contingence dans le premier cas, le tableau 4 dans le second cas.

Tableau 3 Tableaux de contingence pour les deux couples de périodes comparées pour les 4 604 hectomètres de routes numérotées

1994-1996	1991-1993			
		0	1	total
	0	3 850	288	4 138
	1	241	225	466
	total	4 091	513	4 604

1997-1999	1994-1996			
		0	1	Total
	0	3 871	257	4 128
	1	267	209	476
	total	4 138	466	4 604

Tableau 4 Tableaux de contingence pour les deux couples de périodes comparées pour les 941 hectomètres de routes numérotées appartenant à une zone noire pour au moins une période

1994-1996	1991-1993			
		0	1	total
	0	187	288	475
	1	241	**225**	**466**
	total	428	**513**	**941**

1997-1999	1994-1996			
		0	1	total
	0	208	257	465
	1	267	**209**	**476**
	total	475	**466**	**941**

1994-1996	1997-1999			
		0	1	total
	a(0,0)	0	187	187
	c(0,1)	167	74	241
1991-1993	*b*(1,0)	208	80	**288**
	d(1,1)	90	**135**	**225**
	Total	465	**476**	**941**

À la lecture du tableau 4, nous pouvons remarquer une relative stabilité spatio-temporelle entre les périodes : nous observons environ 50 % d'hectomètres qui sont noirs pour chaque période (respectivement 513/941, 466/941, 476/941). Parmi les hectomètres dangereux en 1991-1993, 44 % le restent en 1994-1996 (225/513), tandis que parmi les hectomètres dangereux en 1994-1996, 45 % le restent en 1997-1999 (209/466). De plus, parmi les hectomètres dangereux en 1991-1993, 26 % (135/513) le restent au cours des deux périodes suivantes.

Tableau 5 Indices de similarité pour les deux couples de périodes comparées, pour les 4 604 hectomètres de routes numérotées et pour les 941 hectomètres noirs

	5.3.1.1.1.2 Similarité 1991-1993/1994-1996		**Similarité 1994-1997/1997-1999**	
Intitulé de l'indice *S*	tout le réseau	hectomètres noirs	tout le réseau	hectomètres noirs
Simple concordance	0,89	0,44	0,89	0,44
Jaccard	0,30	0,30	0,29	0,29
Soerensen	0,46	0,46	0,44	0,44
variante de Jaccard	0,56	0,56	0,54	0,54
Sokal & Sneath	0,18	0,18	0,17	0,17
Kulczynski	0,43	0,43	0,40	0,40
variante de Kulczynski	0,46	0,46	0,44	0,44
Ochiai	0,46	0,46	0,44	0,44
Faith	0,47	0,34	0,47	0,33

Les indices de similarité sont présentés dans le tableau 5. L'indice de simple concordance, qui tient compte des doubles absences, révèle 89 % de similarité, si l'on considère tout le réseau routier numéroté pour l'ensemble des périodes, et 44 % si l'on ne considère que les hectomètres appartenant à une zone noire pour une période au moins. Par contre, pour les indices qui ne tiennent pas compte des doubles absences (*a*), les taux de similarité sont plus faibles, mais importants à considérer. Ainsi, l'indice de Jaccard nous donne 30 % et 29 % de similarité entre périodes prises respectivement deux à deux. Lorsque l'on donne un poids plus important aux doubles occurrences (*d*), on obtient 46 et 44 % pour l'indice de similarité de Soerensen et 56 et 54 % avec la variante de Jaccard.

L'indice de Sokal & Sneath, qui donne un double poids aux différences (*b*) et (*c*), c'est-à-dire (1-0) et (0-1), conduit à la similarité la plus faible (indice de 0,18 et 0,17 respectivement). Par contre, quand les différences sont opposées aux doubles présences (*d*), comme dans l'indice de Kulczynski, les périodes de temps présentent des similarités de 43 et 40 % respectivement.

Les indices comparant les doubles présences à leurs totaux marginaux (variante de Kulczynski et indice de Ochiai) indiquent tous deux des similarités de 46 et 44 % respectivement pour les deux comparaisons. Enfin, l'indice de Faith, qui tient comptent des doubles absences (*a*) pondérées par 1/2, établit une similarité de 47 % pour tout le réseau sur toutes les périodes de temps et de 0,34 et 0,33 pour les hectomètres appartenant à une zone noire uniquement.

Ces premiers résultats démontrent l'importance des doubles absences dans les mesures de similarité, surtout au travers de l'indice de simple concordance (près de 90 % de similarité). Tous les autres indices considérés indiquent une similarité globale des périodes d'environ 40 %, ce qui constitue une part non négligeable. Par conséquent, une stabilité des zones noires dans le temps existe et nous suggérons plus loin quels facteurs environnementaux seraient susceptibles d'être à l'origine de cette stabilité.

Le tableau 6 montre une association positive entre les deux couples de périodes pour l'ensemble du réseau routier numéroté, puisque les valeurs observées des doubles présences sont dans chaque cas plus grandes que les valeurs attendues. Ce qui signifie que les hectomètres appartenant à une zone noire pour la première période auront plus de risque de se retrouver comme hectomètres noirs dans la seconde. La relation entre deux périodes est conforme, va dans le même sens et traduit une certaine stabilité dans le temps. Dans ce cas-ci, la relation est essentiellement influencée par *N,* car on note la présence d'un grand nombre d'hectomètres n'appartenant à aucune zone noire.

Pour quantifier quelque peu ces relations entre les observations, nous nous sommes intéressés aux odds ratios, qui sont un facteur multiplicatif du risque d'observer un hectomètre noir (1) ou non (0) au cours d'une période par rapport au risque observé au cours de la période précédente. Le tableau 6 montre qu'il y a 7,45 fois plus de risque pour un hectomètre de présenter la valeur 1 pour la période 1994-1996, s'il présente déjà la valeur 1 plutôt que 0 sur la période 1991-1993. Cet odds ratio est de 7,61 entre 1994-1996 et 1997-1999. Le risque pour un hectomètre d'appartenir à une zone noire sur une période donnée est plus élevé, s'il était déjà dangereux sur la période antérieure.

Tableau 6 : Indices d'association et odds ratios pour les deux couples de périodes comparées, pour les 4 604 hectomètres de routes numérotées

	Association		**Odds ratio**
Période	valeur attendue $(d+b)\cdot(d+c)/N$	valeur observée d	
1991-1993/1994-1996	51,92	225	7,45
1994-1997/1997-1999	48,18	209	7,61

Cette première analyse globale montre des répartitions spatiales relativement similaires dans le temps : la stabilité exprimée par les indices de similarité est confirmée par une relation d'association positive entre les périodes de temps, quantifiée par les odds ratios. Nous pouvons dès lors conclure que l'espace a un pouvoir structurant non négligeable.

Nous présentons à présent quelques caractéristiques physiques du milieu qui appuient cette tendance à la stabilité et qui annoncent une analyse explicative ultérieure.

3.2. Facteurs potentiels de stabilité spatiale

Nous analysons ici la stabilité des zones noires selon quelques variables environnementales. L'objectif est de montrer qu'au travers des années, certaines caractéristiques physiques du milieu induisent des facteurs de risque difficiles à maîtriser : les zones noires perdurent dans l'espace et le temps.

Nous parlerons dorénavant de similarité globale entre périodes de temps en considérant la moyenne des indices de similarité afin de comparer les périodes de temps plus facilement.

3.2.1. Le type de route

Nous considérons ici séparément les autoroutes et les autres routes (tableau 7).

Le nombre total d'hectomètres d'autoroutes du Brabant wallon est de 747 et le nombre d'hectomètres d'autoroutes appartenant au moins une fois à une zone noire est de 245 (soit 32,8 %, alors que ce pourcentage était de 20,4 % toutes routes confondues !).

Tableau 7 Tableaux de contingence pour les deux couples de périodes comparées, pour les 747 hectomètres d'autoroutes d'une part, les 3 857 hectomètres de routes nationales d'autre part

Sources des données : (MET, 2000)

	91-93			
94-96		0	1	
	0	562	50	*612*
	1	90	45	*135*
		652	*95*	**747**

	94-96			
97-99		0	1	
	0	527	57	*584*
	1	85	78	*163*
		612	*135*	**747**

	91-93			
94-96		0	1	
	0	3288	238	*3526*
	1	151	180	*331*
		3439	*418*	**3857**

	94-96			
97-99		0	1	
	0	3344	200	*3544*
	1	182	131	*313*
		3526	*331*	**3857**

Par rapport aux résultats de la section 3.1 qui concernait le réseau dans son ensemble, la similarité globale diminue pour le couple de périodes 1991-1993/1994-1996 de 0,47 (tout le réseau) à 0,40 (autoroutes) si l'on considère tous les indices en moyenne ; de 0,40 (hectomètres noirs) à 0,35 (hectomètres d'autoroute noirs) si l'on considère tous les indices en moyenne. Cela est dû principalement à l'indice de similarité de Kulczynski qui oppose les doubles présences aux différences (*b*) et (*c*), passé de 43 à 32 %, ce qui veut dire que, pour les autoroutes, la part des hectomètres appartenant à une zone noire pour une période et pas pour l'autre (1-0) et vice-versa (0-1) est grande par rapport aux hectomètres de zones noires communes aux deux (1-1). La similarité globale — et donc la stabilité — est diminuée dans ce cas-ci du fait de cette disparité. Par contre, une légère augmentation de la similarité globale est observée entre 1994-1996 et 1997-1999 (de 0,45 à 0,51 en moyenne pour tout le réseau ; de 0,39 à 0,46 en moyenne pour les hectomètres noirs uniquement), c'est-à-dire près de 50 % de stabilité en moyenne.

Pour les mesures d'association, nous avons une relation positive comme pour le cas général. En ce qui concerne les odds ratios, l'interprétation est la même que pour les résultats de la section 3.1, les chiffres étant cependant moins élevés pour les hectomètres d'autoroutes : *OR* de 3,43 et 4,16 respectivement pour les deux couples de périodes, à comparer à 7,45 et 7,61 pour le réseau tout entier ; par rapport au cas général les odds ratios indiquent donc une diminution dans le temps du risque pour un hectomètre de présenter la valeur 1 pour la période 1994-1996 (ou 1997-1999) s'il présente déjà la valeur 1 plutôt que 0 en 1991-1993 (ou 1994-1996).

Le nombre total d'hectomètres de routes nationales numérotées du Brabant wallon est de 3 857 et le nombre d'hectomètres de routes nationales numérotées appartenant au moins une fois à une zone noire est de 696 (soit 18,0 % contre 20,4 % toutes routes confondues).

Pour les routes nationales numérotées, les résultats sont semblables à ceux obtenus avec le réseau entier, puisque la similarité globale passe de 0,47 à 0,49 en moyenne pour le réseau de nationales et de 0,40 à 0,42 en moyenne pour les hectomètres noirs de routes nationales. Par rapport au cas général, nous observons donc que la stabilité pour les routes nationales est un peu meilleure pour 1991-1993/1994-1996 et un peu moins bonne pour 1994-1996/1997-1999.

La relation d'association reste positive pour le réseau de routes nationales. L'odds ratio est de 9,81 pour le réseau de routes nationales pour la comparaison 1991-1993/1994-1996, donc nettement supérieur à la valeur 7,45 obtenue dans le cas général pour la même comparaison. Le risque de rencontrer un hectomètre noir est plus stable entre les deux premières périodes.

En ce qui concerne le caractère stable des zones noires, nous remarquons donc une légère différence selon le type de route : les différents indices de similarité concluent à une stabilité plus élevée pour les autoroutes que pour les routes nationales pour 1994-1996/1997-1999.

3.2.2. Les caractéristiques physiques des routes

Nous avons considéré les autoroutes et les routes à berme centrale (au moins 2 □ 2 voies de circulation) séparément des routes sans berme centrale (tableau 8).

Le nombre d'hectomètres de routes à berme centrale (autoroutes comprises) numérotées du Brabant wallon est de 1 563 et le nombre d'hectomètres de routes à berme centrale appartenant au moins une fois à une zone noire est de 468 (soit 29,9 % contre 20,4 % toutes routes confondues).

Tableau 8 Tableaux de contingence pour les deux couples de périodes comparées, pour les 1 563 hectomètres d'autoroutes et routes à berme centrale d'une part, les 3 041 hectomètres de routes sans berme centrale d'autre part
Source des données : (MET, 2000)

	91-93			
94-96		0	1	
	0	1192	140	*1332*
	1	127	104	*231*
		1319	*244*	**1563**

	94-96			
97-99		0	1	
	0	1190	90	*1280*
	1	142	141	*283*
		1332	*231*	**1563**

	91-93			
94-96		0	1	
	0	2658	148	*2806*
	1	114	121	*235*
		2772	*269*	**3041**

	94-96			
97-99		0	1	
	0	2681	167	*2848*
	1	125	68	*193*
		2806	*235*	**3041**

La plus grande différence par rapport au cas général (tout le réseau) intervient pour la comparaison 1994-1996/1997-1999 : les indices de similarité (Kulczynski et variante de Jaccard entre autres) passent de 0,45 à 0,54 si l'on considère tous les indices en moyenne pour le réseau de routes à berme centrale dans son ensemble et de 0,39 à 0,49 en moyenne pour les hectomètres noirs de routes à berme centrale, soit plus de 50 % de stabilité pour ce couple de périodes. La relation d'association est positive dans ces mêmes conditions.

Le nombre d'hectomètres de routes sans berme centrale numérotées du Brabant wallon est de 3 041 et le nombre d'hectomètres de routes sans berme centrale appartenant au moins une fois à une zone noire est de 473 (soit 15,6 % contre 20,4 % toutes routes confondues).

Pour les routes sans berme centrale, l'odds ratio est de 10,94 pour la comparaison 1991-1993/1994-1996, plus élevé donc que dans le cas général.

Une différenciation fondée sur les caractéristiques physiques des routes conduit donc à identifier plus de stabilité dans la distribution spatiale des accidents des dernières années. Elle est plus marquée encore pour les autoroutes et les routes à berme centrale, pour lesquelles existe un déterminisme spatial non négligeable, principalement pour les dernières années d'étude.

Nous allons vérifier dans la section suivante si l'intensité du trafic renforce cette tendance pour cette période et apporte un changement pour les premières années de l'étude.

3.2.3. L'intensité du trafic

Les données disponibles concernant les flux de véhicules sont des moyennes journalières (de 6 h à 22 h) dans les deux sens de circulation, tous types de véhicules confondus (MET, 2000). Elles ne sont collectées ni au moment, ni sur le lieu de l'accident. Les données de trafic sont donc extrapolées à partir des points de comptage par segment de route, pour chacune des trois périodes de temps. Les flux de trafic sont des moyennes annuelles pour chaque période étudiée. Nous avons considéré, pour construire deux sous-ensembles, un seuil de trafic de 17 300 véhicules qui correspond au deuxième quartile (ou médiane) de la matrice des flux (tableau 9).

Le nombre d'hectomètres numérotés du Brabant wallon supportant un trafic journalier de plus de 17 300 véhicules est de 985 et le nombre d'hectomètres noirs supportant un trafic journalier de plus de 17 300 véhicules est de 347 (soit 35,2 % contre 20,4 % tous trafics confondus).

Le nombre d'hectomètres numérotés du Brabant wallon supportant un trafic journalier de moins de 17 300 véhicules est de 3 619. et le nombre d'hectomètres de routes supportant un trafic journalier de moins de 17 300 véhicules et appartenant au moins une fois à une zone noire est de 594 (soit 16,4 % contre 20,4 % toutes routes confondues).

Tableau 9 Tableaux de contingence pour les deux couples de périodes comparées, pour les 3 619 hectomètres supportant un trafic moyen journalier inférieur à 17 300 véhicules d'une part, pour les 985 hectomètres supportant un trafic moyen journalier supérieur à 17 300 véhicules d'autre part

Source des données : (MET, 2000)

	91-93			
94-96		0	1	
	0	3148	201	*3349*
	1	125	145	*270*
		3273	*346*	**3619**

	94-96			
97-99		0	1	
	0	3179	187	*3366*
	1	170	83	*253*
		3349	*270*	**3619**

	91-93			
94-96		0	1	
	0	702	87	*789*
	1	116	80	*196*
		818	*167*	**985**

	94-96			
97-99		0	1	
	0	692	70	*762*
	1	97	126	*223*
		789	*196*	**985**

Différencier selon le trafic conduit à des résultats intéressants : ainsi le taux moyen de similarité est de près de 60 % pour un seuil de trafic supérieur à 17 300 véhicules, toutes routes confondues et pour 1994-1996/1997-1999. L'indice de Kulczynski entre autres présente un score de 0,75, qui s'explique par le nombre élevé des doubles présences (126 hectomètres présents dans les deux périodes) par rapport aux différences (70 hectomètres présents sur la première période, mais pas sur la seconde et 97 hectomètres absents de la première période, mais présents sur la seconde). La relation d'association est positive pour ce couple de périodes, la valeur observée (*d*), égale à 126, étant supérieure aux valeurs attendues, égales à 44,37 et 125,35 pour tous les hectomètres et les hectomètres noirs respectivement. Les odds ratios les plus élevés sont observés dans le cas d'un trafic inférieur à 17 300 véhicules et pour la comparaison 1991-1993/1994-1996.

Les autoroutes et les autres routes présentent cependant des conditions de trafic très différentes, puisque le trafic moyen journalier varie entre 81 400 et 1 200 véhicules selon le cas. Dès lors, nous avons réalisé une nouvelle analyse en choisissant des seuils de trafic différents selon le type d'aménagement considéré. Cette analyse porte sur les seuls hectomètres appartenant à une zone noire (*N* = 941) en raison de l'insuffisance des données de trafic pour tout le réseau. Ces hectomètres sont répartis en croisant différents types d'aménagement et seuils de trafic : routes à berme centrale et trafic journalier supérieur ou inférieur à 24 000 véhicules (3^{e} quartile de la matrice des flux) d'une part, routes sans berme centrale et trafic journalier supérieur ou inférieur à 9 400 véhicules (1er quartile de la matrice des flux) d'autre part.

Les résultats montrent qu'une forte stabilité existe pour les routes à berme centrale supportant un trafic moyen journalier supérieur à 24 000 véhicules entre 1994-1996 et 1997-1999 (53 % de similarité en moyenne par rapport aux 39 % en moyenne dans le cas général).

Un taux de similarité élevé (soit près de 55 % en moyenne par rapport aux 40 % du cas général) est observé également pour les routes sans berme centrale ayant un trafic supérieur à 9 400 véhicules entre 1991-1993 et 1994-1996. La relation d'association positive renforce cette tendance à la stabilité dans ce dernier cas.

Les seuils de trafic proposés sont donc pertinents pour conforter la tendance d'une stabilité sur les routes les plus importantes. En dessous de ces seuils, les concentrations d'accidents sont considérées comme spatialement instables et donc aléatoires (de l'ordre de 25 à 40 % dans les cas autres que ceux cités ci-dessus).

3.2.4. L'environnement

Certains types d'environnement seraient-ils propices aux ancrages spatiaux ? C'est l'hypothèse que nous voulons tester en considérant les environnements suivants : le bâti dense, le bâti lâche (bâti uniquement le long des routes), le rural boisé (routes reliant deux villages séparées par des bois) et le milieu rural ouvert (routes reliant deux villages séparées par des champs). Au moyen des techniques des systèmes d'information géographique (SIG), nous avons donc attribué à chaque hectomètre du réseau du Brabant wallon un type d'environnement, d'après la carte numérique d'affectation du sol au 1/50 000 (IGN, 2002), couplée à celle du réseau routier (MET, 2000). Le réseau routier est ici considéré dans sa totalité (*N* = 4 604).

Le milieu urbain dense présente la similarité globale moyenne la plus élevée (près de 60 % pour 1991-1993/1994-1996 et près de 50 % pour 1994-1996/1997-1999), du fait du grand nombre de doubles présences par rapport aux autres occurrences pour la première paire de périodes. Les zones noires se stabilisent très fortement dans un milieu urbain dense. Rappelons cependant que ce dernier comprend beaucoup de routes non numérotées que nous n'avons pu prendre en compte ici.

Près de 50 % de tous les hectomètres du Brabant wallon se trouvent en milieu rural ouvert ; nous y observons une bonne stabilité dans le temps, semblable au cas général (soit près de 40 %). La stabilité, moins fortement marquée qu'en milieu urbain cependant, est principalement due aux doubles absences. Les odds ratios de ce milieu sont les plus élevés de tous les environnements considérés (*OR* = 9,99 entre 1991-1993 et 1994-1996 ; *OR* = 9,75 entre 1994-1996 et 1997-1999).

4. Conclusions

Le but de notre étude était de vérifier la stabilité ou l'instabilité dans le temps et dans l'espace des zones noires du Brabant wallon. Si nous nous reportons à la carte réalisée pour les trois périodes de temps que nous avons considérées (Fig. 2), nous constatons visuellement une stabilité des zones noires, notamment sur les grands axes autoroutiers nord-sud, et une part d'instabilité, explicable ou non, sur certaines routes nationales secondaires.

Pour répondre à l'inadéquation d'indicateurs traditionnels de similarité statistique tel le χ^2 pour tester la stabilité temporelle et spatiale des zones noires, nous avons été amenés à utiliser d'autres indices, qui surmontent la difficulté due à la disproportion créée par le grand nombre d'hectomètres n'appartenant à aucune zone noire. Cette méthode a conduit aux résultats suivants.

Globalement, le taux de stabilité moyen observé est de 40 % entre deux périodes successives, ce qui confirme statistiquement la simple comparaison sur carte.

Si l'on se réfère à des facteurs environnementaux et d'infrastructure, les stabilités les plus élevées (entre 50 et 60 %) correspondent aux caractéristiques suivantes :

– autoroutes et périphériques (1994-1996/1997-1999) ;

– trafic supérieur à 17 300 véhicules (1994-1996/1997-1999) ;

– routes à berme centrale supportant un trafic journalier supérieur à 24 000 véhicules (1994-1996/1997-1999) ;

– routes sans berme centrale supportant un trafic journalier supérieur à 9 400 véhicules (1991-1993/1994-1996) ;

– environnement urbain dense (1991-1993/1994-1996) ;

– milieu rural ouvert.

Ces résultats plus fins confortent à nouveau la tendance observée sur les cartes. Ainsi les axes routiers les plus importants constituent-ils une structure spatiale forte pour les deux dernières périodes (1994-1996/1997-1999). En milieu urbain dense et sur les routes secondaires, les similarités — et donc la stabilité — sont plus marquées pour les périodes 1991-1993/1994-1996 : les aménagements réalisés au cours des dernières années semblent avoir amélioré la sécurité dans ces zones.

Les hectomètres du réseau qui appartiennent à des zones noires et qui présentent une stabilité moyenne observée de 40 % entre deux périodes successives constituent un noyau dur qu'il a été jusqu'à présent impossible d'améliorer sur le plan de la sécurité routière. Une autre partie est plus mobile, sans doute liée à la part d'aléatoire qui existe dans l'occurrence des accidents. Il ne faut pas oublier que l'intensité du trafic joue un rôle important : en moyenne, lorsque le trafic augmente, le nombre d'accidents augmente également. Cette relation est particulièrement vérifiée sur les autoroutes belges (Thomas, 1992).

Cette stabilité des zones noires dans le temps nous conforte dans l'idée que le choix de l'année importe peu dans les études géographiques qui traitent des accidents de la route, à paramètres environnementaux stables. Nous pouvons donc rejeter l'hypothèse que *la distribution géographique des zones noires varie fortement en fonction de la période de temps choisie*. Nous pouvons penser qu'il est fort probable que l'environnement explique tout ou partie de l'occurrence des accidents de la route, lesquels ont donc tendance à se concentrer toujours aux mêmes endroits.

Ces éléments sont extrêmement précieux pour l'élaboration d'une politique de mobilité durable, dont la sécurité est un paramètre incontournable. En effet, la sécurité routière présente de multiples aspects, l'un d'entre eux concernant l'environnement dans lequel se produisent les accidents. L'importance de ce facteur a été soulignée ici.

REMERCIEMENTS NOUS REMERCIONS LA POLITIQUE SCIENTIFIQUE FEDERALE POUR LE FINANCEMENT DE CETTE RECHERCHE (CONTRAT N°CP/10/342). NOUS REMERCIONS MME STEENBERGHEN DE LA KATHOLIEKE UNIVERSITEIT VAN LEUVEN ET M. ANTOINE DU MINISTERE WALLON DES ÉQUIPEMENTS ET DES TRANSPORTS POUR UNE RELECTURE EFFICACE ET CRITIQUE D'UNE PREMIERE VERSION DE L'ARTICLE, AINSI QUE M. ARNOLD POUR SES PRECIEUX CONSEILS.

Références

Eckhardt, N., Thomas, I., 2004. Techniques innovatrices d'analyse spatiale en matière de sécurité routière. Rapport des Services fédéraux des affaires scientifiques, techniques et culturelles, Louvain-la-Neuve, Belgique (à paraître).

Everitt, B.S., 1977. The analysis of contingency tables. Chapman and Hall, Grande-Bretagne.

Flahaut, B., Thomas, I., 2002. Identifier les zones noires d'un réseau routier par l'autocorrélation spatiale locale, Analyses de sensibilité et aspects opérationnels. Revue internationale de géomatique, 12(2), 245-261.

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. Accident Analysis & Prevention, 35(6), 991-1004.

Fleiss, J.L., 1981. Statistical methods for rates and proportions. John Wiley & Sons, États-Unis.

Gatrell, A.C., 1983. Distance and space: a geographical perspective. Clarendon Press, Oxford, Grande-Bretagne.

Hauer, E., 1996. Identification of sites with promise. Transportation Research Record, 1542, 54-60.

IGN, 2002. Banques de données et cartes topographiques (Top50r). Institut géographique national, Bruxelles, Belgique.

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

Kent, M., Coker, P., 1996. Vegetation description and analysis, A practical approach. Wiley Chichester, Grande-Bretagne.

Legendre, P., Legendre, L., 1998. Numerical ecology. Elsevier Science, Pays-Bas.

MET, 2000. Données de l'intensité du trafic et des caractéristiques physiques des routes. Rapport du ministère de l'Équipement et des Transports, Bruxelles, Belgique.

Nguyen, T.N., 1991. Identification of accident blackspot locations, an overview. VIC Roads/Safety Division. Research and Development Department, Australie.

Silcock, D.T., Smyth, A.W., 1985. Methods of identifying accidents blackspots. Transport Operations Research Group. Dept. of Civil Engineering, Université de Newcastle upon Tyne, Grande-Bretagne.

Stokes, M.E., Davis, C.S., Koch, G.G., 2000. Categorical data analysis using the SAS system. SAS Institute Inc., États-Unis.

Thomas, I., 1992. La relation trafic-accidents sur autoroute, Approche statistique empirique. Selected Proceedings of the Sixth World Conference on Transport Research, Politiques de transport, Lyon, France.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis & Prevention, 28(2), 251-264.

Thomas, I., Tulkens, H., Berquin, P., 2000. Quelles frontières pour Bruxelles ? La réponse d'un exercice statistique, géographique et économique. In : Combes, P.-P., Thomas, I. (Eds.), Les forces d'agglomération dans la métropolisation de l'économie. CIFOP, Rapport Commission 3, Quatorzième congrès des Économistes belges de langue française, 73-88.

van der Laan, L., Vogelzang, J., Schalke, R., 1998. Commuting in multi-modal systems: an empirical comparison of three alternative models. Journal of Economic and Social Geography (former TESG), 89(4), 384-400.

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.

H. Spatial nested scales for road accidents in the periphery of Brussels

Nathalie ECKHARDT [1],
Isabelle THOMAS [1,2,3]

October 6th 2004
Submitted to IATSS

[1] Department of Geography, U.C.L, Louvain-la-Neuve, Belgium.
[2] National Fund for Scientific Research, Brussels, Belgium
[3] Centre of Operation Research and Econometrics, Louvain-la-Neuve, Belgium

Corresponding author: isabelle@geog.ucl.ac.be

Abstract
This paper examines, by means of a multilevel model (MLM), how far the characteristics of the geographical environment influence the occurrence of road accidents at two levels of spatial aggregation. The results are compared to those obtained from a more classical logistic regression. The analysis is performed on data from the southern periphery of Brussels (Belgium). The main findings are: (1) that MLM is a potentially useful technique for modelling road accidents, but that hierarchical levels are not easy to define for spatial data and so MLM are less useful than other regression techniques; (2) that the characteristics of the environment and the road itself significantly influence the occurrence of road accidents, and changes in these characteristics are quite important elements in the explanation, leading to the suggestion that road users do not adapt their behaviour sufficiently to changes in road conditions. Hence, concentrations of road accidents often correspond to places where improvements could be made in terms of road design, signalling and land-use planning.

Key words: Multilevel model, geography, environment, accidents, Brussels.

1. Introduction

The main research objective of our team is to analyse the spatial aspects of road-accident occurrence at several scales of analysis [1, 2, 3, 4]. The present contribution aims at modelling the spatial occurrence of road accidents by considering different nested levels of spatial data aggregation.

Multilevel analysis is a recent technique that enables the relationships/interactions between variables at several levels of data aggregation to be examined simultaneously [5, 6]. It attempts to solve the dilemma of the spatial scale, that is to say the use of scales in analysing the characteristics of accidents at a local level, taking into account the broader spatial context in which these accidents occur. Up to now, applications have mainly been limited to the social and behavioural sciences. Let us mention here the well-known example of school test results, which can be explained by the individual characteristics of the scholars, but also by the characteristics of the class (group) as well as the school or even its environment [7]. Multilevel models are specifically dedicated to the concept of the integration of contextual effects and so to hierarchical models [6, 8, 9]. In spatial analysis, multilevel models enable the researcher to go beyond the scale defined *a priori* by territorial executives (for example), and to attempt to capture the continuous character of space, taking into account the nested nature of spatial scales [6, 10].

In the field of road safety, two recent papers have used multilevel modelling. Jones and Jorgensen [11] consider road accident casualties, and show that the risk of fatality is associated with casualty age and sex, as well as the type of vehicles involved, characteristics of the impact, attributes of the road section on which it took place, time of day, and whether alcohol was suspected. The multilevel analysis shows that 16% of the unexplained variation in casualty outcomes was between accidents, whilst approximately 1% was associated with the area in which each incident occurred. Gee and Takeuchi [12] analyse the cross-sectional relationship between traffic stress and neighbourhood conditions, depression and health status by means of multilevel analyses. They show that perceived traffic stress is associated with both general health status and depression in multilevel models, with people reporting traffic stress having lower health status and more depressive symptoms.

In this paper, we aim to show the utility of multilevel analysis for understanding the spatial aspects of road safety, and more particularly to show how far the characteristics of space (environment and infrastructure) can influence the location of accidents at different levels of measurements. An analysis is conducted on data from the suburbs of Brussels (southern periphery); the modelling results are interpreted in terms of operational results and are also compared to those obtained by means of a more "classical" logistic regression [3].

This paper is organised as follows. The model is described in Section 2 and choices related to the data are discussed in Section 3. Empirical results are reported in Section 4. Two levels of aggregation are considered: the hectometre (100 m) of road and the commune; we aim to understand why accidents are concentrated in some hectometres and why these hectometres are located in specific communal environments. Our conclusions and discussion are to be found in Section 5.

2. Multilevel analysis

We will first justify our modelling choices by briefly defining the type of model (Section 2.1), its advantages in terms of accident modelling (Section 2.2) as well as the methods used for estimating the parameters (Section 2.3). We refer to the literature for further model definition and formulation [6, 7, 9, 13, 14].

2.1 Definition

Multilevel modelling (MLM) is a type of regression that has mainly been developed since the 1980s; it is designed to handle hierarchical and clustered data. Such data involve group effects on individuals, which may not be assessed validly by traditional statistical techniques. That is, when grouping is present, observations within a group are often more similar than would be predicted on a pooled-data basis, and hence the assumption of independence of observations is violated. MLM uses variables at several levels of aggregation to adjust the regression of the base-level dependent variables on the base-level independent variables. In our case, for instance, we could predict accident occurrence from the characteristics of the hectometres or from larger environments such as communes. MLM is related to structural equation modelling in that it fits regression equations to the data, then tests alternative models using a likelihood ratio test.

MLM specifies the expected direct effects of variables on each other within any one level, and cross-level interaction effects between variables located at different levels. Hence, mediating mechanisms are postulated; they allow variables at one level to influence variables at another level (e.g. better road infrastructure may influence road-user behaviour and hence prevent accidents at places other than that where the road enhancement has been effected). MLM tests multilevel theories statistically, by simultaneously modelling variables at different levels without having recourse to aggregation or disaggregation. In our case, the global problem is to model the relationship between the place of an accident and the context in which it occurs. We hence aim to detect the amount of context contribution and its effect on the total variation of the "individual behaviour", and to identify which macro characteristics are responsible for the context effect. This means conceptually introducing a multilevel approach in which road accidents are grouped together at different spatial levels; in our case, variables from two levels will be jointly analysed in a unified framework.

The multilevel model is therefore a model with a single dependent variable (*Y*) measured at the base level (e.g. the accidents are taken individually). As in ordinary least squares (OLS) regression, there may be one or more independent variables collected at the base level. In addition, there will be at least one broader level of aggregation, with at least one explanatory variable (for instance, the environment or the socio-economic characteristics of the ward). In an OLS model, base-level data are analysed for all groups pooled together (e.g. all hectometres, all communes). In a MLM, the regression is performed separately for each group. This produces different regression coefficients and different intercepts (e.g. for each black zone or each commune) and also explains why MMLs are called "random coefficient models". Such models usually use maximum likelihood algorithms to estimate the parameters (coefficients) (see http://www2.chass.ncsu.edu/garson/pa765/multilevel.htm).

Let us briefly summarise the formulation. Y_{ij} is the variable to be explained. We consider here a model at two levels of observation (see Section 2.2): the hectometre *i* (Level 1) localised in a municipality *j* (Level 2). The general shape of the linear model includes an explanatory variable at Level 1 related to the hectometre (X_i), and a contextual variable at Level 2 related to the commune (municipality) and denoted Z_j. Then

$$Y_{ij}=\beta_0+\beta_i X_i+\Gamma_j Z_j+(\mu_{0j}+\mu_{ij}*X_{ij}+\varepsilon_{ij})$$

where $\beta_0+\beta_i X_i+\Gamma_j Z_j$ is the *fixed part* of the model and $(\mu_{0j}+\mu_{ij}*X_{ij}+\varepsilon_{ij})$ is the *random part*. β_0 is the intercept. β_i is the angular coefficient of the right-hand side of the regression; it is the coefficient of the explanatory variables at Level 1 (X_i); Γ_j is the coefficient of the Level 2 explanatory variable (Z_j). *Error terms* (residuals) are associated with β_0 and β_i at the contextual level (μ_{0j} and μ_{ij}); they represent the deviation of the municipality *j* from the average coefficient. These contextual residuals are assumed to follow a normal distribution law with null averages, variances $\sigma_{0\mu}^2$ and $\sigma_{i\mu}^2$ and covariance $\sigma_{0i\mu}$. Level 1 residuals (ε_{ij}) have a null-average and equal $\sigma_{0\varepsilon}^2$ variances. Hence, this formulation allows every municipality *j* to have its own constant (β_0)

and angular coefficient (β_i). This heterogeneity of the regression coefficients between municipalities can later be tested and explained by Level 1 and Level 2 variables. In other words, the *outputs* of a multilevel model include: (1) a fixed part containing the regression coefficients and the corresponding *p*-value for the significance levels, at Level 1, Level 2, and for the cross-level interactions; (2) a random part containing regression coefficients and *p*-values for estimating the variances of Level 1 variables and intercepts; and (3) standardised (beta) coefficients which, as in OLS regression, allow us to compare the relative importance of the independent variables and interactions. Most software packages s also produce a measure of deviance.

2.2 Advantages of spatial analysis

Two types of errors are avoided when different levels of aggregation are considered simultaneously [6, 15, 16, 17]: the *ecological error* which consists of using a global aggregated statistical measure to reveal individual behaviour; and the *atomist error*, which considers the characteristics of the individual but ignores the context in which the human behaviour occurs. This is also true for road accidents: indeed it seems fallacious to isolate the accident from its environment, or the society in which it occurred. The purpose of this paper is to determine the direct effect of the explanatory variables measured at a low level of aggregation and at a higher level of aggregation, and to see if the explanatory variables at the aggregated level moderate the relationships occurring at the individual level, or vice versa.

Levels of analysis are often hierarchically organised: items at one level interact and create a higher homogeneity [8]. This hierarchical structure leads to a correlation of the observations that violates the hypothesis of the independence of residuals (classical regression techniques) and leads to an underestimation of the standard deviations of the regression coefficients. MLMs take this correlation into account in the estimation of the standard deviations of the regression coefficients by including terms of error at the contextual level. With regard to our study, this dependence is spatial and means that the observations supply less information than if they were randomly distributed, as is assumed in the OLS method [18].

Additional advantages of MLM are: (1) regression coefficients are specific to each level of analysis thanks to the contextual residuals; (2) it is possible to test whether the variance in the contextual terms of error is significantly different from 0 (likelihood tests); (3) coefficients of determination (R^2) can be computed for each level of analysis by comparing the residual variances with the variances of an "empty model" (without explanatory variable) for each level; and (4) it is possible to attribute the residual variance of the classical multiple regression to various levels of analysis [6, 10]. The assumptions for the application of MLMs are in general less restrictive than for more traditional regression techniques [6, 7, 9, 10]; the main difficulty with MLMs is the definition of the hierarchical levels of observation and the associated variables (see Section 3.2).

2.3 Estimating the parameters

Homoscedasticity (equal variances) is often violated in hierarchical situations: OLS techniques are hence inappropriate. Multilevel linear models are often best estimated by the Newton-Raphson algorithm that is based on maximum likelihood and generalised least squares. The restricted maximum likelihood method is often used. The values of the regression coefficients are first computed on the basis of the first analysis of the matrix of variance-covariance. The matrix is then re-estimated, using the first values of the coefficients. Finally, the estimation of the coefficients is improved by the new variance-covariance matrix until convergence. In this way, the fixed and random parts of the model are more effectively estimated than with OLS.

Several software packages, such as *HLM* and *MlwiN*, are available [19, 20]. SAS was used (*Proc Mixed, Proc Nlmixed*) for the research reported in this paper; this package enables the more classic regression models to be considered as well as the hierarchical formulations. Several models can be implemented with *Proc Mixed*: simple random-effects only, simple mixed with a single fixed and random effect, split-plot, multilocation, repeated measures, analysis of covariance, random coefficients, and spatial correlation.

3. Methodological choices

3.1 Road safety in the area studied

Brussels is the capital city of Belgium, located in the centre of the country and containing approximately 1 million inhabitants. As in most urban areas, the city sprawls far beyond its administrative boundaries. Walloon Brabant corresponds to an administrative entity (province) located in the south of the city; it is mainly peri-urban but its landscape results from its historical evolution. Between the 15th and 18th centuries it was mainly rural, with many small villages. During the 19th century, industries located in the west (ironworks) and centre (paper mills). Railways and better roads later increased the accessibility. In the first half of the 20th century, industries started closing one after the other. In the sixties, the region was increasingly polarised by Brussels: as in many European cities, people started to move from the centre of Brussels to the countryside, while keeping their jobs in the city. Later on a university was created in Louvain-la-Neuve and several industrial parks were planned all over the area. Hence, the southern periphery of Brussels is now characterised by old villages and small towns, a new town and many allotments, old industrial locations as well as new planned ones (industrial parks), highly urbanised communes close to Brussels as well as more residential areas, woods and agriculture as well as employment and commercial centres. The result was a mosaic of landscapes, a polynuclear structure and quite an interesting spatial pattern [21, 22].

In Belgium, any road accident that occurs on a public road and that involves casualties must be officially reported. Location is accurately known on numbered roads: there is a stone marker every hectometre (100 meters); numbered roads are motorways, national and provincial roads linking towns together. On other roads, location is identified by postal addresses that are often less accurate. This analysis is limited to accidents with casualties on numbered roads; the hectometre is the smallest spatial unit for which accident data are spatially and officially available.

Walloon Brabant has 460.4 km of numbered roads including 37.7 km of motorway. In this paper, a black zone is defined by means of local spatial autocorrelation indices [2, 23, 24] as a set of contiguous hectometres with a high number of accidents. Black zones vary in length and intensity, and some black zones may include a hectometre with no or very few accidents. These black zones seem characterise the same place from year to year [25], despite the many socio-economic changes. In the area being studied, 47% of the road hectometres did not register any accidents. Black zones represent 38% of the total number of accidents but only 12% of the total number of hectometres. In this paper, the period under study is 1998–2000, a period long enough to minimise random fluctuations but short enough to limit changes in road traffic conditions. Some 2,363 accidents were registered on numbered roads during this period; in total, 1,388 hectometres out of 4,604 experienced at least one accident with casualties.

3.2 Levels of analysis and dependent variables

The definition of the hierarchical levels is not as easy as for some topics in human sciences. It is a compromise between the significance and the availability of the data.

Ideally, accidents should be considered as the first level of analysis. However, accident data are collected by hectometre (no GPS was available). Moreover, most environmental variables are also only available for hectometres or segments of roads rather than for pinpoint locations on the road. Let us here add that some environmental characteristics are collected at the time and place of an accident and are reported on the statistical form filled in for each accident with casualties. These could have been used for characterising the places at which accidents took place, but their quality is doubtful [26] and we do not have any comparable information about places where no accidents occurred. Hence the lowest aggregated level of analysis used here is the hectometre.

Initially, we decided to take the road segment (several hectometres long) as the second level of analysis. However, it was not easy to choose the best length for this segment or to justify it. Thresholds would have been quite artificial and such segmentation does not correspond to any official definition; explanatory data are not available at this level, and they would have had to be a combination of data available at the hectometre level. Hence, the municipality *j* (commune) was chosen as the second level of analysis. It is not directly related to the road network itself, but to its global environment. However a lot of data are available at this level of administrative aggregation, which corresponds quite well to mobility patterns in Belgium.

Three dependent variables are used and modelled separately: *Y1* takes the value 1 when a hectometre belongs to a black zone and 0 otherwise. It enables us to understand why some hectometres are more dangerous than others. *Y2* is the total number of accidents observed in each hectometre of road, and *Y3* is a measure of the risk of accidents, roughly estimated as the total number of accidents divided by the average daily traffic intensity. *Y2* has the advantage of giving the real number of accidents: some hectometres can belong to a black zone (*Y1*) without any accidents being recording [2], and, on the contrary, a hectometre with several accidents can be surrounded by hectometres with no accidents and hence not belong to a black zone. The absolute number of accidents (*Y2*) is interesting for some public authorities (such as the emergency services), while for others, the relative number (*Y3*) is important. For each *Y* variable, the situation on motorways is modelled separately because traffic has a different structure (two separate lanes, etc.) on motorways.

3.3 Explanatory variables

The purpose here is to identify environmental conditions associated with road safety/danger at two nested levels of analysis. Most explanatory variables are selected from official data bases (Belgian National Institute of Statistics, Ministry of Equipment and Transport) or constructed by means of GIS techniques from official IGN maps [27]; some of these variables are identical with or close to those used by Flahaut [3]. They are briefly described below.

Most explanatory variables used at Level 1 describe the environment of the hectometre itself: the physical characteristics of the road, land use or natural environment of the hectometre (see Table 1). Changes in environmental conditions are also considered by means of so-called "transition variables" which pinpoint road and environmental discontinuities that could influence drivers' ability (or inability!) to cope with changes. These variables are referenced by the absolute distance to the spatial discontinuity and are recorded as *_dist*. If the change is within the hectometre, the distance is recorded as 0 metres. A maximum of 200 m is considered on numbered roads and 300 m on motorways. Information on traffic, speed limits and the physical characteristics of the roads (type of road, type of surface, adherence of the road, presence of rutting or obstacles) were made available by the Ministry of Equipment and Transport (MET). Traffic density is measured by the 1999 average daily number of vehicles in both directions, including all types of vehicle, between 6 a.m. and 10 p.m. Traffic density is measured for each hectometre and extrapolated to larger road segments, but is never known at the time and place of the accident. It is a daily average. Land use variables refer to human activities or to measurable characteristics of the physical environment (e.g. % of area built-up, level of afforestation along the road at a distance of 50 m). These are constructed on the basis of the 1:50,000 digital maps provided by the National Geographic Institute. Finally, the orientation, bends and slopes of the roads are obtained from measurements made on digitalised topographic maps by means of appropriate GIS techniques. These characteristics of the roads can certainly influence the visibility and the behaviour of road users. Most variables are (0,1) variables.

	Variables	*Description*
Road use	TRAFFIC	Logarithm of the average daily volume of traffic (Source: MET)
	VMAX	Maximum speed limit (Source: MET)
	VMAX_*dist*	Distance to a change in speed limit (in meters): >200; 200; 100; 0 m and 300 m on motorways. Written as VMAX(>200)_*dist* etc.
Physical characteristics of the road	LANES	Number of lanes: (1) 1+1; (2) 2+2; (3) 2; (4) 3
	LANES_*dist*	Distance to a change in road type in terms of number of lanes: >200; 200; 100; 0 m and 300 m on motorways. Written as LANES(100m)_*dist* etc.
	SURFACE	Type of surfacing (1) BE: concrete; (2) HFN: conventional asphalt; (3) HOM: thin asphalt; (4) HON: draining asphalt
	SURFACE_*dist*	Distance to a change in surfacing: >200; 200; 100; 0 m for numbered roads and 300 m on motorways
	RUT	Presence (1) / absence (0) of ruts
	JUNCT_*dist*	Distance to major crossroad: >200; 200; 100; 0 m on numbered roads and >300 ; 300 m on motorways
	ADHERENCE	Adherence of the road surface: (0) good or normal; (1) bad or very bad
	PROX_ACCES	Proximity (<300 m) to entry/exit. (0,1) On motorways only
	BZONE	Hectometre belonging (1) or not (0) to a black zone
Land-use	BUILT	Estimated % of built-up area (from the road to 50 m from it): <20; 20; 25; 30; 40; 50. Written as Built30 etc.
	BUILT_*dist*	Distance to transition in density of built-up area (BUILT): >200; 200; 100; 0 m (>300; 300 m on motorways)
	WOODS	Estimated % of wooded area along the road (at 50 m): < 20; 20; 25; 30; 40; 50. Written as Woods30 etc.
	FIRMS	Proximity of firms or large supermarkets (50 m) (0,1)
	OBSTA _*dist*	Distance to an obstacle such as a bridge pillar: >200; 200; 100; 0 m (>300; 300 m on motorways)
Landscape	AGGLO	Inside/outside an urban agglomeration (F1/F3 road sign)
	DIRECTION	Road segment direction: (1) Others (N-S); (2) E–W ≤ 22.5°; (3) 22.5°<E–W ≤ 45°
	RELIEF	At the top of a slope; at 100 m from the top; at the bottom of a slope; at 100m from the bottom; other

Table 1: Explanatory variables at Level 1

At level 2, explanatory variables mainly pertain to the socio-economic characteristics of the population living, working or shopping in the commune (Census data). Table 2 contains the definitions of the variables used in this paper. As in most human geography studies, it is difficult to estimate the population at risk who are really present in the municipality or traversing it. We limited ourselves to proxies that are officially available. Structural variables were also taken into account: the level of urbanisation[28] and the rurality (as expressed by the ratio of agricultural areas to the total land area of a municipality). We also feel that the morphology of the built-up environment could influence accident occurrence; morphology was estimated by fractal dimension *D* [21, 22]. Let us remember that D describes the extent to which a mass (here the built-up area) is concentrated within a zone. Fractal dimension is not equal to density [22]. Thus for spatial mass distributions, *D* can be interpreted as a measure of mass concentration in a given area. It can be shown that a value of *D* close to 2.0 describes a fairly homogeneous distribution. The lower the value of *D*, the more the mass is concentrated: thus a dimension close to 0.0 corresponds to a concentration of the mass in one isolated point, while the value 1.0 corresponds to a line, but also characterises a hierarchical spatial distribution of masses. No value smaller that 1.0 was obtained in this study. Such values would refer to structures composed of a disconnected set of points. It is possible that road safety is affected by whether built-up areas are distributed homogeneously or heterogeneously throughout the neighbourhood.

Variable	*Description*
MIXITY	Number of jobs/number of inhabitants in a commune in 2001 (= level of mixing of the activities and hence traffic).
ATTRACT	(Numbers of jobs + number of inhabitants)/surface of the commune in 2001 (= attractivity of the commune)
EMPLOYDENS	(Working population residing in a commune + population working in the commune) / total surface of the commune
MOBILITY	Working population residing in commune *i* / Total resident population in *i*
ROADLENGTH	Total number of hm of roads in the commune (logarithm)
URBE	Level of urbanisation
DCORR	Fractal dimension of the built-up area obtained by correlation
RURALITE	Area devoted to agriculture in 2001 / total area (%)

Table 2: Explanatory variables at Level 2

4. Modelling results

Section 4.1 and Table 3 contain the results for accidents occurring on numbered roads (regional roads) with the exception of motorways, which are reported in Section 4.2 and Table 4. The three dependant variables are analysed separately. The exploratory data analysis conducted on the explanatory variables is not reported here; it was mainly based on correlation coefficients and odd ratios, and enabled the number of explanatory variables to be reduced and their best formulation to be selected.

4.1 Accidents on numbered roads

Proc Nlmixed uses a non-linear formulation of the model. It requires initial values for the parameters to be specified, so that the model converges. Convergence is due to the fact that maximum likelihood estimation is an iterative algorithm which may require many runs before reaching stable coefficient estimates. Our first estimate was β_0 = –2.05; this initial value was introduced into the "empty" model (Table 3, Column 2). The intra-municipal variance (Level 1) is an indicator of the variability between hectometres, whereas the inter-municipalities variance (Level 2) concerns the variability between municipalities. The total residual variance of the empty model was 7.77. The intra-class correlation $\rho_\mu=\frac{\sigma_{\mu 0}^2}{\sigma_{\mu 0}^2+\sigma_{\varepsilon 0}^2}$ indicates the proportion of the variance due to the communes (Level 2); here ρ equals 24.95% ($\sigma_{0\mu}^{2}$=1.94). For the hectometres (Level 1) $\rho_\varepsilon=\frac{\sigma_{\varepsilon 0}^2}{\sigma_{\mu 0}^2+\sigma_{\varepsilon 0}^2}$ = 75.05% ($\sigma_{0\varepsilon}^{2}$=5.83). The smaller the intra-class correlation, the better the standard errors of the parameters are estimated [11]. Let us also mention the pseudo-coefficient of determination (R^2), which indicates the weight of the explanatory variables at every level by comparing the residual variances to those of the empty model ($R_1^2 = 1-\frac{\sigma_{\varepsilon 0 actuel}^2}{\sigma_{\varepsilon 0 ancien}^2}$ and $R_2^2=1-\frac{\sigma_{\mu 0 actuel}^2}{\sigma_{\mu 0 ancien}^2}$). Pseudo *R*-squared (denoted R^2) should not be compared to the R^2s obtained by OLS. There are analogous but not equivalent.

The empty model contains no explanatory variables. It simply results in a partition of the total variation between the intra- and inter-level constituents. The purpose is then to reduce these variances by introducing explanatory variables. The explanatory variables that make a significant contribution to the equation related to *Y1* (being or not being part of a black zone) only explain 5.15% of the variance (Table 3, Column 3). However the decomposition by level of observation leads to a larger R^2 at the municipal level (R^2_2 = 23.19%). Let us here note that, quite surprisingly, the variance observed at the hm level (5.88) is almost the same (in fact slightly larger!) than that observed in the empty model (5.83). At this stage of the analysis, this is not easy to explain. Perhaps MLM models are not the best way to model *Y1*. Interpretation of the coefficients may be meaningless.

	Y1		*Y2*		*Y3*	
Variables	*Empty Model*	*MLM*	*Empty Model*	*MLM*	*Empty Model*	*MLM*
β_0	–2.05***	–13.71***	0.21***	–1.97***	0.47***	–1.47**
Level 1 variables						
AGGLO				0.38***		0.18***
TRAFFIC		1.52**		0.53***		
VMAX (0m) _dist				0.19***		
LANES(3)		–0.75***				
LANES (0m) _dist				0.34**		
LANES(100m)_dist				0.31***		
RUT						–0.31**

JUNCT (0m) _*dist*		1.33***		0.32**		2.15***
JUNCT(100m)_*dist*				0.35***		
ADHERENCE		0.75***		0.23***		0.35***
BZONE						0.75***
BUILT30		0.32***				0.54***
WOODS25						0.50*
FIRMS		0.47**				
DIRECTION (3)		–0.29**				
Level 2 variables						
EMPLOYDENS		2.64***				–0.97***
DCORR						2.82***
Variance $\sigma_{0\varepsilon}^2$ *(level 1)*	5.83	5.88	1.11	1.01	3.79	2.98
Variance $\sigma_{0\mu}^2$ *(level 2)*	1.94	1.49	0.0003	0.0001	0.0002	0.0002
Total variance	7.77	7.37	1.11	1.0088	3.79	2.98
ρ_ε	75.05%	79.78%	99.98%	99.99%	99.99%	99.99%
ρ_μ	24.95%	20.21%	0.02%	0.01%	0.01%	0.01%
R^2_1	–	–	–	9.05%	–	21.34%
R^2_2	–	23.19%	–	66.3%	–	–
R^2_{total}	–	5.15%	–	9.07%	–	21.34%

***significant at 99.9%; ** significant at 99%; * significant at 95%

Table 3: MLM for accidents on numbered roads

Y2 is the total number of accidents observed on each hectometre of road. Table 3 shows that its variability between communes is small ($\sigma_{0\mu}{}^2$= 0.0001; ρ_μ = 0.01%); communes can therefore be considered homogeneous in this respect. Compared to the empty model, the *X* variables have a greater explanatory power at the municipal level (R^2_2 = 66.30%) than at the hectometre level (R^2_1 = 9.07%). Several variables explain the variation of *Y2*, including the quality of the adherence of the road surface (ADHERENCE). We suspect the role of this variable in explaining *Y2* to be associated with users' behaviour; given the many recent technical improvements to motor vehicles, drivers may take greater risks because they feel more confident when driving vehicles equipped with these safety features even in difficult conditions [29]. When adherence is not good, road conditions seem not to be manageable by road users for one reason or another (weather conditions, infrastructural or behavioural circumstances), leading to an accident. *Y2* is also high when the hectometre is located in the vicinity of a crossroads (JUNCT_*dist*). This confirms earlier results [30, 31]. Better signalled intersections could be a short-term solution for avoiding black zones, with roundabouts being a solution in the longer term. Traffic density (TRAFFIC) also has a significant positive influence on *Y2*, confirming the results of authors such as Hiselius[32] or Golob[33]. The total number of accidents (*Y2*) is higher in urban agglomerations (AGGLO), which confirms the effect of density or mobility, and also close to places where driving conditions change (LANES 0, 100_*dist*; VMAX 0_*dist*). These latter are quite interesting because they identify road discontinuities that are potentially dangerous: narrowing of the road, agglomerations etc. Better warning signs could be an easy short-term solution for reducing risks at these transition places. However, in the longer term, they should be accompanied by changes in infrastructure and the environment.

Y3 is a simple measure of risk: the total number of accidents divided by the average daily traffic intensity. The scale effect is here due to hectometres only. Some 21.34% of the variance is explained. This means that infrastructure and the environment play a limited but significant role in accident risk. *Y3* is also explained by distance to crossroads (JUNCT_*dist*), location within urban agglomeration (AGGLO) or within a black zone (BZONE), bad adherence of the road surface (ADHERENCE), built-up (BUILT30) and wooded (WOODS25) environments. There is also an unexpected negative effect of rutting (RUT) (β=–0.31): the risk of accidents is small where rutting occurs. This might be explained by the fact that, in the area we studied, rutted roads may correspond to roads with dense traffic and hence congestion. We know that congestion leads to more damage-only accidents (fewer casualties). But rutted roads may also correspond to small roads between hamlets with little traffic, where vehicles do not necessarily adapt their speed. Variables measured at the municipality level have a very small but significant explanatory power for the risk of accidents. However DCORR, which is an index of urban morphology, plays an interesting role: the greater the uniformity in the built space (no hierarchy), the greater the risk of accidents. This is quite an important and novel finding in terms of planning. Specific spatial organisations (morphologies) may affect the relative speed of vehicles. Moreover, the risk of accidents also decreases when population density increases: the greater the density, the greater the activity and traffic and hence congestion or speed limits. These latter reduce the risk of accidents. This should be taken into account when considering mobility problems within urban areas [34].

We can conclude that the utility of MLM for modelling the occurrence of accidents is limited: effects of scale are small. In all cases one level of analysis is predominant, although this may differ for different dependent variables. In the model for the number of accidents (*Y2*), the significant variables explain 66.3% of the variance at the level of the commune, whereas in the model for the risk of accidents (*Y3*), the explanatory variables account for 21.34% of the variance at the level of the hectometre. In the first model, for the occurrence of black spots (*Y1*), the variation is mainly at the level of the hectometre, but the explanatory variables account for 23.19% of the variance at the municipal level. The effect of scale in this case is shared between hectometres (79.78%) and municipalities (20.21%). Although MLM is known to be helpful in revealing differences in variance among units of analysis at different levels, it is less interesting than expected in this frame of application (see Section 5).

The explanatory variables that are significant in all three models are proximity to crossroads (JUNCT_*dist*), the adherence of the road (ADHERENCE), and measures of traffic density (TRAFFIC or EMPLOYDENS). This confirms the results of previous studies. The quality of the road surface is a risk factor that deserves particular attention in the maintenance of roads, especially in a peri-urban context. At the municipal level the risk of accidents increases with the urban morphological indicator (DCORR) and decreases with population density.

4.2 Accidents on motorways

Given the characteristics of motorway traffic (separate lanes, few entries/exists, minimum/maximum speed conditions, etc.) and the specificities of road accidents on motorways, models were computed separately for this type of road accidents. The results are shown in Table 4.

The first analysis, for the existence of a black spot (*Y1*), shows an increase in the total residual variance for the MLM compared to the empty model (from 17.81 to 55.46) (Table 4, Columns 2 and 3); this means that, overall, the significant explanatory variables do not improve the explanation of whether or not a particular hectometre is part of a black zones of the motorway. The fact that an hectometre belongs to a black zone is not to be explained by the here used explanatory environmental variables.

	Y1		*Y2*		*Y3*	
Variables	*Empty Mode*	*MLM*	*Empty Model*	*MLM*	*Empty Model*	*MLM*
β_0	–5.27**	92.91**	0.75***	–11.22***	1.54***	4.85*
Level 1 variables						
TRAFFIC		–21.47**		2.55***		
VMAX (0m)_*dist*				2.38***		4.07***
VMAX (100m) _*dist*				1.83**		3.71**
LANES(4)		–2.44**		3.74***		
SURFACE (HFN)				–0.41***		–0.49*
SURFACE(200m)_*dist*				1.06**		
PROX_ACCES (300m)		1.29**				
BZONE						1.70***
BUILT (20%)				–2.63**		
FIRMS		2.46**		0.48*		
Level 2 variables						
EMPLOYDENS						7.19*
MIXITY						–5.31**
ATTRACT						–6.88*
Variance $\sigma_{0\varepsilon}^2$ *(niv 1)*	12.02	30.14	1.43	1.23	4.78	4.15
Variance $\sigma_{0\mu}^2$ *(niv 2)*	5.79	25.32	0.0002	0.0002	0.0002	0
Total variance	17.81	55.46	1.43	1.23	4.78	4.15
ρ_ε	67.47%	54.35%	99.98%	99.98%	99.96%	100%
ρ_μ	32.51%	45.65%	0.02%	0.02%	0.04%	–
R^2_1	–	–	–	14.24%	–	13.14%
R^2_2	–	–	–	–	–	–
R^2_{total}	–	–	–	14.24%	–	13.14%

***significant at 99.9%; ** significant at 99%; * significant at 95%

Table 4: MLM for accidents on motorways

The model related to the number of accidents by hectometre (*Y2*) is reported in Columns 4 and 5 of Table 4. The introduction of explanatory variables in the model now reduces the total variance (from 1.43 to 1.23); this is exclusively due to Level 1 variables (R^2_1=14.24%). Hence, the effect of scale is here mainly due to the hectometres. This is quite obvious for motorways where entries/exists are sparse compared to the scale of the communes. Let us have a look at the explanatory variables. Traffic density has a positive relationship to accidents: the larger the traffic volume, the larger the number of accidents. We know that his relationship is true on average: traffic varies with the time of the day leading to congestion at some periods. Let us remember that accidents on motorways often only involve one road user and are often associated with a loss of control and/or travelling in excess of the speed limit [35, 36, 37], which occur when traffic is not dense (often at night). Distance to a change in surface (SURFACE200m_*dist*), distance to a change in maximum speed (VMAX 0 and 100m_*dist*) are sources of changes in road behaviour; they should be the focus of planners' attention, as they normally correspond to roadworks on motorways or places close to cities. Better road signs could avoid these situations. The nearness of firms/department stores (FIRMS) is also a risk factor at the entrance of cities; this risk factor could be avoided by better land use/infrastructure planning. Surfacing of type HFN and the variable BUILT<20 have negative coefficients: asphalt with the conventional texture and sparsely built-up environments are factors of road safety rather than danger.

Modelling of the risk of accidents (*Y3*) is reported in Columns 6 and 7 of Table 4. It turns out that the effect of scale is once again only due to hectometres, and that 13.14% of the variance is explained by the explanatory variables introduced. The MLM adds very little in this case because the variance associated with one of the levels (here Level 2) is almost equal to 0. Significant explanatory variables are the distance to a maximum speed (VMAX 0 or 100_*dist*), membership of the hectometre of a black zone (BZONE) and the type of surfacing (HFN).

Hence, on motorways, Level 2 is useless in modelling *Y2* and *Y3* because variables describing the municipal level only apply to those sections of the motorways that are located near Brussels. It might have been better to choose another contextual environment because the municipal effect only represents a small part of the effects of scale on motorways. For *Y1*, the effects of scale are shared between hectometres (54.35%) and municipalities (45.65%), but no variable explains the total variation, which increases when explanatory variables are introduced. None of the explanatory variables recurs in each model, but each is useful in defining sensitive places on motorways (such as, for example, the zones with different speed limits, the changes in type of road, the surfacing of roads, etc.). To sum up, when modelling the presence of a hectometre in a black zone on motorways (*Y1*), variations are shared in a more or less equivalent way between hectometres (ρ_ε = 54.35%) and municipalities (ρ_μ = 45.65%). The introduction of explanatory variables does not improve the predictive power of the model. When modelling the number of accidents (*Y2*), the variation occurs at the level of the hectometres and 14.24% (R^2_{total}) is explained. When the dependent variable is the number of accidents divided by the average traffic intensity (*Y3*), the effect of scale is only due to hectometres but only 13.14% of the total variation is explained.

4.3 Comparing MLM to logistic regression results

Multilevel results are not really comparable to other regression results[12]. MLM relies on complex, particular distributions of relationships across and within levels. MLM outcomes are hence less general since each best-fitting model may be very specific to the dataset used. Let us here roughly compare *Y1* modelling results obtained on numbered roads with those obtained by Flahaut[3] by means of a more classical logistic regression (Table 5) using the same area of study and almost (but not exactly) the same explanatory variables. We see that (1) infrastructure and land-use have a smaller power of explanation in the MLM model (smaller pseudo R^2), and that (2) the explanatory effect of the variables is also slightly different. Similar effects are to be found for traffic density, crossroads, adherence, built-up areas, and proximity of firms/department stores. MLM modelling adds variables such as the orientation of the roads and the number of lanes; at the municipal level, MLM adds population density, which confirms other density effects measured at the hectometre level. In the logistic model, distances to changes in the speed limit or in the number of lanes as well as the type of road surface play a more determining role.

Variables	*Multilevel model*	*Logistic model*
TRAFFIC	++	+++
VMAX (0m)– _*dist*		+++
LANES	–––	
LANES (0m) _*dist*		+
SURFACE		++
JUNCT (0m) _*dist*	+++	+++

ADHERENCE	+++	+
BUILT (30%)	+++	+++
FIRMS	++	+++
DIRECTION	––	
EMPLOYDENS (level 2)	+++	

(+ a positive relationship; – a negative relationship.
+++/–––significant at 99.9%; ++/–– significant at 99%; +/– significant at 95%).

Table 5: Comparing multilevel and logistic modelling of *Y1* on numbered roads

Operationally, both modelling approaches lead to specific results, but, on average, road accidents data in Brabant Walloon seem to have a strong spatial structure that comes through in both modelling procedures. Results are quite stable and should be better integrated into land use planning and road infrastructure enhancement policies.

5. Conclusions

The importance of MLM in understanding contextual effects on road safety lies in its ability to meaningfully specify the latent structure of relationships, which involve individuals and their environments. This paper has modelled the spatial occurrence of road accidents by means of MLM. Three dependant variables were studied separately. Explanatory components were limited to infrastructure and environment. Due to the nature of the road accident and data limitations, only two levels of analysis were taken into account: the hectometre and the commune. Given these limitations, we can conclude that:

(1) MLM enables the relative importance of spatial levels in the explanatory process to be assessed. In our case, the commune has, on average, less importance in the explanation than the hectometre: road accidents occur at micro-locations (hm) which can be analysed in a broader spatial context, but this context does not seem to correspond to the commune. If communes are useful official statistical and administrative units, they vary in size and shape (modifiable areal unit problem) and are not suitable for explaining the locations and spatial concentrations of road accidents. This level is not appropriate for taking into account the complex relationships in which a car/ road user is involved in an accident. Unfortunately, in road accident analysis, the choice of the level of observation is not as straightforward as in human and behavioural sciences. In both road accident analysis and mobility schemes there is often a compromise between data availability and meaning. In this case hectometres and communes were the only possible choices.

(2) If the level of spatial explanation is not high, it is however significant and corroborates former results. Environment and infrastructure explain between 5% and 21% of the total observed variation in road accidents in Brabant Walloon. We are conscious that our models are mis-specified: we didn't take into account the many other factors that could interact (user behaviour, mobility patterns, etc.). Given these results, the physical characteristics of the road, as well as its environment, should be better integrated into safety and land-use policies.

(3) Three different dependent variables were analysed here: whether or not a hectometre belongs to a black zone; the number of accidents per hectometre; and the risk of accidents, defined as the number of accidents divided by the average traffic volume. Each *Y* variable has a specific meaning for police forces, emergency services or road engineers/planners. Hence, each model has a specific form, with a difference combination of independent variables.

(4) Most explanatory variables are associated with hectometres. Many are related to changes in road conditions. The importance of these changes in road conditions in the explanation reveals the inability of the road user to adapt his or her behaviour to changes in road conditions and road infrastructure [38]. In-depth analysis of each sub-type of accident should

increase our understanding of each type of circumstance, but this is far beyond the scope of this paper. All these associations show that spatial concentrations of road accidents often correspond to places where improvements could be made in terms of road design, signalling and land-use planning. This corroborates previous results on road accidents and road geometry [31, 39, 40, 41].

(5) The introduction of a morphological index (fractal dimension) is quite novel in measuring land use and more particularly in explaining road safety. Further analyses will be performed with this variable. In this paper we showed that homogeneity in texture leads to greater danger. Specific spatial organisation (morphology) may affect the relative speed of vehicles as well as the mobility patterns. This should be taken into consideration when considering safety and mobility problems within urbanised areas.

(6) Multilevel results are not really comparable to other regression results. MLM outcomes are less general since each best-fitting model may be very specific for the dataset used. In our case, logistic regression seems to be easier to use and could be extended to cope with autocorrelation [3]. Other analytical methods, such as a weighted geographic regression [17] (which is based on the hypothesis that the variations between variables measured in different places cannot be constant in the space), may however also be interesting to use. Instead of considering the local variations as averages and as unobservable, weighted regression allows us to measure local variations and to map them. However, the quality of the data collected for this type of analysis is preliminary.

The findings of this paper are both suggestive and limited in that they are based on only one data set, and only consider the environment and infrastructure as explanatory variables. Our modelling results depend strongly upon the many choices made, and these are strongly related to data availability. Neither the social characteristics of the road users nor the technical characteristics of vehicles are considered here. The interactions between social, technical and spatial variables are not taken into account. Our paper shows the importance of the hectometre as a basic spatial unit and the limited usefulness of multilevel models in analysing road accident locations. Other statistical techniques may be better suited to this task. Spatial concentrations of accidents are characterised by specific accident circumstances, which require different counter-measures to reduce their number (e.g. improvements in terms of road design, signalling, and local environment). There is no unique combination of characteristics associated with road accident locations: it is a complex phenomenon of which only a very few aspects have been considered here.

Acknowledgements. This work was funded by the Federal Scientific Policy in Belgium (project n° CP/10/342). We would like to thank Benoît Flahaut and the anonymous referees for helpful comments on earlier versions of this paper.

6. References

1. Thomas I. 'Spatial data aggregation: exploratory analysis of road accidents', *Accident Analysis and Prevention*, 28:2, 251–264, 1996.

2. Flahaut B., Mouchart M., San Martin E. and Thomas I. 'Identifying black zones with a local spatial autocorrelation index and a kernel method: a comparative approach'; *Accident Analysis and Prevention*, 35:6, 991–1004, 2003.

3. Flahaut B. 'Impact of infrastructure and local environment on road insecurity: logistic modelling with spatial autocorrelation'; *Accident Analysis and Prevention*, (in press).
4. Lassarre S. and Thomas I. 'Exploring road mortality ratios in Europe: national versus regional realities', *Journal of the Royal Statistical Society A* (in press)
5. Courgeau D. and Baccaïni B. 'Analyse multiniveau en sciences sociales', *Population,* 4, 831–864, 1997.
6. *Mathian H. and Piron M. 'Echelles géographiques et méthodes statistiques multidimensionnelles' in Sanders L. (ed.)* Modèles en analyse spatiale*, Paris, Hermès Science Publications, pp. 62 –103,.2000.*
7. Goldstein H. *Multilevel Statistical Models*, London, Arnold, 178 pp., 1995.
8. Jones K. *Concepts and Techniques in Modern Geography: Multi-level Models for Geographical Research,* Norwich, Environmental Publications, University of East Anglia, 50 pp., 1991.
9. Bryk A. S. and Raudenbsuch S. *Hierarchical Linear Models: Applications and Data Analysis Methods*, second edition, California, Sage Publications, 485 pp., 2002.
10. Jones K. and Duncan, C. 'People and places: the multilevel model as a general framework for the quantitative analysis of geographical data', in Longley and Batty, pp 79–104, 1996.
11. Jones A. and Jorgensen S. 'The use of multilevel models for the prediction of road accident outcomes', *Accident Analysis and Prevention*, 35, 59–69, 2003.
12. Gee G. and Takeuchi D. 'Traffic stress, vehicular burden and well-being: a multilevel analysis, *Social Science & Medicine*, (in press).
13. Kreft I. and de Leeuw J. Introducing Multilevel Modelling, *London: Sage, 1998.*
14. Snijders T. and Bosker R. *An Introduction to Basic and Advanced Multilevel Modelling*, London, Sage Publications, 266 pp., 1999.
15. Hox J. *Applied Multilevel Analysis*, Amsterdam, TT-Publikaties, 119 pp., 1995.
16. Jones K. and Duncan C. 'Individuals and their ecologies: analysing the geography of chronic illness within a multilevel modelling framework', *Health & Place*, 1:1, 27–40, 1995.
17. Fotheringhem A., Brunsdon C. and Charlton M. 'Multilevel modelling' in Fotheringham A. (ed) *Quantitative Geography: Perspectives on Spatial Data Analysis*, London, Sage Publications, pp. 103–106, 2000.
18. Polsky C. and Easterling W. 'Adaptation to climate variability and change in the US Great Plains: a multi-scale analysis of Ricardian climate sensitivities', *Agriculture, Ecosystems and Environment*, 85, 133–144, 2001.
19. Singer J. 'Fitting multilevel models using SAS PROC MIXED', *Multilevel Modelling Newsletter*, 10:2, 5-9, 1998.
20. Singer J. 'Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models, *Journal of Educational and Behavioral Statistics*, 24:4, 323–355, 1998.
21. De Keersmaecker M. L., Frankhauser P. and Thomas I. 'Analyse de la réalité fractale périurbaine: l'exemple de Bruxelles', *L'Espace Géographique,* (in press)

22. Thomas I., Frankhauser P. and De Keersmaecker M.L. 'Fractal dimension versus density of the built-up surfaces in the periphery of Brussels', paper presented at the ERSA 2004 Conference in Porto, refereed session.

23. Flahaut B. and Thomas I. 'Identifier les zones noires d'un réseau routier par l'autocorrélation spatiale locale. Analyses de sensibilité et aspects opérationnels', *Revue Internationale de Géomatique*, 12:2, pp. 245–261, 2002.

24. Steengergen T., Dufays T., Thomas I. and Flahaut B. 'Intra-urban location of road accident black zones: a Belgian example', *International Journal of Geographical Information Science*, 18, 2, 169–181, 2004.

25. Eckardt N., Flahaut B. and Thomas I. 'Spatio-temporalité des accidents de la route en périphérie urbaine. L'exemple de Bruxelles', *Recherche Transports Sécurité* , 82, 35–46, 2004.

26. Casaer F., Eckardt N., Steenberghen T., Thomas I., Wets G. and Wijnants J. 'Een onderzoek naar de kwaliteit van de Belgische ongevallendata', Working paper, LUC (Diepenbeek), 2002.

27. Institut Géographique National *Banques de données et cartes topographiques* (Top50r), Bruxelles, IGN, 2002.

28. Halleux J., Derwael F. and Merenne B. *Urbanisation. Monographie 11A*, Recensement Général de la Population et des Logements au 1er mars1998.

29. Edwards J. 'Weather-related road accidents in England and Wales: a spatial analysis', *Accident Analysis and Prevention,* 4:3, 201–212, 1996.

30. Larsen L. and Klines P. 'Multidisciplinary in-depth investigations of head-on and left-turn road collisions',. *Accident Analysis and Prevention*, 34, 367–380, 2002.

31. Geurts K., Thomas I. and Wets G. 'Understanding accidents at black zones using frequent item sets', *Accident Analysis and Prevention,* (submitted).

32. Hiselius L. 'Estimating the relationship between accident frequency and homogeneous and inhomogeneous traffic flows', *Accident Analysis and Prevention*, (in press).

33. Golob T., Recker W. and Alvarez V. 'Freeway safety as a function of traffic flow', *Accident Analysis and Prevention,* (in press).

34. Shefer D. and Rietveld P. 'Congestion and safety on motorways: towards an analytical model', *Urban Studies*, 34, 679–692, 1997.

35. Lee J. and Mannering F. 'Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis', *Accident Analysis and Prevention*, 34, 149–161, 2002.

36. Navon D. 'The paradox of driving speed: two adverse effects on motorway accident rate', *Accident Analysis and Prevention*, 35:3, 361–367, 2003.

37. Thiffault P. and Bergeron J. 'Monotony of road environment and driver fatigue: a simulator study', *Accident Analysis and Prevention*, 35:3, 381-391, 2003.

38. Elvik R. 'To what extent can theory account for the findings of road safety evaluation studies?', *Accident Analysis and Prevention,* (in press).

39. Agent K. and Deen R. 'Relationship between roadway geometrics and accidents', *Transportation Research Record,* 541, Washington D.C., 1975.

40. Wong Y. and Nicholson, A. 'Driver behaviour at horizontal curves: risk compensation and the margin of safety', *Accident Analysis and Prevention,* 24:4, 425–436, 1992.

41. Greibe, P. 'Accident prediction models for urban roads', *Accident Analysis and Prevention*, 35, pp. 273–285, 2003.