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

Part 1:
Sustainable production and consumption patterns

SUMMARY

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)

January 2005

Table of Contents

1 Theme and objectives of the research project

The project "Innovative spatial analysis techniques for traffic safety" consists of three interrelated subprojects, linked into a multidisciplinary thematic network. The objective of this research was to improve the explanatory model for traffic safety, in order to clarify the interactions between safety factors, and progress in an explanatory model for traffic safety.
In order to reach the objective, this project explored the potential of new data sources and analysis techniques. The following approaches were used by the research partners:

Three innovative approaches were explored:

- **Remote sensing analysis of very high resolution satellite images (KUL)**:
  The potential of very high resolution satellite imagery for identification of land use, infrastructure and traffic characteristics was explored.
- **Spatial statistics (KUL)**
  This part of the research included the analysis of statistically significant risk zones, and spatial regression of socio-economic characteristics of neighbourhoods, which might be related to accident concentrations
- **Model-based clustering of accidents (LUC)**
  The accident data consisted of a database of all the accidents with casualties occurred in Belgium in the period 1991-1999. Per accident, about 100 attributes are recorded. Clustering techniques were used to structure, and to find patterns and relations in this large dataset.
- **Association rules (LUC)**
  Association rules were used to identify the relevance of data, and to identify typical characteristics and circumstances of accidents in black zones.
- **Analysis of time-space variations in the spatial distribution of accidents (UCL)**
  The stability and the spatio-temporal variations of black zones in a peri-urban environment were examined.
- **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.

The development of a single explanatory model soon appeared to be very complex, and could not be realised. However, the methods did contribute to a statistically more sound analysis of accident concentrations.

2 Summary of the research results

2.1. Accident data

2.1.1 Actualisation of the accident database, and localisation of the accidents (KUL)

The first step consisted of actualisation of the accident database of all the accidents with casualties from 1991 to 1999, in Belgium. In Flanders and Wallonia only accidents on numbered roads were located on a digital map, while in Brussels accidents on local streets were also located. We used 'linear referencing' for accidents on numbered roads and 'address matching' techniques for accidents on local roads.

2.1.2 Data quality assessment (KUL,UCL,LUC)

The quality of the accident database was evaluated by looking at empty fields, inconsistencies and reliability. Empty fields can occur for two different reasons: the value for the field is unknown or forgotten (missing value) or the field is non-applicable. Unfortunately there is no distinction in the database between these two situations. Inconsistencies are another source of errors in the database and they show up when some information about the traffic accident is registered several times in different fields and these fields show contradictions. Apart from the intrinsic accuracy and consistency of the data, it is also important to have an eye for the reliability of the data. The reliability of different fields was discussed with experts. According to them, several fields are unreliable and can thus better not be used in analysis.

2.1.3 Data relevance assessment

Assessment of the relevance of the explanatory variables is also important. This was achieved with the data mining technique of association rules. This 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.

The study concentrated on high frequency accident locations in Brussels, for the period of 1991-1996. 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.
- Priority to the right and making a left turn are the most important factors in sideways collisions.

2.2. Remote sensing

Very high resolution imagery like Ikonos – we used an Ikonos image of Brussels with a spatial resolution of 1 x 1 meter – offers a lot of detailed information for mapping and interpretation or classification applications. However, automated extraction of infrastructure and traffic indicators (e.g. traffic volume) quickly gets complex and several constraints and technical problems show up making it practically impossible to extract such information in an automatic way. Thus, it seemed more feasible to apply a multispectral classification technique on the image and produce a land use map. With very high resolution (VHR) imagery this is not anymore a standardized procedure as VHR imagery necessitates specific attention because the imagery contains a huge amount of detailed information which was not present (visible) on the traditional images with a lower resolution. Accordingly, a multispectral per-pixel classification will result in a fairly overloaded and diffuse land cover map (Figure 1). This map is difficult to interpret and has a low readability. Therefore, the land cover map will have to be classified further, based on land parcels instead of pixels because parcels are meaningful objects – at least in the context of land use classification. Another difficulty concerns the identification of *land use*. Whereas *land cover* is related to the physical characteristics of the earth's surface and can rather easily be classified, land use is related to the socio-economic occupation of the earth's surface and its classification is more problematic. We were able to profile and classify different land uses by means of a statistical analysis of the land cover objects within the land parcels. Five land use classes were classified automatically (high density built area or central city land use, medium density residential area, low density residential area, offices and industrial area and 'green' area) and a few extra classes were added manually on the map. Figure 2 shows an extract of the final land use map.

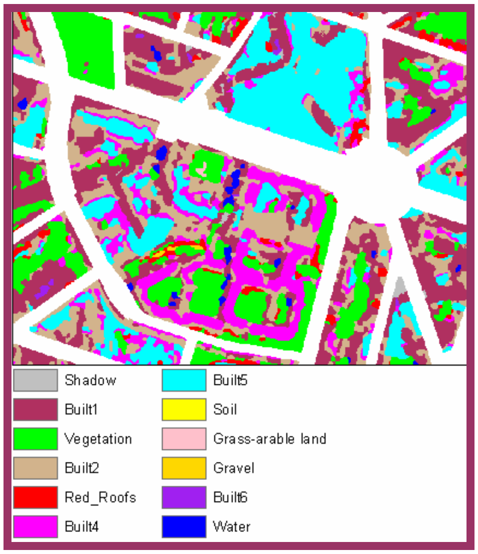


Figure 1: land cover classification (detail of European district)

Figure 2: Land use map, per-parcel classification, European District

This technique – parcel based land use classification – is capable of classifying the high amount of detailed information of very high resolution imagery like Ikonos. However, the overall accuracy is still quite low (75%), with the typical urban land use classes being the most difficult to classify.

2.3. Spatial statistics

Explorative: Two-dimensional significant black zones

In the next phase, the researched turned to exploration and modelling of the spatial accident pattern in Brussels. First, an explorative analysis was done to delineate significant two-dimensional black zones. 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 like the city of Brussels, it is more interesting to search for two-dimensional black zones. In order to be sure that spatial clusters exist in the pattern, Ripley's K-function was plotted and this plot confirmed the hypothesis that clusters are present. Next a spatial clustering technique (developed in the epidemiologic research domain) was applied to the spatial pattern of road accidents. Essentially, the technique calculates Poisson probabilities, expressing the likelihood for the occurrence of the observed number of accidents. On the resulting probability surface, the black zones are then delimited for a specified significance level (e.g. 0.05). The results are maps which show zones with a significantly high accident frequency (Figure 3) or accident risk (Figure 4). Whereas accident frequency represents the absolute accident count, accident risk takes the accident exposure into account and thus measures the risk.

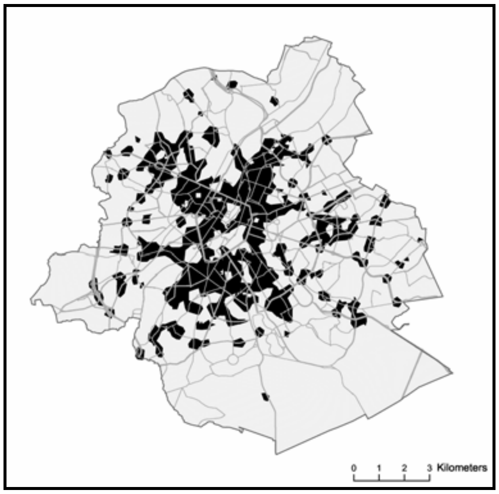


Figure 3: absolute black zones, all accidents (1997-99), bandwidth 250 m

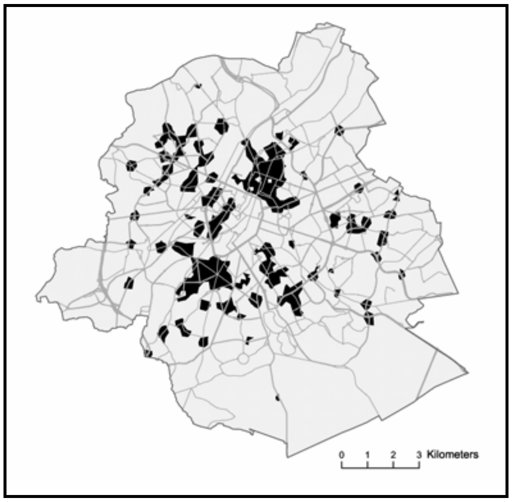


Figure 4:risk black zones, all accidents (1997-99), bandwidth 250 m

Explanatory: spatial regression at neighbourhood level

Finally, an explanatory model for accident frequency was designed in order to quantify the effect of the neighbourhood (the statistical district) on accident occurrence in these neighbourhoods. The model thus looks only at the broad spatial context and not at traffic or infrastructure related explanatory variables. Specifically, our explanatory variables described demographic, socio-economic, housing comfort and building type information and also an exposure factor (traffic volume estimation). We were well aware of the spatial autocorrelation effects, which were present in the data, and thus used a model which was corrected for this effect. The results indicate that accident exposure is very important in explaining accident frequency but the demographic and socio-economic factors can't improve the predictive power of the model. Thus, for Brussels we can conclude that the neighbourhood characteristics, as we described them, have no effect on the accident frequency.

2.4. Model based clustering on traffic accident data

One of the main goals of the LUC research was to develop a method for clustering of traffic accidents. Due to the complex interplay between accident causes, we decided to deploy an unsupervised traffic accident examination based on a broad clustering of all available attributes, without any assumption or hypothesis on either the existence of some typical in advanced known group or relationships between contributing factors. A model-based or latent class clustering analysis was preferred for this research for several reasons. The main reason is that most traditional clustering techniques are based on a distance measure or similarity index but these methods are not designed to compare categorical data such as our accident data. For this study the 1997-1999 accident records of two different regions were used: Walloon Brabant and Brussels and for both regions, a black zone attribute was available. In Walloon Brabant the model performed optimal with 5 clusters and in Brussels 7 clusters were needed for the model to reach its optimum.

In Walloon Brabant, black zone accidents are mostly represented in two clusters, thus a future focus on black zone accidents will have to consider these clusters: 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. 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. Concerning the age factor, it is highly important to consider the collision type: 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.

In Brussels, many conclusions differ from the ones for the Walloon Brabant region which is not remarkable because it is a very different region, geographically as well as demographically. The fact that 7 clusters were optimal is an indication that the accidents are far more heterogeneous. 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 Brussels. Additionally, it seems that this type of traffic accident often occurs at a crossroad with traffic lights where the pedestrian ignores a red light.

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. In order to expose the heterogeneity, the latent class clustering technique proved itself very useful.

2.5. Understanding accidents in black zones using association rules

During the last part of the research, we tried to develop and apply a data mining technique in order to profile black spots. 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. Again, accident data from 1997-1999 for Walloon Brabant was used.

The most important result of this research is that road accidents in black zones correspond to specific association rules:

- taking a left turn on intersections with traffic lights
- 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.
- 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.

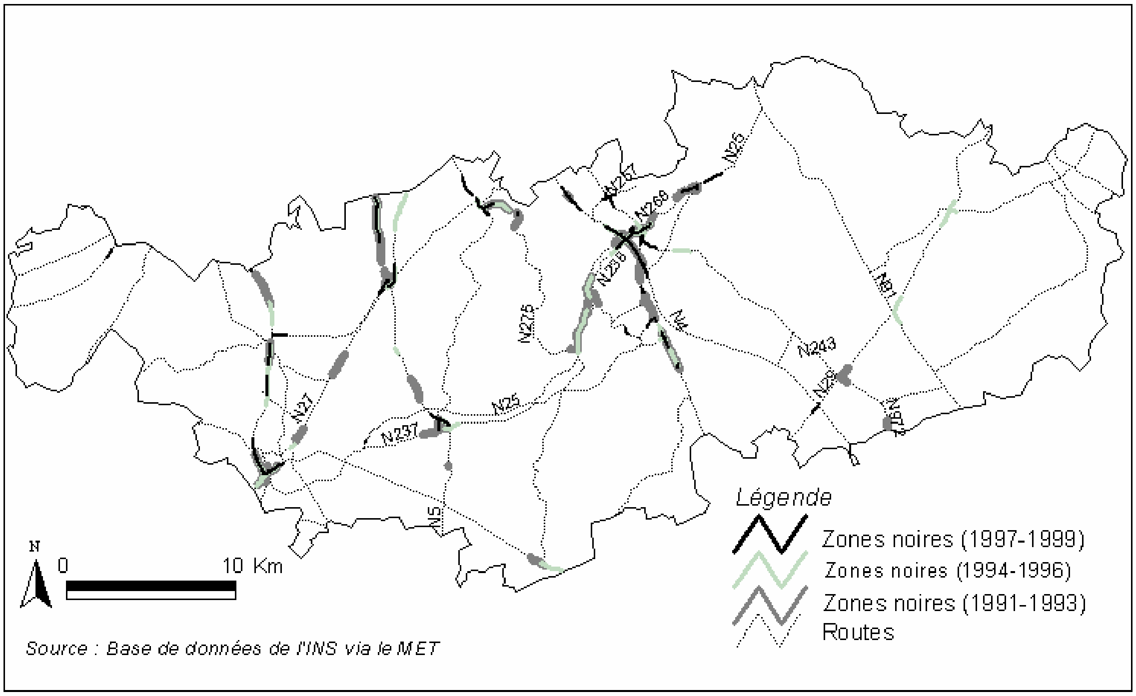
This confirms that pedestrian injury collisions often occur when and where large numbers of pedestrians travel within complex roadway systems with high traffic flows.

- Finally, loss of control over the steering wheel and the resulting collision with a crash barrier is a frequently occurring accident pattern and is related to an inadequacy of the speed and/or behaviour of the user to the driving circumstances.

2.6. Spatio-temporal distribution of road traffic accidents in a peri-urban environment. The case of Brussels

The aim of this preliminary research is to analyze the spatio-temporal variations in the spatial concentrations of traffic accidents in a peri-urban 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 peri-urban. 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.



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

First, it is important to know whether black zones are fairly stable in space and time. If so, then 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. We compared three consecutive 3-year periods: 1991 to 1993 and 1994 to 1996 on the one hand and 1994 to 1996 and 1997 to 1999 on the other.

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 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 (indices calculated from contingency tables). 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).

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). 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. We can therefore reject the hypothesis that *the geographical distribution of black zones varies as a function of the period time selected.*

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

The importance of Multi Level Modelling (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. The explanatory factors 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.

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.

5) 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. Other analytical methods, such as a weighted geographic regression (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 analysis 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;

data availability strongly constrains the model building. The results show 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.

3 General conclusion

In this project, we tried to reach better understanding of the spatial components of traffic accidents, by using different approaches: remote sensing, spatial statistics, data-mining and multi-level analysis.

Although the accident database is large, the potential for analysis is reduced due to the moderate data quality. The lack of information about the actual accident circumstances (such as the speed of the vehicles involved), is another disadvantage.

Ikonos satellite images were used to classify urban land use in urban environment, by performing parcel-based classification. This classification makes it possible to link morphological characteristics and statistics at parcel level.

Statistically significant 'neighbourhoods' (two-dimensional accident concentration zones) can be determined, both in absolute terms and in terms of accident risk. An explanatory model of these concentrations was developed by means of spatial regression, using statistical data of the NIS sector. The model results confirm the importance of traffic volume as explanatory variable for unsafety. Introducing neighbourhood factors did not increase the explanatory value of the model.

Model-based cluster analysis supports the fact that traffic accidents are a very heterogeneous phenomenon. Prior to any explanatory research of accidents, clustering should be performed. To reach a thorough understanding of the complexity of the accident data, the latent class clustering technique appeared very powerful.

Association rule techniques were used to identify the relevance of accident data, and to identify accident factors and –circumstances typical for black zones.

The stability and the spatio-temporal variations of black zones was examined for a 9 year period (1991-1999) in a peri-urban environment. The results confirm the image of the map. Major roads are an influential spatial structure for the last two periods (1994-1996 and 1997-1999). Most (40%) black zones on the network are stable in time. There is a hard core of black zones where neither adaptations to infrastructure, nor optimisation of signalisation, could increase safety. The hypothesis that black zones vary through time, was rejected.

Explanatory factors for traffic (un)safety appear to be very scale-dependent. Instead of constructing different models for different scale levels, interactions between at different scales, and between these scales, were studied and modelled. This study confirms the importance of the hectometre (road section) as spatial base unit and the limited usability of multi-level models for traffic safety. Other statistical techniques may be preferable to MLM. The results of this analysis require careful interpretation because they are based on a single dataset and only environment and infrastructure are used a explanatory variable.