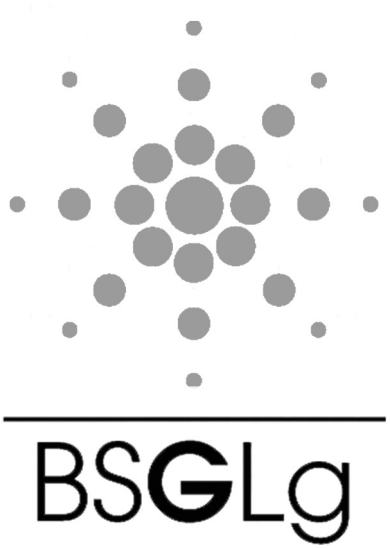
BULLETIN DE LA SOCIÉTÉ GÉOGRAPHIQUE DE LIÈGE, 2013, 60

Publié avec l'aide financière du
FONDS NATIONAL DE LA RECHERCHE SCIENTIFIQUE - FNRS

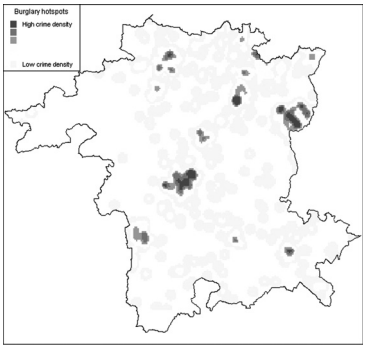


Crime mapping & modelling

Marie TROTTA, Jean-Paul KASPRZYK & Jean-Paul DONNAY
Éditeurs scientifiques

Édition de la
SOCIÉTÉ GÉOGRAPHIQUE DE LIÈGE
Sart Tilman - B11 - B 4000 LIÈGE (Belgique)
2013

Composition du Comité de lecture



Document de couverture :

Densité de cambriolages dans le comté de Worcesteshire (UK)
Auteur : Spencer Chainey

Mise en page : Imprimerie BJ

ISBN 2-87298-019-9
ISSN 0770-7576

BSGLg, 60, 2013, 3-4

EDITORIAL
CRIME MAPPING & MODELLING

For the first time the SAGEO (Spatial Analysis and Geomatics) conference – supported by the GDR MAGIS (CNRS) – was held out of France in 2012. On this occasion, the hosting University of Liege (Belgium) organized a full week of events dedicated to all geomatics professionals between the 5th and 9th November 2012. While SAGEO conference took place during the last three days of the week, three parallel workshops were first held on Tuesday 6, respectively on "urban GIS", "geo -marketing" and "crime mapping & modelling".

Several hundred participants from various public gathered throughout the week: faculty members, graduate and PhD students, professionals from the public and private sectors. It was obviously a good opportunity to make participants aware of innovative fields of research that are still underdeveloped outside the Anglo-sphere. This is precisely the case of crime mapping to which was dedicated a complete one-day workshop.

Crime mapping is not strictly speaking a recent discipline (cf. Guerry's and Quetelet's pioneering works in the nineteenth century). However, it was, for a long time, restricted to location of crime events on a map thanks to pins (pin mapping). As the development of any discipline needs to be rooted in a robust theoretical framework, the decade of the eighties saw the emergence of environmental criminology: "*the scientific study of spatial patterns in crime, the perception and awareness of space potential criminals, criminal mobility patterns, and the process of target selection and decision to commit the crime*" (Brantingham & Brantingham, 1981, p.7). The environmental criminology, part of the positive school of criminology, generated new questions and challenges about the modelling of relationships between offences and the places where they take place.

For example: are there general laws to explain offender's spatial choices? What is the relationship between the offender's activity space and the offence locations? How can we model such relationships? Are there specific built-up environments that favour the concentration of criminal activities? Digital mapping and geographic information systems allowed answering to such questions thanks to the access to large databases describing the environment features and statistics, interactive geo-visualization of several sources of information and genuine algorithms in spatial analysis.

The workshop in Liège successfully brought together researchers from the most specialized European centres in the domain: the Jill Dando Institute (University College London) , the Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), the Institute Forensics of Lausanne (University of Lausanne), the Criminal Intelligence Service of Austria, several representatives of the Strategic Analysis and Operational Analysis Services of the Belgian Federal Police, and of course representatives of several departments concerned in the University of Liege.

The workshop showed that researches from those centres are strongly diversified with applications such as detection of hot spots, geographical profiling or crime prevention through environmental design. In order to share good practices, a round table was also organized during the workshop with two recurrent issues: the integration of the temporal dimension in crime mapping and the collaborations between the operational and strategic aspects of the discipline.

Five authors agreed to draft a paper about their contribution. They are gathered in this special issue of the Bulletin of the Geographical Society of Liege.

Two authors focus on the detection of hot spots in order to identify places concentrating the offences. Spencer Chainey is the principal research associate at the University College London, department of Security and Crime Science. He examines the influence that resolution and bandwidth size have on hotspot maps built with a Kernel density

estimation. David Dabin, Christiane Dickens and Paul Wouters are strategic analysts for the federal police of Belgium. They propose a methodology to evaluate hotspots of car accidents restricted to the road network.

In addition to the spatial distribution of crimes, the issue of crime temporality arises. Do the patterns change according to the chosen temporal granularity and resolution? As criminal activities involve mobile offenders and/or mobile victims, the distribution of criminal opportunities presents varying attractiveness across both time and space (Ratcliffe, 2010). The temporal dimension can be integrated through many granularities: hour of the day, week/weekend, seasons and processes: evolution, diffusion, periodicity, etc…

The contribution of Quentin Rossy, lecturer and senior researcher at the school of forensic sciences of the University of Lausanne, precisely proposes a visualization methodology for the spatiotemporal analysis of crime data. He illustrates his method with two analytical tasks frequently applied: the analysis of traces left by digital devices like mobile phone or GPS devices and the detection of crime series.

The spatiotemporal pattern is also used in geographical profiling, where the series of crime locations – potentially attributed to a same unknown offender – are operated to delineate a prior search area. Marie Trotta is research fellow for the National Fund for Scientific Research. Together with her colleagues, André Lemaître and Jean-Paul Donnay respectively from the criminology and geomatics departments of the University of Liège, she brings a theoretical reflexion about the constraints and factors enabling the computation of an effective geographic profile, especially in the spatiotemporal aspects of the crime series.

In the previous contributions, crime mapping is mainly used to identify and act on risky areas. But it is also helpful to study evolutions, by comparing for example crime patterns before and after the implementation of crime reduction policies. Christian Kreis is a post-doc researcher at the Netherlands Institute for the Study of Crime and Law Enforcement (NSCR). In his paper, he uses geospatial data mining to enhance the validity of an observational design in order to evaluate community policing in major Swiss urban areas.

These contributions reflect the multifaceted aspect of crime mapping but also the degree of scientific and technical sophistication achieved by the discipline to answer a fundamentally geographical question: "*why do crimes happen there, and not elsewhere*?"

BRANTINGHAM, P. & BRANTINGHAM, P. 1981. *Environmental Criminology.* Prospect Heights (IL): Waveland Press.

RATCLIFFE, J. 2010. Crime Mapping: Spatial and Temporal Challenges. *In:* PIQUERO, A. R. & WEISBURD, D. (eds.) *Handbook of Quantitative Criminology.* New York : Springer.

Editors

Marie TROTTA,
Jean-Paul KASPRZYK
& Jean-Paul DONNAY
Geomatics Unit of the
University of Liège

TABLE DES MATIÈRES

BSGLg, 60, 2013, 7-19

EXAMINING THE INFLUENCE OF CELL SIZE AND BANDWIDTH SIZE ON KERNEL DENSITY ESTIMATION CRIME HOTSPOT MAPS FOR PREDICTING SPATIAL PATTERNS OF CRIME

Spencer CHAINEY

Abstract
Hotspot mapping is a popular technique used for helping to target police patrols and other crime reduction initiatives. There are a number of spatial analysis techniques that can be used for identifying hotspots, but the most popular in recent years is kernel density estimation (KDE). KDE is popular because of the visually appealing way it represents the spatial distribution of crime, and because it is considered to be the most accurate of the commonly used hotspot mapping techniques. To produce KDE outputs, the researcher is required to enter values for two main parameters: the cell size and bandwidth size. To date little research has been conducted on the influence these parameters have on KDE hotspot mapping output, and none has been conducted on the influence these parameter settings have on a hotspot map's central purpose – to identify where crime may occur in the future. We fill this gap with this research by conducting a number of experiments using different cell size and bandwidth values with crime data on residential burglary and violent assaults. We show that cell size has little influence on KDE crime hotspot maps for predicting spatial patterns of crime, but bandwidth size does have an influence. We conclude by discussing how the findings from this research can help inform police practitioners and researchers make better use of KDE for targeting policing and crime prevention initiatives.

Keywords:
hotspot analysis, kernel density estimation, crime prediction, cell size, bandwidth, burglary, violent crime

Résumé
La cartographie des points chauds (hotspots) est une technique populaire pour orienter les patrouilles de police et assister d'autres initiatives visant à la réduction de la criminalité. Il existe un certain nombre de techniques d'analyse spatiale qui peuvent être utilisées pour identifier les points chauds, mais la plus populaire au cours des dernières années est l'estimation à noyau de densité (Kernel Density Estimation – KDE). KDE est très populaire en raison de la manière visuellement attrayante dont elle représente la distribution spatiale de la criminalité, et parce que la méthode est considérée comme la plus précise parmi les techniques de cartographie des points chauds couramment utilisées. Pour produire des résultats avec KDE, le chercheur est tenu de fixer les valeurs de deux paramètres principaux : la taille des cellules et la taille de la fenêtre de convolution. A ce jour, peu de recherches ont été menées sur l'influence qu'ont ces paramètres sur l'interprétation finale d'une carte des points chauds – à savoir, identifier où la criminalité peut se produire dans l'avenir. Nous comblons cette lacune avec cette recherche, en effectuant un certain nombre d'expériences en utilisant différentes tailles de cellules et de valeurs de fenêtre, avec des données de la criminalité sur les cambriolages résidentiels et les agressions violentes. Nous montrons que la taille des cellules a peu d'influence sur les cartes de points chauds de criminalité issues de KDE pour prédire la répartition spatiale de la criminalité, mais par contre la taille de la fenêtre a une influence. Nous concluons en discutant de la manière dont les résultats de cette recherche peuvent aider à informer les praticiens de la police et assister les chercheurs dans une meilleure utilisation de KDE permettant de mieux cibler les initiatives de prévention du crime et de maintien de l'ordre.

Mots-clés
Analyse des points chauds, estimateur à noyau de densité, prédiction de criminalité, taille de cellule, taille de fenêtre, cambriolage, agression violente

I. INTRODUCTION

The mapping of hotspots of crime has become common practice in police agencies across the world. A hotspot is defined as being an area of high concentration of crime relative to the distribution of crime across the entire study area (Home Office, 2005; Chainey and Ratcliffe, 2005; Sherman, 2009). In these terms, hotspots can exist at different geographic scales of interest, whether it is at the city level for exploring localities where crime is highest, or at a local residential housing estate level, identifying particular streets or clusters of buildings where crime is seen to highly concentrate. Hotspot analysis has been applied to many forms of crime: from the analysis of gang-related murders in Belo Horizonte, Brazil (Beato, 2008), violent crime in Philadelphia (Ratcliffe et al., 2011), residential burglary, street robbery and vehicle crime in London (Eck et al., 2005), and street assaults in Melbourne, Australia (Mashford, 2008).

The use of hotspot mapping has also helped initiate the concept of hotspot policing – the targeting of police patrol strategies to crime hotspots in an effort to reduce the high volume of crime that is committed at these locations (see Braga, 2007 and Ratcliffe et al. 2011 for examples). Hotspot maps are also routine outputs that feed into *Compstat* style meetings (for a description of *Compstat* and examples see Chainey and Ratcliffe, 2005 and Home Office, 2005) and the intelligence production process of the UK's National Intelligence Model (NPIA, 2010). Hotspot mapping has therefore become a ubiquitous application in contemporary policing.

There are many spatial analysis techniques that can be applied to produce hotspot maps of crime. These include the use of spatial ellipses as shown by Block and Block (2000) when analysing hotspots around rapid transit stations in Chicago, applying a thematic (or choropleth) mapping approach to geographic administrative units as illustrated by Ratcliffe and McCullagh (2001) in their analysis of burglary across a study area's census zones, and grid thematic mapping as used by LeBeau (2001) to map patterns of emergency calls and violent offences in North Carolina. However, it is the use of kernel density estimation that in recent times has become the technique of choice by police practitioners and researchers (Chainey et al., 2008a), and as illustrated by examples of hotspot maps presented at the 2012 International Crime and Intelligence Analysis Conference (Figure 1).

Kernel density estimation (KDE) is also considered to be the most accurate of these *common* hotspot mapping techniques. This was illustrated by Chainey et al. (2008a) in a study that compared the hotspot mapping outputs generated using spatial ellipses, thematic mapping of census areas, grid thematic mapping and KDE for their ability to predict spatial patterns of crime. That is, based on the principle that hotspot mapping is used as a basic form of crime prediction – it uses data on past incidents to determine where crime may occur in the future - they showed that KDE outputs consistently produced better prediction results in comparison to the other techniques.

Like many spatial analysis techniques KDE requires the researcher to determine the values to enter for certain technical parameters in order to produce mapping

(a) (b)

Figure 1. Examples of kernel density hotspot maps of crime, as presented at the 2012 International Crime and Intelligence Analysis Conference (a) hotspot map of burglary for the county of Worcestershire (UK) and (b) hotspot map of criminal damage in Hawkesley, Birmingham (UK).

output. These two main parameters are the value for the cell size (sometimes referred to as the resolution) and the value for the bandwidth (often also referred to as the search radius). An alternative method to specifying a fixed bandwidth is the adaptive KDE approach where the bandwidth varies based on a user determined number of neighbours to include in the kernel density calculation. This adaptive kernel approach is rarely used by crime mapping practitioners (Chainey et al., 2008), hence the focus of this research was towards the more commonly used fixed kernel bandwidth approach. There is currently very little guidance on cell size and bandwidth size selection for the practical application of KDE hotspot mapping in policing, with the researcher either giving little thought to these values and their influence, settling for the default values determined by their KDE software application, or drawing from their own particular whims, fancies or experience (Eck et al., 2005; Chainey and Ratcliffe, 2005).

In this paper we aim to better inform practitioners and researchers by examining the influence that cell size and bandwidth size have on KDE hotspot mapping outputs. We follow as a guide the methodology used by Chainey et al. (2008a) that compared the spatial prediction measures of different hotspot mapping techniques, by comparing the influence that different cell size and bandwidth size values have on the spatial prediction abilities of KDE hotspot mapping outputs.

Section 2 describes in further detail the kernel density estimation function and how cell size and bandwidth values fit mathematically into its formulation. Section 3 describes the methodology, with results (section 4), a discussion and conclusions then following.

II. KERNEL DENSITY ESTIMATION

The spatial application of kernel density estimation emerged as a popular technique in spatial epidemiology to assist the study of disease patterns (for an early example see Bithell, 1990). Similar to disease, crime incidents are most usually geographically referenced as points. The kernel density estimation function is applied to these points to obtain a smooth surface estimate representing the density of the point distribution. In mathematical terms, KDE is expressed as:

$$f(x,y) = \frac{1}{nh^2} \sum_{i=1}^{n} k\left(\frac{d_i}{h}\right) \qquad (1)$$

Where f(x,y) is the density value at location *(x,y)*, *n* is the number of incidents/points, *h* is the bandwidth, d_i is the geographical distance between incident *i* and location *(x, y)* and *k* is a density function, known as the kernel. *k* can take many forms although the results between different functions produce very similar density values (Bailey and Gatrell, 1995). A common choice for *k* is the quartic function (Bailey and Gatrell, 1995; Ratcliffe, 2002; Levine, 2004).

Evaluation of the components of the KDE equation show that the density value for each location is affected by the number of points, their spatial distribution, and the bandwidth. For the purpose of generating a hotspot map of crime for a single study area, using data for a particular retrospective snapshot of previous incidents, the number of crime incidents across the area would remain the same (and hence not influence changes in the density estimate), the spatial distribution of the crime incidents is static, therefore it is the bandwidth that will influence different values of *f* at each *x,y* location. Each *x,y* location is represented spatially as a grid cell (the coordinates referring to the centroid of that cell), with the calculated density value *f* attributed to each cell. The cell size chosen by the researcher can vary, resulting in many calculations of *f* if the cell size is small or much fewer calculations if the cell size is large. Whilst cell size is not an input to the KDE equation, the representation of these density values for areas of different size will be subject to the Modifiable Areal Unit Problem (Openshaw, 1984) – different size cells may produce different results of the spatial KDE distribution of crime.

There is currently very little guidance on the choice of cell size a researcher should select and no research that we are aware of that investigates comprehensively the impact it can have on a crime hotspots central aim – to accurately identify areas where there have been high concentrations of crime, using the hotspot mapping output to determine where policing resources should then be targeted. The little guidance that is offered is by Chainey and Ratcliffe (2005) who recommend that a suitable KDE cell size to choose for crime hotspot mapping is to divide the shorter side of the study area's minimum bounding rectangle (MBR) by 150. Whilst simple to calculate and used to determine the default cell size in the Hotspot Detective MapInfo add-on (Ratcliffe, 2002), this approach has not been rigorously evaluated.

The choice of bandwidth size value for crime researchers to select is similarly uninformed. Whilst there are several bandwidth size optimisation routines such as the Mean Integrated Square Error (Fotheringham et al., 2000; Bowman and Azzelini, 1997), Akaike Correlation Coefficient and the Cross Validation method (Silverman, 1986; Brunsdon, 1995; Fotheringham et al., 2002), these tend to produce large bandwidths and are considered unsuitable for the purposes of exploring local spatial patterns of the density distribution of crime (Uhlig, 2005). Bailey and Gatrell suggest a value derived from calculating $h = 0.68n^{-0.2}$ as

a *rough choice* (Bailey and Gatrell, 1995, 86) for the bandwidth (where *n* is the number of observed events across the study area), but again experimentation of this approach tends to produce bandwidth values that are much larger than those used by crime researchers in practice (Uhlig, 2005). Chainey (2011) recommends a good starting bandwidth is to measure the shorter side of the study area's MBR, divide by 150, and multiply this value by 5. Whilst simple to calculate, the choice of this bandwidth size has not been evaluated, but is common applied - Hotspot Detective for MapInfo uses a very similar procedure for calculating bandwidth default values (Ratcliffe, 2002). Many others suggest an approach of experimenting with different sizes of bandwidth (Bailey and Gatrell, 1995; Eck et al., 2005; Chainey and Ratcliffe, 2005). Whilst this encourages the researcher to explore their data under different bandwidth conditions it often leaves the researcher choosing the mapping output that *looks the best* (Chainey and Ratcliffe, 2005, 159), rather than being more scientifically informed on the influence that bandwidth size selection may have on the hotspot map's central purpose – to accurately assist the targeting of police interventions by helping determine where crime is likely to occur in the future.

III. METHODOLOGY

Kernel density hotspot maps were created using MapInfo Professional version 10.5 and the MapInfo add-on programme Hotspot Detective (Ratcliffe, 2002). The study area chosen was the district of Newcastle-upon-Tyne in North East England (Figure 2). Newcastle is one of England's largest ten cities and therefore includes many of the urban geographical features and amenities that one would expect in a typical city. This includes a vibrant shopping and entertainment area in the centre of the city, a large number of economic and commerce functions, a mainline train station, a metro system, and two large universities. The district also includes rural areas towards the north. The district population was 292,000 at the time of the 2011 Census of England and Wales.

Geocoded crime point data was provided by Northum-

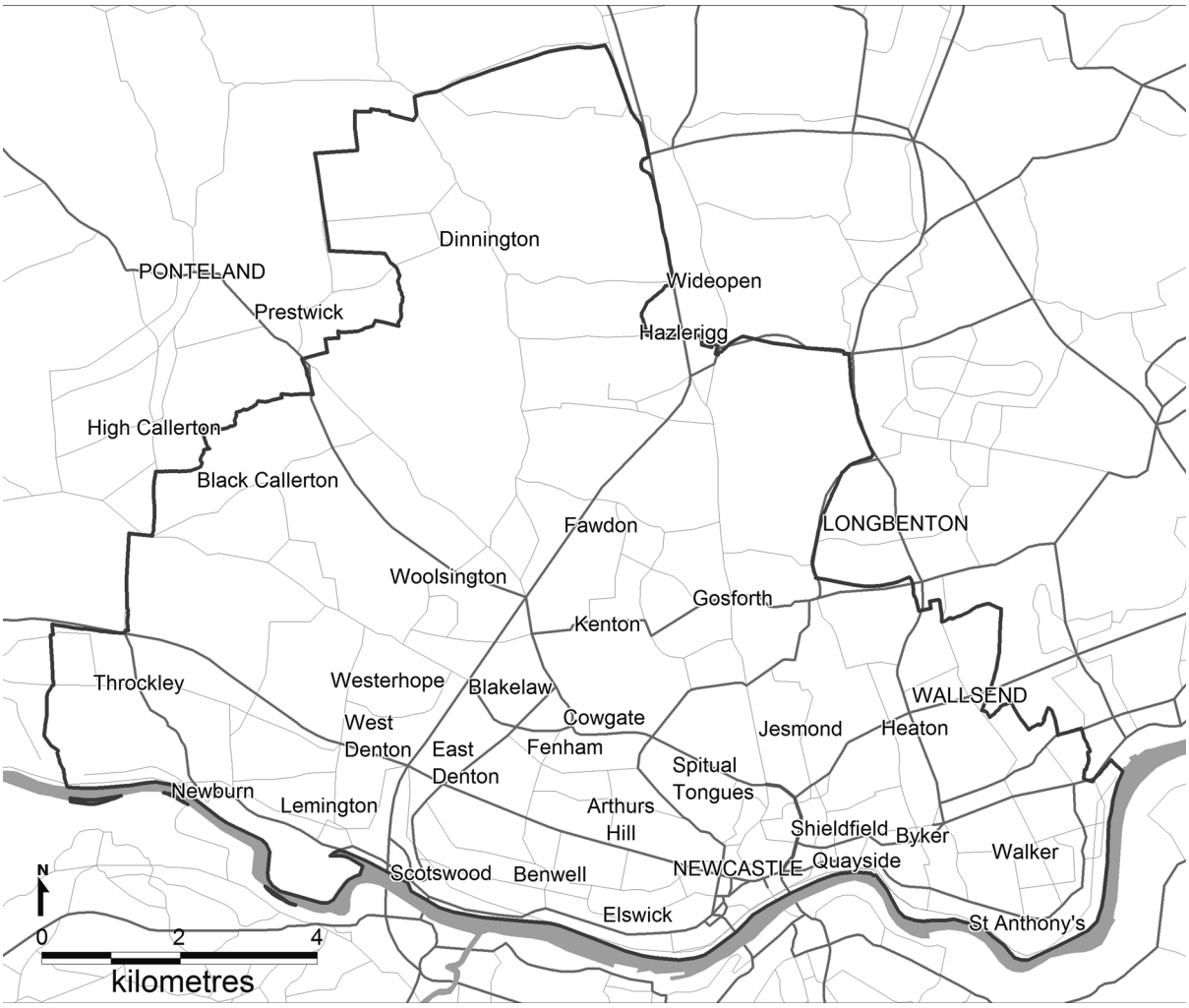


Figure 2. Newcastle-upon-Tyne study area

bria Police for a one year period (1st October 2009 to 30th September 2010). Burglary to a residential dwelling and violence with injury were the two subsets of data that were chosen for analysis. Two types of crime were selected to explore consistencies between the results. Previous research into the spatial prediction of hotspot maps has used residential burglary (Chainey et al., 2008a), but to date no study has used violent crime data. The analysis of residential burglary would therefore enable comparisons with previous research, with the analysis of violent crime offering a new perspective of spatial prediction patterns using KDE hotspot maps. These crime types were also chosen because they are groupings that are regularly analysed by police and crime reduction practitioners - therefore the implications of the research would be of practical interest. Table 1 lists the number of crimes in each sample crime type dataset. The two sets of geocoded crime data were validated using a methodology for geocoding accuracy analysis as reported in Chainey and Ratcliffe (2005, 61-63). This revealed the crime data to be more than 95% accurate to the street address level and fit for purpose for this research.

Crime type	**Number of incidents** (1st October 2009 to 30th September 2010)
Residential burglary	1304
Assaults with injury	1838

Table 1. The number of crime incidents for each crime type in Newcastle-upon-Tyne

In following the methodology used by Chainey et al. (2008a), a suitable date had to be chosen within the data time period as the day on which retrospective data were selected to generate hotspot maps against which *future* events could be compared. For simplicity, the 1st April 2010 was selected in order to maximise the use of 6 months of retrospective data for generating KDE hotspot maps, and to use the complete set of 6 months of data after this date for measuring the hotspot maps' abilities for predicting future events. In their analysis that compared two different measurement dates (1st January and 13th March), Chainey et al. (2008a) found no difference in their results. We were therefore confident that the selection of the 1st April 2010 would offer a measurement date that generated representative results.

The retrospective time data was sliced into six time periods and used as input data to generate KDE hotspot maps. This meant that rather than using just one retrospective time period (e.g. the three months prior to the measurement date) which may generate an anomalous result, the use of a number of retrospective time periods would form a more reliable basis on which to draw conclusions. Retrospective input data was sliced into the time periods shown in Table 2a, for each crime type. This concept of using different slices of data as the input data was also followed through to the analysis against measurement data. Six time periods of measurement data were used. This meant that rather than using just one measurement data period for the research (e.g. the three months after the measurement date), the use of a number of measurement data time periods would generate results from which more reliable conclusions could be made. Measurement data was sliced into the time periods shown in Table 2b. This meant that KDE hotspot maps that were generated for each period of input data would be measured for their ability to predict spatial patterns of crime, when the prediction period was the next month, the next two months, and to the next six months.

In their study that compared common hotspot techniques, Chainey et al. (2008a) introduced the Prediction Accuracy Index (PAI). The index was devised as a simple method to allow comparisons between different types of hotspot maps. The index considers the hit rate value (the proportion of crime that occurs within the areas where crimes were predicted to occur i.e. the hots-

Time periods of data used to create KDE hotspot maps					
1 month	2 months	3 months	4 months	5 months	6 months
01 March 2010 - 31 March 2010	01 February 2010 - 31 March 2010	01 January 2010 - 31 March 2010	01 December 2009 - 31 March 2010	01 November 2009 - 31 March 2010	01 October 2009 - 31 March 2010

a

Time periods of data used to measure the spatial prediction abilities of KDE hotspot maps					
1 month	2 months	3 months	4 months	5 months	6 months
01 April 2010 - 30 April 2010	01 April 2010 - 30 May 2010	01 April 2010 – 31 June 2010	01 April 2010 - 31 July 2010	01 April 2010 - 31 August 2010	01 April 2010 - 30 September 2010

b

Table 2. (a) The temporal slices of input data for generating hotspot maps, for a measurement date of the 1st April 2010 and (b) the temporal slices of measurement data for calculating the ability of KDE hotspot maps to predict spatial patterns of crime

pots) against the size of the areas where crimes were predicted to occur (i.e. the areas determined as hotspots), relative to the size of the study area. The PAI is calculated by dividing the hit rate percentage by the area percentage (the area of the hotspots in relation to the whole study area (see Equation 2)).

$$\frac{\left(\frac{n}{N}\right)*100}{\left(\frac{a}{A}\right)*100} = \frac{HitRate}{AreaPercentage} = \text{Prediction Accuracy Index} \quad (2)$$

n: number of crimes in areas where crimes are predicted to occur (e.g. hotspots)
N: number of crimes in study area
a: area (e.g. km^2) of areas where crimes are predicted to occur (e.g. area of hotspots)
A: area (e.g. km^2) of study area

For example, if 25% of future crime events took place in 50% of the study area, the PAI value would equal 0.5; if 20% of future crime events took place in 10% of the area, the PAI would equal 2. Therefore, the higher the PAI, the better the hotspot map for predicting spatial patterns of crime.

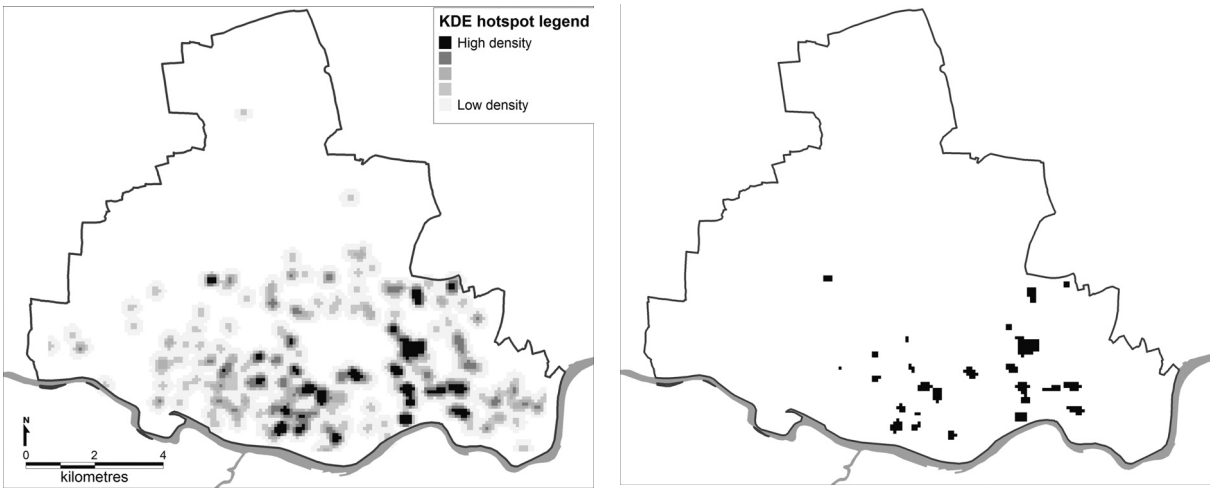
Since the PAI was introduced, other approaches for measuring the predictive abilities of mapping output have been developed. Perhaps the most rigorous of these is proposed by Johnson et al. (2009). The problem with a single measure such as the PAI is that it only offers a comparison between one hit rate and one defined hotspot area, and no comparison against chance expectation. Johnson et al. (2009) proposed the use of an accuracy concentration curve. This is generated by plotting the percentage of crimes that have been accurately predicted (i.e. the hit rate) against the incremental risk ordered percentage of the study area i.e. comparing the number of *future* crimes in 1% of the study area, with this 1% area containing the highest KDE values; comparing the number of *future* crimes in the areas containing the top 2% of KDE values in the study area; comparing the number of *future* crimes in the areas containing the top 3% of KDE values in the study area ..., to comparing the number of *future* crimes in the areas containing 100% of the study area.

However, in Johnson et al.' study (2009) they only compared results between mapping techniques for one input data period (two months) and one output data period (seven days). Calculating an accuracy concentration curve and comparing it against a Monte Carlo simulated result (produced after running at least 19 simulations in order to use a 0.05 level of significance) is practical for comparing one set of data input and output for two different techniques (i.e. two experiments). In our study that uses six different input datasets, and six different output datasets for eight different bandwidth settings and eight different cell size settings (more details on bandwidth and cell size settings are described below), for two types of crime (therefore involving 1152 experiments), and generating 19 Monte Carlo simulations for each experiment, this approach is not practical nor proportionate to the aims of this research – to explore differences between cell size value and bandwidth value, for the same study area, using the same hotspot mapping technique. The use of the PAI has since been discussed further by Pezzuchi (2008), Levine (2008) and Chainey et al. (2008b; 2008c), with researchers concluding it to be a useful measure for comparing multiple hotspot mapping outputs. This has included minimising chance expectation by using the mean PAI results and observing the variation in the standard deviation generated from the many experiments.

Eight cell size values were chosen for comparison: 30 m, 60 m, 90 m, 120 m, 150 m, 180 m, 210 m and 240 m. A value that is often used for the cell size (as referred to in section 2) is the result from measuring the shortest side of the minimum bounding rectangle of the study area, and dividing this distance by 150. Although the choice of 150 is rather arbitrary, in practice it provides a useful starting measure and is the procedure that is used to calculate cell size in the popular Hotspot Detective for MapInfo software (Ratcliffe, 2002). This gave the value of 89.6 (rounded upto 90 m). We therefore felt it useful to generate results for this measure in comparison to other cell size values, using multiples of 30 m in our cell size experiments. For each cell size experiment, the bandwidth was controlled to a single size: a bandwidth of 450 m was used, as per the guidance described in section 2.

Eight bandwidth size values were chosen for comparison: 100 m, 200 m, 300 m, 400 m, 500 m, 600 m, 700 m and 800 m. If we had followed the recommendations of Chainey (2011) (i.e. five times the cell size) this would have suggested a bandwidth value of 450 m. Rather than use multiples of 150 m, we decided to use multiples of 100 m in order to explore the influence of a small bandwidth (100 m), to help more simply present results, but still enable a comparison between the outputs generated between 400 m and 500 m as an indication of the effectiveness of this rather crude approach for determining bandwidth size. For each bandwidth size experiment, the cell size was controlled to a single size: a cell size of 90 m was used, as per the guidance described in section 2.

To identify if predicted spatial patterns of crime generated by KDE hotspot maps under different cell size and bandwidth settings differed, Prediction Accuracy Index measures were aggregated and averaged for the periods of input data and for the periods of measurement data. This meant that the PAI measures could be compared, with any differences being explained in relation to the



(a) KDE hotspot map (b) Top thematic class of KDE hotspot map

Figure 3. Hotspots were determined by selecting the top thematic class calculated using five classes and the default values generated from applying the quantile thematic range method in MapInfo

cell size and bandwidth size rather than different periods of input and measurement data. This approach was applied separately to the two crime datasets: residential burglary and assault with injury. The standard deviation and coefficient of variation of the PAI for each crime type across the eight different cell size values and eight bandwidth values were also calculated.

During the data time period (1st October 2009 to 30th September 2010) there could have been police operations and crime reduction initiatives that had an impact on crime levels, plus there could have been an impact from seasonal influences. For this study, because the focus was on comparing KDE parameter settings against the same data, any changes in crime patterns would have the same impact on different cell size and bandwidth size parameter entries and would not affect the ability to examine results and draw conclusions on the analyses.

A final parameter to consider for KDE hotspot map generation is a threshold value for determining which areas are *hot*. For purposes of research comparison, we followed the methodology used by Chainey et al. (2008a). This involved using five thematic classes and default values generated from using the quantile thematic classification method in MapInfo. *Hot* was then determined by the top thematic class (Figure 3).

IV. RESULTS

A. The influence of cell size on KDE hotspot maps for predicting where crime may occur

Table 3 shows the PAI results for residential burglary and assaults with injury for different cell sizes. The PAI results for residential burglary varied between 6.6 for a cell size of 240 m to 7.1 for 30 m and 60 m cell sizes. The PAI results for assaults with injury were much higher than those for residential burglary, but

	Residential burglary			Assaults with injury		
Cell size (m)	**PAI**	**SD**	**CV**	**PAI**	**SD**	**CV**
30	7.1	0.60	0.08	68.4	3.06	0.04
60	7.1	0.66	0.09	68.5	2.96	0.04
90	6.7	0.53	0.08	65.1	2.66	0.04
120	6.9	0.57	0.08	64.5	2.50	0.04
150	6.7	0.53	0.08	63.0	3.17	0.05
180	6.8	0.64	0.10	64.3	2.95	0.05
210	6.7	0.46	0.07	59.9	2.39	0.04
240	6.6	0.48	0.07	60.2	2.85	0.05

Table 3. KDE hotspot map PAI, standard deviation (SD) and coefficient of variation (CV) results for residential burglary and assaults with injury for different cell sizes

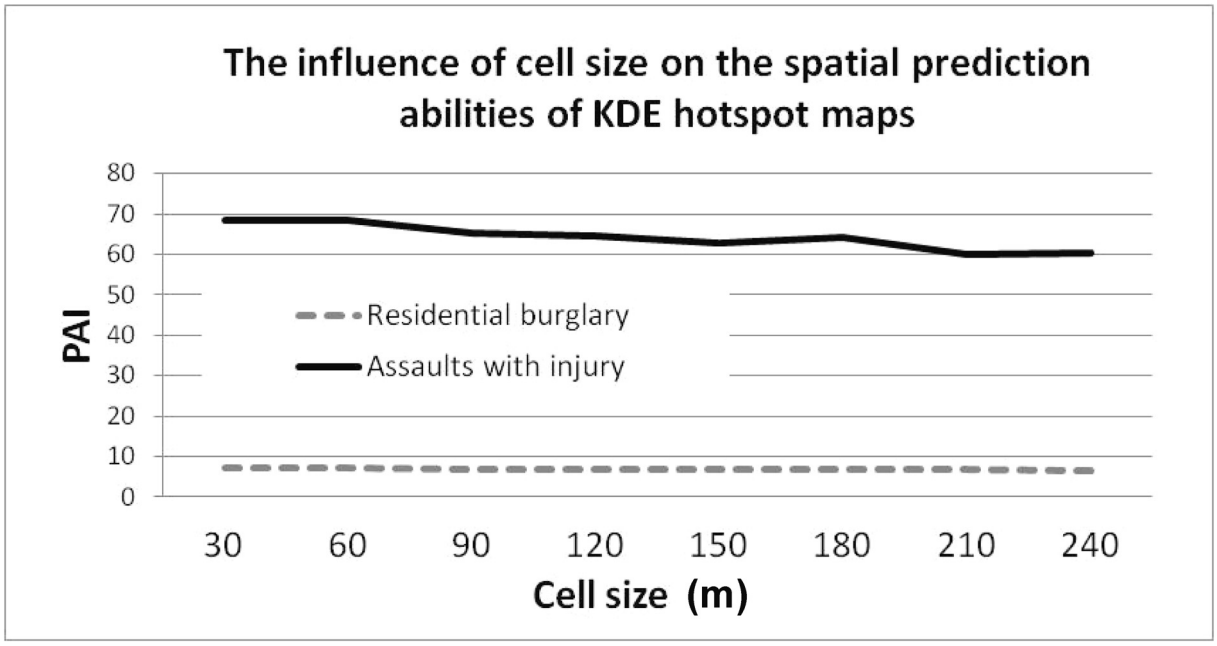


Figure 4. The influence of cell size on KDE hotspot map PAI values for residential burglary and assaults with injury

again showed only a small amount of relative variation from 59.9 for a cell size of 210 m to 68.5 for a cell size of 60 m. These results suggest that although PAI values decrease with increases in cell size, this difference is marginal. These results are also shown in Figure 4. There was little statistical variation in the results for each cell size, as indicated by the low coefficient of variation (CV) values, and little difference in the CV values between cell sizes.

The similarity in results for different cell sizes is further illustrated by the difference in the number of crimes that maps of different cell sizes predict in KDE generated hotspot areas (Table 4). When the KDE hotspot areas were controlled to identify 1% of the total study area (i.e. the 1% of areas with the highest KDE values), generated from 3 months of input data using cell sizes of 30 m and 240 m to predict where crimes would occur in the next 3 months, very similar results were produced: for residential burglary, KDE outputs generated using a 30 m cell size predicted 29 crimes, in comparison to 28 crimes using a cell size of 240 m; for assaults with injury, KDE outputs generated using a 30 m cell size predicted 158 crimes, in comparison to 153 crimes using a cell size of 240 m. That is, as the spatial resolution of the KDE hotspot map begins to degrade, the ability of the map to predict where crime occurs in the future reduces only slightly.

A. The influence of bandwidth size on KDE hotspot maps for predicting where crime may occur

Table 5 shows the PAI results for residential burglary and assaults with injury for different bandwidth sizes. The PAI results for residential burglary varied between 5.6 for bandwidth sizes of 700 m to 13.1 for 100 m bandwidth sizes. The PAI results for assaults

Crime type and cell size (m)	Crimes committed April – June 2010	Number of crimes in hotspots (1% of area)	Percentage of crimes in hotspots
Residential burglary: 30 m	329	29	8.8%
Residential burglary: 240 m	329	28	8.5%
Assaults with injury: 30 m	459	158	34.3%
Assaults with injury: 240 m	459	153	33.3%

Table 4. Crimes predicted using kernel density estimation outputs of difference cell sizes for residential burglary and assaults with injury, based on using three months of input crime data (January – March 2010) and 3 months of measurement data (April – June 2010). The area determined as *hot* was controlled to cover 1% of the study area's total area.

	Residential burglary			Assaults with injury		
Bandwidth size (m)	**PAI**	**SD**	**CV**	**PAI**	**SD**	**CV**
100	13.1	2.8	0.22	142.8	11.53	0.08
200	11.1	1.3	0.12	91.7	4.65	0.05
300	8.7	1.0	0.12	79.4	3.03	0.04
400	7.1	0.7	0.10	68.3	3.54	0.05
500	6.5	0.6	0.09	60.2	2.60	0.04
600	5.9	0.6	0.11	54.3	2.52	0.05
700	5.6	0.6	0.11	48.6	2.23	0.05
800	5.7	0.5	0.09	42.9	1.98	0.05

Table 5. KDE hotspot map PAI, standard deviation (SD) and coefficient of variation (CV) values for residential burglary and assaults with injury for different bandwidth sizes

with injury were much higher than those for residential burglary, but also showed large variation from 42.9 for bandwidth sizes of 800 m to 142.8 for bandwidth sizes of 100 m. These results suggest that as bandwidth size increases, the power of the KDE hotspot map to predict spatial patterns of crime degrades. These results are also shown in Figure 5. With the exception of residential burglary KDE hotspot maps generated using a bandwidth of 100 m, there was little statistical variation in the results for each bandwidth size and little difference in the CV values between cell sizes.

The difference in results for different bandwidth sizes is further illustrated by the difference in the number of crimes that maps of different bandwidth sizes predict in hotspots generated using KDE (Table 6). To illustrate this (and to allow for easier comparisons with future research) we controlled the KDE hotspot areas to identify only the top 1% of density values (i.e. the 1% of areas with the highest KDE values), generated from 3 months of input data using bandwidth sizes of 100 m and 800 m to predict where crimes would occur in the next 3 months. For residential burglary, KDE outputs generated using a 100 m bandwidth size predicted 35 crimes (i.e. 11% of all burglaries in just 1% of the study area), in comparison to 22 crimes using a bandwidth size of 800 m; for assaults with injury, KDE outputs generated using a 100 m bandwidth size predicted 166 crimes (i.e.

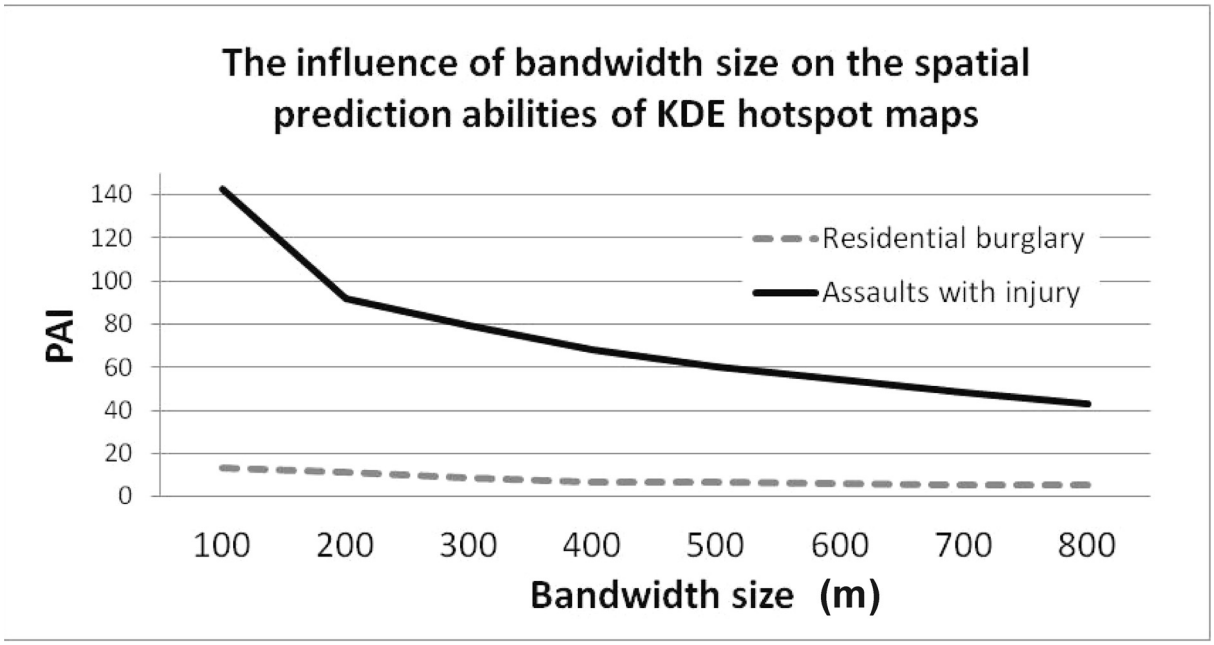


Figure 5. The influence of bandwidth (m) size on KDE hotspot map PAI values for residential burglary and assaults with injury.

Crime type and bandwidth size (m)	Crimes committed April – June 2010	Number of crimes in hotspots (1% of area)	Percentage of crimes in hotspots
Residential burglary: 100 m	329	35	10.6%
Residential burglary: 800 m	329	22	6.7%
Assaults with injury: 100 m	459	166	36.2%
Assaults with injury: 800 m	459	137	29.8%

Table 6. Crimes predicted using kernel density estimation outputs of difference bandwidth sizes for residential burglary and assaults with injury, based on using three months of input crime data (January – March 2010) and 3 months of measurement data (April – June 2010). The area determined as *hot* was controlled to cover 1% of the study area's total area.

36% of all violent assaults in 1% of the study area), in comparison to 137 crimes using a bandwidth size of 800 m. That is, as the smoothing of the KDE hotspot map increases, the ability of the map to predict where crime occurs degrades. These results also illustrate the proportion of crime that KDE hotspot maps can predict.

V. DISCUSSION AND IMPLICATIONS

The findings from this research show that KDE hotspot maps generated using different cell sizes have little impact on the mapping outputs ability to predict spatial patterns of crime, but that different bandwidth sizes do have an impact. Cell size mainly impacts on the visual appeal of the KDE mapping output, with higher resolutions producing maps that avoid the *blocky* pixilation of outputs generated using larger cell sizes. For example, the maps shown in Figure 6 are equally as good as each other for predicting where crime may occur in the future, but Figure 6a is the more preferable output due to its better visual appeal. While smaller cell sizes require greater computer processing due to the larger number of calculations that are required, in our experiments this extra length of processing was not a significant impairment.

Bandwidth size does though affect the ability of KDE hotspot maps to predict spatial patterns of crime. For example, the maps shown in Figure 7 were generated using the same period of input data but have very different PAI values. That is, the smaller the bandwidth, the better the KDE map is at predicting spatial patterns of crime.

The research has also shown the large variation that exists between the ability to predict different types of crime using KDE. This was initially shown by Chainey et al. (2008a), with street robbery KDE maps generating higher PAI values than KDE hotspot maps of residential burglary, and vehicle crime. The PAI results for residential burglary in this study of crime in Newcastle-upon-Tyne are higher than those found by Chainey et al. (2008a) for residential burglary in London, indicating differences between areas. However, it is the high PAI values generated for violent assaults that offer new insights into the spatial prediction of KDE hotspot maps. This is reflected by the manner in which violent assaults

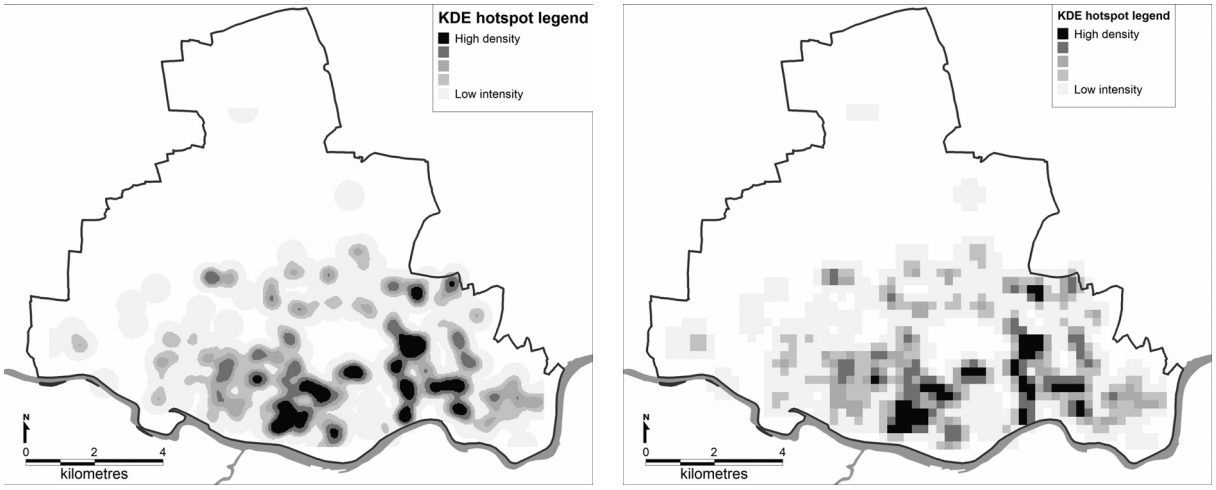


Figure 6. A comparison of KDE hotspot maps generated using the same bandwidth but with different cell sizes (a) 30 metres (PAI of 7.0) and (b) 240 metres (PAI of 7.0)

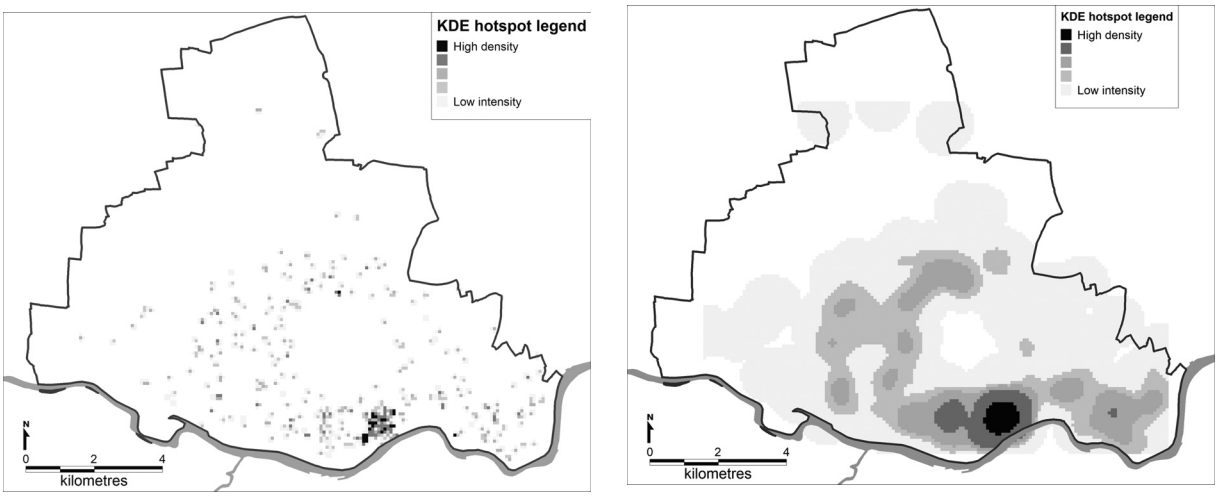
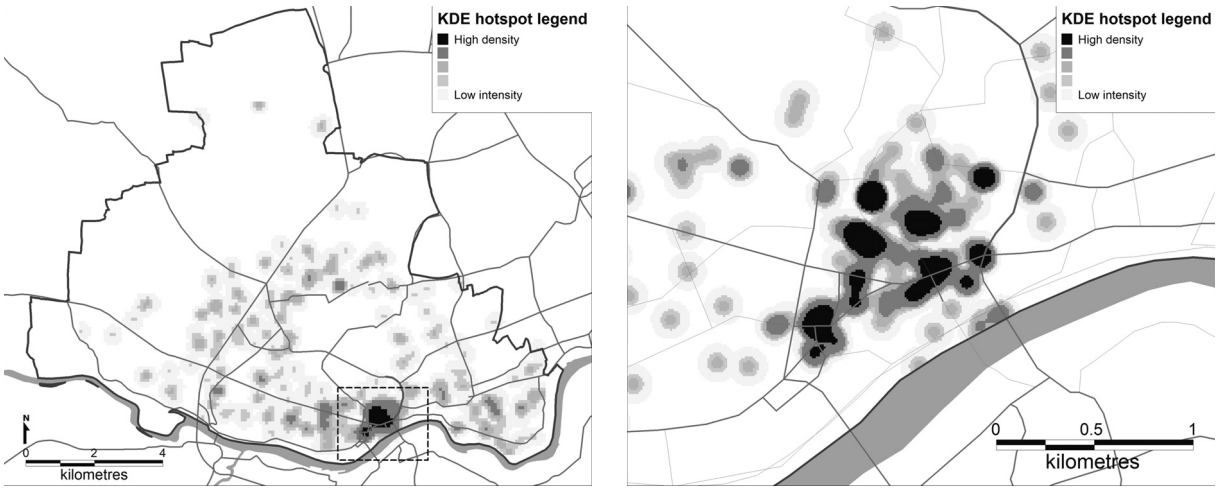


Figure 7. A comparison of KDE hotspot maps generated using the same cell size but with different bandwidth sizes (a) 100 metres (PAI of 119.3) and (b) 800 metres (PAI of 40.4)

cluster in comparison to burglary. Whilst burglary does concentrate spatially, these hotspots tend to be larger in number and more dispersed. This is most likely due to the wider (in spatial terms) opportunity for burglary, with residential properties spread geographically across areas. Areas where violent assaults take place tend to be highly concentrated in areas that are associated with alcohol and the night-time economy (Maguire and Hopkins, 2003; Babor et al., 2003; Graham and Homel, 2008). Newcastle's night-time economy is heavily concentrated in the city centre, therefore the occasional violent interaction between people in this highly compact area heavily influences the spatial distribution of this type of crime. That is, the highly compact nature of the night-time economy has a direct impact on the highly compact, and predictable nature of where violent assaults are most concentrated.

The analysis of different cell sizes and bandwidths also offers practitioners the means to better qualify the default parameter values that are determined by Geographical Information System products such as ESRI's ArcGIS Spatial Analyst, Crime Analyst, and Hotspot Detective for MapInfo. Our results indicate that defaults for cell size such as those generated using Hotspot Detective (which involves dividing the shorter side of the MBR by 150) offer a useful starting point, but that reducing this value further will generate maps of greater visual appeal without affecting the maps ability to predict where crime is likely to occur in the future. However, bandwidth default values need further scrutiny by practitioners to ensure they are not too large and impair the purpose of the KDE hotspot mapping output. For example, the default Hotspot Detective KDE bandwidth size for three months of violent assaults data for



(a) (b)

Figure 8. A procedure for creating precise and practical KDE hotspot maps for accurately assisting in the targeting of policing and crime reduction resources: (a) is a KDE hotspot map generated for a large area for identifying the key strategic areas for focus (bandwidth 300 m; cell size 90 m). Once a focus area is identified data for this area is selected, and a KDE hotspot map is generated using a smaller bandwidth (100 m) and cell size (10m).

Newcastle-upon-Tyne was 450 m – a bandwidth size that generated a PAI value of 60 compared to a PAI of 143 if a bandwidth of 100 m was used.

However, low bandwidth values produce KDE hotspot maps that appear *spikey*, with many small areas identified as hotspots. In practice, this type of hotspot map is often considered unsuitable because it does not identify a small number of areas for strategic attention. Therefore, it is argued that a balance is required between KDE hotspot prediction accuracy, and output that is useful in practice. A way in which this can be overcome is to use a bandwidth size that is large enough to initially identify key hotspot areas for strategic attention, with these areas then being focused upon in more detail with a second hotspot map generated based on the distribution of crime in this focus area. Figure 8 shows an example of this – Figure 8a uses a bandwidth size of 300 m and cell size of 90 m to identify the main assaults hotspots in Newcastle-upon-Tyne. The main hotspot then becomes the area of attention, with a second KDE hotspot map generated for this area to more precisely identify the areas that are required for police attention. Figure 8b was generated using a bandwidth of 100 m and a cell size of 10m.

KDE is though not without its weaknesses. The procedure described above would fail to identify areas where there is a high and compact concentration of crime because larger bandwidths have the tendency to smooth these out over the area it generates density values for. An additional weakness is that the use of KDE requires the researcher to determine what is *hot* by deciding the value for the top thematic class. In this research we standardised this procedure by using the quantile thematic classification method in all experiments. However, most GIS software offer several options for the user to determine a thematic classification method preference, leading to subjectivity in hotspot mapping output. This calls for further research that identifies hotspot mapping methods that can overcome these KDE weaknesses.

VI. CONCLUSION

Hotspot analysis is a basic form of crime prediction – using crime data from the past to predict where crime may occur in the future, with the outputs from hotspot mapping being used by in practice for determining where police patrols and other crime prevention initiatives should be targeted. Kernel density estimation has become the most popular technique used in practice for identifying hotspots of crime.

Cell size and bandwidth size are the two main parameters that the user is required to enter in order to generate KDE hotspot mapping output. The findings from this research illustrate that cell size has little impact on a KDE hotspot map's ability to predict spatial patterns of crime, but that smaller sizes generate hotspots maps of greater visual appeal. Bandwidth size does though have an impact on a KDE hotspot output's ability to predict spatial patterns of crime, with the spatial prediction ability of the KDE hotspot map degrading as bandwidth size increases. To date, most users of KDE for hotspot mapping make use of default settings for cell size and bandwidth size, without qualifying these values. This research has helped to identify the influence these parameters have, and in so doing offer practitioners and researchers a more informed basis on which to qualify the values they should use for cell size and bandwidths for producing KDE hotspot maps.

REFERENCES

Babor, T., Caetano, R., Casswell, S., Edwards, G., Giesbrecht, N., Graham, K., Grube, J., Gruenewald, P., Hill, L., Holder, H., Homel, R., Osterberg, E., Rehm, J., Room, R., Rossow, I. (2003). *Alcohol : No ordinary commodity. Research and public policy*. Oxford: Oxford University Press.

Bailey, T.C., and Gatrell, A.C. (1995) *Interactive Spatial Data Analysis*. Reading, Massachusetts: Addison-Wesley.

Beato, C. (2008). *Comprehendo e Avaliando: Projetos de Segurança Publica*. Belo Horizonte: Editora UFMG.

Bithell, J.F. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine,* 9, 691-701.

Block, R. and Block, R.B. (2000) The Bronx and Chicago – Street Robbery and the Environs of Rapid Transit Stations. In Goldsmith, V., McGuire, P.G., Mollenkopf, J.H. and Ross, T.A. (Eds.) *Analysing Crime Patterns: Frontiers and Practice* (p. 137-152). Thousand Oaks (CA): Sage.

Bowman, A. and Azzelini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Braga, A. (2007). *Effects of Hot Spots Policing on Crime: A Campbell Collaboration Systematic Review.* http://www.aic.gov.au/campbellcj/reviews/titles.html

Brunsdon, C. (1995). Analysis of Univariate Census Data. In S. Openshaw (ed.) *Census Users Handbook* (p. 213-238). Cambridge: Geoinformation International.

Chainey, S.P. (2011). Identifying hotspots: an assessment of common techniques. Presentation at the International Crime and Intelligence Analysis Conference November 2011, Manchester, England.

Chainey, S.P. and Ratcliffe, J.H. (2005) *GIS and Crime Mapping*. London: Wiley.

Chainey,S.P., Tompson,L., Uhlig,S. (2008a). The utility

of hotspot mapping for predicting spatial patterns of crime. *Security Journal*. 21, 1-2.

Chainey,S.P., Tompson,L., Uhlig,S. (2008b). Response to Pezzuchi. Security Journal, 21:4.

Chainey,S.P., Tompson,L., Uhlig,S. (2008c). Response to Levine. Security Journal, 21:4.

Eck, J.E., Chainey, S.P., Cameron, J.G., Leitner, M. and Wilson, R.E. (2005) *Mapping Crime: Understanding Hot Spots*. USA: National Institute of Justice.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis.* London: Sage.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships.* Chichester: Wiley.

Graham, K. and Homel, R. (2008). *Raising the bar: preventing aggression in an around bars, clubs and pubs*. Cullompton: Willan Publishing.

Home Office (2005) *Crime Mapping: Improving Performance, A Good Practice Guide for Front Line Officers*. London: Home Office.

Johnson, S.D., Bowers, K.J., Birks, D.J., and Pease, K. (2009). Predictive mapping of crime by ProMap: accuracy, units of analysis, and the environmental backcloth. In D. Weisburd, W. Bernasco and G.J.N. Bruinsma eds., *Putting Crime in its Place* (p. 165–192). Dordrecht: Springer.

LeBeau, J.L. (2001) Mapping Out Hazardous Space for Police Work. In Bowers, K. and Hirschfield, A. *Mapping and Analysing Crime Data – Lessons from Research and Practice*. London: Taylor and Francis.

Levine, N. (2004) *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Houston (TX): Ned Levine and Associates. Washington (DC): National Institute of Justice.

Levine, N. (2008). The Hottest Part of a Hotspot: comments on The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 21, 4.

Maguire, M. and Hopkins, M. (2003). Data analysis for problem-solving: alcohol and city centre violence. In Bullock, K. and Tilley, N. (eds) *Crime Reduction and Problem-Oriented Policing* (p. 24-38). Cullompton: Willan Publishing.

Mashford, T. (2008). Methods for implementing crime mapping within a large law enforcement agency; experiences from Victoria, Australia. In S.P. Chainey and L. Tompson, eds., Crime Mapping Case Studies: Practice and Research (p. 19-26). London: Wiley.

National Policing Improvement Agency (2010). The analysis of geographic information - workbook. Wyboston: NPIA.

Openshaw, S. (1984) *The Modifiable Areal Unit Problem*. Concepts and Techniques in Modern Geography 38. Norwich: Geobooks

Pezzuchi, G. (2008). A brief commentary on The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, 21,4.

Ratcliffe, J., (2002) *HotSpot Detective 2.0 for MapInfo Professional 7.0*. Philadelphia: JHR Systems.

Ratcliffe, J. and McCullagh, M. (2001) Crime, Repeat Victimisation and GIS. In Bowers, K. and Hirschfield, A. *Mapping and Analysing Crime Data – Lessons from Research and Practice* (p. 61-92). London: Taylor and Francis.

Ratcliffe, JH, Taniguchi, T, Groff, E and Wood, J (2011) The Philadelphia Foot Patrol Experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology*. 49(3): 795-831.

Sherman, L. (2009). Hot spots. In Wakefield, A. and Fleming, J. *The SAGE Dictionary of Policing*. London: Sage Publications.

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

Uhlig, S. (2005). *Examining the Prediction Accuracy of Crime Hotspots* (Unpublished Master Thesis in GIS, University College London).

Coordonnées de l'auteur:

Spencer CHAINEY
University College London,
Department of Security and Crime Science,
35 Tavistock Square, London, WC1E 9EZ.
England.
s.chainey@ucl.ac.uk

BSGLg, 60, 2013, 21-31

ESTIMATEUR À NOYAU (KDE) SUR RÉSEAUX: UNE APPLICATION AUX ACCIDENTS DE LA ROUTE BELGES

David DABIN, Christiane DICKENS et Paul WOUTERS

Résumé
Le problème de la détection des zones de haute concentration des accidents de la route est un sujet de première importance tant pour les décideurs que pour les gestionnaires des infrastructures routières. Cependant, la notion même de point noir reste sujette à de vives discussions entre les experts du domaine bien que des définitions fonctionnelles existent dans plusieurs pays (Elvik, 2008). Les différents outils utilisés par les autorités en Belgique négligent plusieurs dimensions importantes des données d'accident notamment l'aspect réseau dont elles proviennent, l'aspect stochastique des accidents, les éventuelles erreurs de localisation de ceux-ci et rendent impossible la définition de zones noires d'accidents de la route. Dans ce travail, nous proposons une méthodologie mixte de trois étapes: (i) la fonction de densité des événements ponctuels sur le réseau est évaluée par un KDE, (ii) la significativité des valeurs lissées observées est calculée par des simulations MC et (iii) des outils d'analyse des réseaux permettent de classifier les éléments significatifs en *hotspots* et *hotzones*. La méthodologie est testée sur l'entièreté des routes numérotées belges pour la période 2006-2009.

Mots clés
accident de la route - point noir - zone noire - estimateur de densité à noyau - réseau

Abstract
The detection of high accident concentrations on roads is of vital importance for stakeholders but also for road safety managers. However experts in the field do not agree about the notion of black spots itself despite the existence of different practical definitions in many countries (Elvik, 2008). The different tools currently in use at the Belgian scale neglect some important dimensions of the accidents data such as the fact that accidents occur only on the road network, the spatial random component of some accidents, the localization problem and the impossibility to define black zones of road accidents. We propose here a new methodology based on three steps: (i) the density function of point event is estimated using a Kernel Density Estimator (KDE), (ii) the p-value of observed values against Complete Spatial Randomness (CSR) is computed through Monte-Carlo simulations and (iii) the significant items are classified into hotspots and hotzones by network analysis tools. The methodology is then evaluated on all road accidents with casualties on the Belgian numbered road network.

Keywords
road accident - hotspot - hotzone - KDE - network

1. INTRODUCTION

Les États Généraux de la Sécurite Routière de 2011 ont fixé un objectif de 620 tués maximum sur les routes belges en 2015 et de 420 tués en 2020. Un effort conséquent doit donc encore être consenti en matière de sécurité routière afin de tenir ces objectifs d'autant que le travail devient de plus en plus complexe à mesure que les progrès sont engrangés car plus le nombre de tués se réduit, plus les nouvelles diminutions de mortalité sont difficiles à obtenir (Geurts et Wets, 2003 ; Casteels *et al.*, 2010). Afin de porter ses fruits, cet effort supplémentaire doit viser au minimum chacune des 3 composantes clés de la sécurité routière que sont (i) les infrastructures, (ii) le niveau de sécurité et de sûreté des véhicules et, finalement, (iii) le conducteur (Iversen et Rundmo, 2002 ; Castellà et Pérez, 2004). Dans le contexte de ressources budgétaires limitées, la question est de savoir où doivent être portés ces efforts pour obtenir un effet maximal.

En Belgique, plusieurs méthodes sont actuellement utilisées par les autorités afin d'identifier les éléments ponctuels problématiques nommés classiquement points noirs ou *hotspots* dans le domaine de la circulation routière.

- Comptage du nombre absolu d'accidents par km. Toutes les bornes hectométriques dont la fréquence des accidents dépasse une valeur seuil sont alors considérées comme des points noirs.

- Indice de dangerosité pondéré par la gravité des blessures encourues défini par la relation

$$S(i) = LI(i) + 3\ SI(i) + 5\ DI(i)$$

où $S(i)$ est l'indice de dangerosité de la borne i, LI est le nombre de blessés légers, SI est le nombre de blessés graves et DI le nombre de tués (Geurts et Wets, 2003). Cet indice est calculé pour toutes les bornes hectométriques dont le nombre d'accidents de roulage avec lésions corporelles est de minimum 3 pour les 3 dernières années. Les bornes hectométriques qui présentent une valeur supérieure à 15 sont alors considérées comme dangereuses et prioritaires.

- Indice de risque spatio-temporel (Romano et Heuchenne, 1996 ; Romano, 1997 ; Antoine, 2010). Route par route, un indice mensuel est d'abord calculé sur une matrice espace-temps qui représente une route sur la période de temps donnée. Cet indice se construit par une fenêtre mobile de 200 m et de 5 mois c'est-à-dire que l'indice est fonction du nombre d'accidents à la borne hectométrique elle-même et de ses deux voisines de part et d'autre mais aussi en fonction des accidents sur les 5 mois précédents et des 5 mois suivants. L'indice de risque pour une borne hectométrique est ensuite calculé comme la moyenne des indices mensuels à cette borne. Des valeurs supérieures à 2,4 sont classifiées en tant que zones à haut risque alors que des valeurs comprises entre 1,2 et 2,3 sont considérées comme des zones à risque moyen.

L'exploitation des résultats de ces analyses *hotspots* démontre cependant plusieurs limitations.

– L'impossibilité de définir des zones noires ou *hotzones*. Les autorités responsables des infrastructures ainsi que celles répondant des services de police visent un maximum d'efficience. Comme il est plus facile de traiter des segments de route moins nombreux mais plus longs, une approche par zone noire semble plus appropriée que celle par point noir.

– La négation du réseau routier. En effet, les indices de dangerosité de Geurts et Wets (2003) et de risque de Romano et Heuchenne (1996) sont calculés route par route en se basant sur les bornes hectométriques mais en négligeant totalement les carrefours. Les relations entre les routes disparaissent donc totalement.

– La nature stochastique des accidents de roulage. Un certain nombre d'accidents sont provoqués par des causes qui ne dépendent pas directement du milieu environnant. Le site de l'accident et ses caractéristiques propres telles que les infrastructures ne permettent pas d'expliquer entièrement ou en partie le sinistre. Les accidents revêtent alors un caractère purement aléatoire du point de vue spatial. Ils ont été observés à la borne hectométrique i cette année mais pourraient très bien se produire à la borne i+1 ou i−1 l'année prochaine. Par une approche purement fréquentiste, il est impossible de tenir compte de cette nature aléatoire.

– La prise en compte des erreurs d'allocation. La qualité de l'encodage du lieu des accidents laisse à désirer. En effet, sur la période 2006-2009, ce sont près de 1/5 des accidents qui n'ont pas pu être géocodés. Plusieurs raisons expliquent ce défaut d'encodage des coordonnées spatiales des accidents dont notamment les difficultés à trouver la borne hectométrique (BH) ou l'adresse la plus proche lors de l'enregistrement (BH cachée, absente, accidents en rase campagne sans maison proche), la méconnaissance du lieu et du nom des rues, une modification de la structure du réseau routier par l'introduction de nouvelles routes, le manque de temps en opération, l'absence de motivation à remplir ces données…

Des développements récents mettent en avant de nouvelles possibilités afin de dépasser ces limitations. Citons entre autres, les travaux de Flahaut *et al*. (2003), Elvik (2008), Xie et Yan (2008), Moons *et al*. (2009), Okabe *et al*. (2009), Steenberghen *et al*. (2010), Manepalli *et al*. (2011) et Truong et Somenahalli (2011). Parmi toutes ces méthodologies, l'estimateur de densité à noyau (Kernel Density Estimator, KDE) est couramment utilisé pour étudier la distribution d'évènements ponctuels et identifier les zones de hautes concentrations ou *hotspots* au sein d'un espace homogène par l'ajustement d'une surface de densité (Xie et Yan, 2008 ; Anderson, 2009). Cette technique est même considérée comme la carte la plus populaire en analyse criminelle après la *pin map* (Smith et Bruce, 2008) et est évaluée comme une des méthodes les plus performantes pour prédire les patterns spatiaux de crimes futurs (Chainey *et al*., 2008).

Cependant, de nombreux phénomènes ponctuels ne se déroulent pas librement dans l'espace mais sont naturellement contraints à un réseau. Au niveau de la problématique de sécurité, citons par exemple les vols de cuivre sur le réseau ferroviaire, le problème des pickpockets dans les transports publics, le vandalisme dans les espaces publics, les vols de voiture ou encore les accidents de la circulation. Pour ces exemples, l'utilisation de méthodes bidimensionnelles classiques qui supposent l'homogénéité de l'espace dans toutes les directions semble donc hasardeuse (Xie et Yan, 2008). Un réseau est en effet un espace particulier dont la dimension de 1,5 est comprise entre la ligne (1 D) et le plan (2 D) (Steenberghen *et al*., 2010).

Dans ce cadre, les objectifs de ce travail sont triples: (i) établir une méthodologie pour détecter les *hotzones* d'accidents de la circulation sur un réseau routier complexe, (ii) déterminer les paramètres optimaux de cette méthodologie pour l'analyse au niveau de la Belgique et (iii) tester cette méthodologie sur base des Accidents de Roulage avec Lésions Corporelles (AccRLC) sur les routes principales belges.

II. DONNÉES

L'analyse porte sur l'entièreté du réseau des routes principales de Belgique, c'est-à-dire toutes les routes du réseau numéroté. Ce réseau représente environ 16.718 km répartis en autoroutes, routes nationales et routes régionales. Notons finalement que les routes numérotées disposent en Belgique d'un système de référencement nommé « borne hectométrique » car il prend la forme de bornes placées approximativement tous les 100 m.

Les données d'accidents de roulage sont issues des Formulaires d'Accidents de la Circulation (FAC) rédigés par les forces de Police pour tous les AccRLC. Dans ces FAC, les accidents de roulage sont localisés selon un des 3 systèmes suivants illustrés en Figure 1:

- Borne Hectométrique (BH): commune, numéro de route, borne hectométrique (BH) la plus proche;
- Adresse: commune, nom de rue et numéro de maison la plus proche;
- Carrefour: commune, nom de rue 1 et nom de rue 2.

Cependant parmi ces accidents, ceux localisés par l'adresse la plus proche sont alloués lors du post-traitement à la borne hectométrique ou au carrefour le plus proche pour des raisons de cohérence et de consistance des données. Le référentiel spatial de localisation des AccRLC sur le réseau des routes numérotées belges est alors composé de 185.475 bornes hectométriques et 11.404 carrefours soit 196.879 points de mesure. Le processus d'allocation au point de mesure le plus proche fait apparaître des aires de chalandise autour de chacun de ces points. Chaque aire de chalandise, également connue sous le nom de segment « Voronoï », correspond à la portion du réseau qui verra tous les accidents s'y déroulant être alloués au point de mesure central (Figure 1). Notons que ce phénomène est constaté pour tous les processus nécessitant un géocodage. D'un phénomène purement continu pouvant se dérouler partout sur le réseau, le géocodage retourne seulement un nombre fini de points passant d'un problème continu à un problème discret.

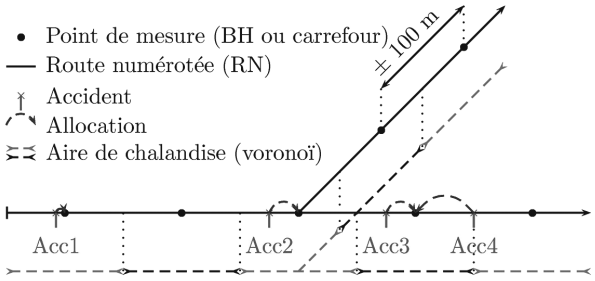


Figure 1. Système de localisation des accidents par les bornes hectométriques (BH) et les carrefours et segments « Voronoï » induits sur les routes nationales (RN)

III. MÉTHODE

Considérons un réseau G=(N, L) composé de N noeuds (*vertices*) et L lignes (*edges*) illustré en Figure 2. Chaque noeud i appartenant à l'ensemble N possède des caractéristiques propres telles que son aire de chalandise (segment « Voronoï » sous-jacent) et le nombre d'évènements ponctuels qui y ont pris place. Chaque noeud se trouve connecté à ses voisins j par des lignes l_{ij} de longueur $|l_{ij}|$. Le parallèle avec la localisation des accidents par les points de mesure est évident.

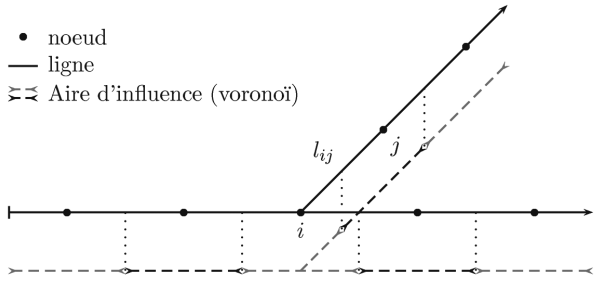


Figure 2. Description des éléments constitutifs du réseau

Pour un noeud i quelconque, nous pouvons définir un voisinage V_i qui contient tous les noeuds se trouvant à une distance inférieure ou égale à h de ce noeud. Dans un estimateur à noyau de densité (KDE pour *Kernel Density Estimator*) classique, cette distance est de type euclidien alors qu'elle sera calculée le long du réseau dans ce cas-ci. L'impact de ce changement de distance est évident pour l'exemple de la Figure 3 où le point p

fait partie du voisinage selon la distance euclidienne et en est exclu selon la distance le long du réseau. La distance le long du réseau paraît également plus pertinente avec une application sur des routes européennes qui présentent un tracé moins régulier que le réseau routier en damier des Etats-Unis par exemple.

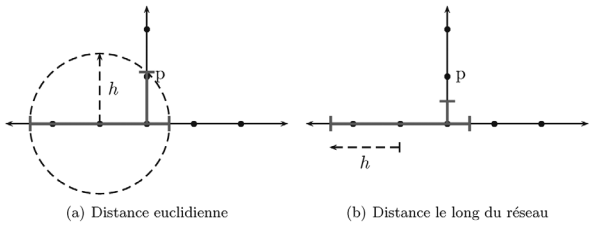


Figure 3. Calcul du voisinage sur base des (a) distances euclidiennes et des (b) distances le long du réseau

Nous pouvons également définir une fonction k(x_i) quelconque qui satisfait à deux conditions

$$k(x_i) \begin{cases} \geq 0 & \forall x \in V_i \\ = 0 & \forall x \notin V_i \end{cases} \quad (1)$$

et,

$$\int_{x \in V_i} k(x_i)dx = 1 \quad (2)$$

Classiquement, la distance h qui défini le voisinage V est dénommée fenêtre ou *bandwidth* et $k(x_i)$ noyau ou *kernel function*. L'estimation de la densité à un point j est alors la somme sur l'ensemble du voisinage V_j des noyaux multipliés par le nombre d'occurrences n_i.

$$K(x_j) = \sum_{i \in V_j} k(x_i) n_i \quad (3)$$

Au delà de sa formulation mathématique, un estimateur à noyau peut se concevoir simplement comme le remplacement des évènements ponctuels par un noyau dont la masse vaut 1 mais répartie sur une fenêtre de largeur h. À chaque point du réseau, la somme des noyaux définit la densité lissée. Dans l'Équation (3), 2 paramètres apparaissent: (i) la forme de la fonction $k(x_i)$ et (ii) la fenêtre h qui définit le voisinage V.

A. Choix du noyau

Les noyaux classiques incluent notamment le noyau uniforme, le noyau d'Epanechnikov, le noyau gaussien ou encore le noyau triangulaire. À fenêtre égale, le noyau gaussien présente la densité la plus élevée à l'origine par rapport au noyau triangulaire et au noyau d'Epanechnikov. Cette relation s'inverse logiquement à une certaine distance en raison de la contrainte de l'Équation (2). Le noyau gaussien donne plus de poids aux points très proches alors que l'Epenechnikov tend à distribuer la densité plus loin et donc à lisser plus fort les données. Cependant bien que différents, de nombreux auteurs (Silverman, 1986; Bailey et Gatrell, 1995; O'Sullivan et Unwin, 2002; Schabenberger et Gotway, 2005; O'Sullivan et Wong, 2007 cités par Xie et Yan, 2008) suggèrent que la forme du noyau est asymptotiquement de moindre importance que la fenêtre sur le résultat final.

En plus de la distance mesurée le long du réseau, l'utilisation d'un noyau classique nécessite toutefois une seconde correction pour une application sur réseau. En effet, l'aire totale sous la courbe devient une fonction du nombre de branches du carrefour comme illustré en Figure 4. L'intégrale de la fonction de densité vaut 1 pour les sections de routes sans carrefour mais descend jusqu'à 0,5 pour les voies sans issue et monte à 1,5 pour un carrefour à 3 routes. L'Équation (3) est donc modifiée par un facteur de correction basé sur la longueur l des segments sous-jacents pour assurer que l'aire sous la courbe reste à 1.

$$K^*(x_j) = \frac{1}{c(x_j)} \sum_{i \in V_j} k(x_i) n_i \quad (4)$$

avec

$$c(x_j) = \sum_{i \in V_j} l_i k(x_i)$$

Une comparaison des valeurs $K(x)$ et $K^*(x)$ est présentée en Figure 4 pour des simulations de réseaux avec 1 accident par point de mesure. Bien que les accidents soient uniformément distribués à une distance constante de 100 m, l'image fournie par l'estimateur à noyau $K(x)$ tend à exagérer la densité aux carrefours et à sous-estimer la densité sur les voies sans issue. Cette exagération de la densité aux carrefours est d'autant plus forte que le nombre de voies qui se croisent est élevé. À l'inverse, l'estimateur $K^*(x)$ corrigé par le facteur d'échelle se comporte plus conformément aux attentes et permet de mieux rendre compte de la distribution uniforme des accidents. Notons finalement que $c(x)$ correspond à la somme de Riemann ou aire sous la courbe estimée par la somme des aires des rectangles de longueur égale à la longueur des segments et de largeur égale à la valeur de densité à une distance d.

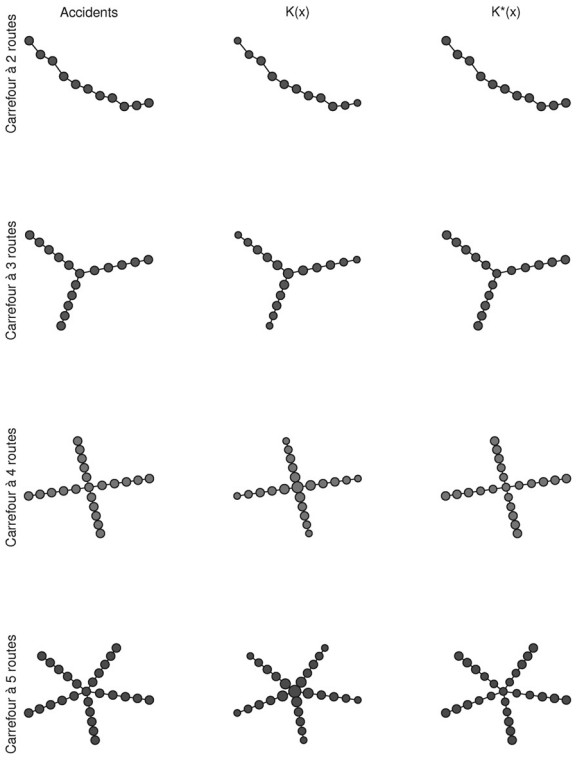


Figure 4. Fréquence des accidents, $K(x)$ et $K^*(x)$ sur base d'un noyau gaussien et une fenêtre de 1,5. Les accidents présentent une distribution uniforme et sont répartis de manière équidistante le long des réseaux simulés. Les bouts de branches du réseau représentent des voies sans issues. La surface du symbole est proportionnelle linéairement à la valeur représentée.

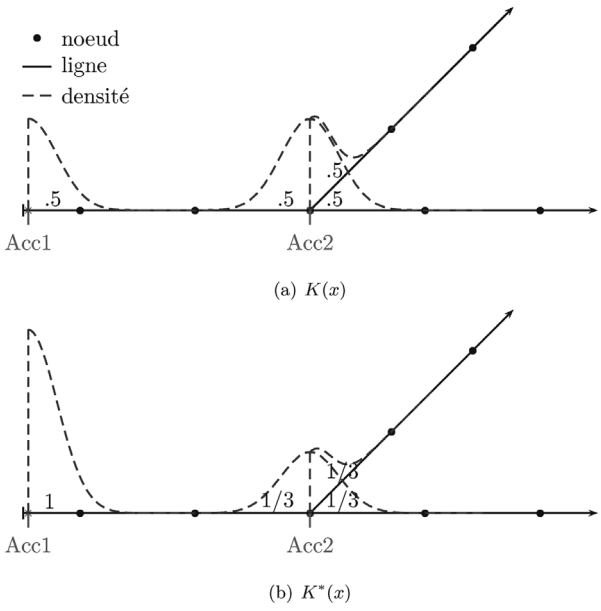


Figure 5. Noyau gaussien lors de l'application à un réseau avec carrefour et voie sans issue. Les aires sous la courbe de l'accident 1 sur la voie sans issue et de l'accident 2 au carrefour de 3 routes sont respectivement de 0,5 et 1,5 pour l'estimateur $K(x)$ alors que les valeurs sont de 1 dans les deux cas pour $K^*(x)$.

Cette correction pose toutefois problème. Okabe *et al*. (2009) démontrent en effet que les estimateurs à noyau sur réseau appartenant à la famille des *similar shape kernel function* dont fait partie $K^*(x)$, sont tous biaisés mais qu'ils présentent des caractéristiques intéressantes dont l'unimodalité, l'égalité de la densité à distance égale, l'isotropie et la symétrie par rapport au centre du noyau. Les alternatives proposées par Okabe *et al*. (2009) pour obtenir un estimateur non biaisé consistent en la définition de deux nouvelles fonctions: (i) l'*equal split kernel* et (ii) l'*equal split continuous kernel*.

L'*equal split kernel* voit sa densité divisée par le nombre de routes présentes à chaque carrefour. Il présente donc des discontinuités et des asymétries aux carrefours avec des sauts dans la densité. Comme nous ne disposons pas d'informations sur le risque d'accident sur chaque bras du carrefour ou encore du nombre de véhicules sur chacun des éléments du carrefour, nous n'avons aucune raison d'allouer des densités différentes à des points équidistants de l'accident. Une solution pourrait être trouvée en estimant l'importance du trafic selon la hiérarchie des routes et constitue certainement une évolution possible de notre méthode.

L'*equal split continuous kernel* est une version corrigée de l'*equal split kernel* pour atteindre la continuité. Cette correction conduit à la définition d'un estimateur dont le mode n'est pas nécessairement le point d'occurrence de l'accident, dont la forme n'est pas symétrique par rapport au site de l'accident et surtout dont l'implémentation est très lourde en termes de temps de calcul. Pour l'instant, cette solution n'a pas encore été appliquée de manière extensive sur des données et est difficilement implémentable techniquement à notre niveau pour l'entièreté du réseau numéroté belge. Cependant, elle mériterait certainement d'être investiguée plus avant dans des évolutions futures de notre méthodologie.

B. Choix de la fenêtre

Plusieurs éléments bibliographiques permettent d'identifier un intervalle de variation pour la fenêtre optimale. Pour des accidents de la route, Xie et Yan (2008) testent des valeurs entre 20 et 2.000 m mais n'avancent pas de choix optimal. Okabe *et al.* (2009) optent pour une valeur de 200 m sans aucune justification ni motivation pour ce choix. Finalement, Steenberghen *et al.* (2010) travaillent sur des données belges avec des valeurs de 25 à 500 m et concluent que tout dépend de l'application suivant qu'elle soit locale ou plus globale. L'utilisateur est donc laissé libre de choisir dans l'intervalle [25, 500] m selon l'échelle de travail.

Des éléments logiques tels que la distance d'arrêt sur route humide ou encore la distance utilisée en conception routière sont également informatifs pour déterminer la fenêtre optimale. La distance d'arrêt est fonction

non seulement de la vitesse du véhicule qui détermine la distance de freinage mais aussi du temps de réaction du conducteur avant d'actionner le système de freinage et de plusieurs autres variables telles que notamment les conditions de la chaussée, les réflexes et l'attention du conducteur ou encore la qualité et l'usure de la gomme des pneus. Plusieurs relations empiriques existent pour déterminer ces différentes distances, citons notamment les relations utilisées par le Sétra (Vertet et Giausserand, 2006) pour des routes planes

$$\begin{aligned} D_{\text{réaction}} &= 2\text{v} \\ D_{\text{freinage}} &= \frac{\text{v}^2}{2gf} \\ D_{\text{arrêt}} &= D_{\text{réaction}} + D_{\text{freinage}} \end{aligned} \quad (5)$$

où v est la vitesse exprimée en [m/s], g est l'accélération de la pesanteur et f le coefficient de frottement longitudinal de la route. Par convention en conception routière, une route humide peu adhérente avec des pneumatiques usagés est considérée (Vertet et Giausserand, 2006) et f prend alors une valeur comprise entre 0,31 et 0,46. Le Tableau 1 donne un aperçu de ces distances de freinage, des temps de réaction du conducteur et finalement de la distance d'arrêt selon différentes conditions de chaussée. Une distance d'arrêt de 300 m sur autoroute n'est donc pas totalement inconcevable si le temps de réaction du conducteur augmente à cause d'inattentions diverses telles que GPS et GSM, que la route est détrempée et en mauvaise état et que le couple freins/pneus du véhicule est en piètre état également.

Finalement, le variogramme constitue également une indication pour choisir la taille de la fenêtre à partir d'éléments empiriques. Pour rappel, le variogramme décrit la dépendance spatiale d'une variable par le biais d'une mesure de dissimilitude entre toutes les paires de points séparés par un intervalle de distance de 0 à l'infini. Dans notre cas, nous étudierons donc la différence du nombre d'accidents observés entre deux points de mesure en fonction de la distance. L'estimateur classique du semi-variogramme proposé par Matheron (1963) est défini par

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i,j \in N(h)} \left[Z(s_i) - Z(s_j) \right]^2$$

avec $N(h)$ le nombre de paires de points séparés par une distance h et $Z(s_i)$ la valeur de la variable Z mesurée au site s_i.

L'estimateur du semi-variogramme pour les AccRLC est donné en Figure 6. Avant tout, les points de mesure se trouvant à moins de 100 m l'un de l'autre sont, par définition, relativement rares pour des bornes hectométriques et induisent donc un saut entre 0 et 100 m sur l'axe des abscisses. Pour vérifier la robustesse du variogramme, nous avons estimé sa valeur pour des intervalles de distance de 50 à 150 m par pas de 25 m. A première vue, la variance entre deux sites de mesure augmente jusqu'à atteindre un palier aux alentours de 300 m. Notons également que l'effet pépite, c'est-à-dire la valeur du variogramme à une distance nulle, semble très important.

Qualité de la chaussée	f	70 km/h	90 km/h	120 km/h
Béton bitumeux propre et sec	0.8	63.0	89.8	137.5
Revêtement sec de qualité moyenne	0.7	66.4	95.5	147.6
Pavé sec	0.6	71.0	103.1	161.1
Béton bitumeux humide	0.4	87.1	129.6	208.2
Revêtement humide de qualité moyenne	0.35	93.9	141.0	228.5
Pavé humide	0.3	103.1	156.2	255.4

Tableau 1. Distance d'arrêt [m] en fonction de la vitesse pour un temps de réaction moyen de 2 s et des conditions variables de chaussée

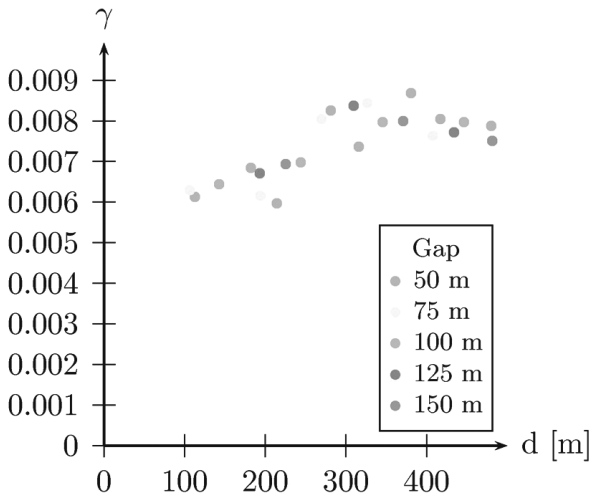


Figure 6. Variogramme empirique
De l'ensemble des éléments, qu'ils soient empiriques, bibliographiques ou bien théoriques, une distance de 300 m semble pertinente dans le cas de l'étude des AccRLC sur les routes numérotées.

C. Simulations de Monte-Carlo

Comme nous l'avons déjà avancé dans ce travail, les accidents de la circulation peuvent présenter un comportement stochastique. En outre, comme le soulèvent Xie et Yan (2008), Truong et Somenahalli (2011) et Chainey *et al.* (2008), le KDE présente le désavantage de ne pas estimer de significativité des valeurs calculées et par là même ne pas offrir de seuils naturels au-dessus desquels une valeur induit un *hotspots*. Pour vaincre cette faiblesse et tester la significativité des valeurs de K observées vis-à-vis d'une situation spatialement aléatoire (Complete Spatial Randomness, CSR), des simulations de Monte-Carlo (MC) sont utilisées. Ces simulations permettent de générer des situations dans lesquelles les accidents respectent l'hypothèse aléatoire spatiale sur le réseau. Il devient alors possible de confronter les valeurs observées de densité lissée avec les valeurs attendues sous l'hypothèse de distribution aléatoire des accidents. À chaque itération du processus :

- le nombre total d'accidents est réparti aléatoirement sur le maillage des points de mesure par un algorithme d'échantillonnage aléatoire avec remise, pondéré selon la longueur des segments;
- les valeurs $K^*(x)$ sont calculées pour tous les points de mesure et enregistrées.

Ensuite, la valeur observée est comparée à la distribution des simulations pour chaque point et fournit ainsi une p-valeur. Le choix du seuil de significativité est fonction premièrement de l'aversion au risque de l'analyste mais aussi d'un compromis entre d'autres paramètres dont le nombre de points significatifs désirés par les décideurs en fonction des moyens disponibles, du nombre d'accidents inclus dans ces éléments et de la longueur du réseau couverte par les points noirs.

D. Chaînage des points noirs

Le calcul de significativité via les simulations de Monte-Carlo permet d'identifier les points de mesure dont la valeur observée est significativement différente d'une distribution aléatoire des accidents. Ces éléments portent classiquement le nom de *hotspots*. Ils sont indépendants les uns des autres même si le calcul utilise un estimateur à noyau qui inclut le voisinage de chaque point. Via l'analyse de réseau, l'information topologique des données spatiales sert à unir les points significatifs contigus les uns aux autres en un seul et même objet que nous nommerons *hotzone*. Les points de mesures avec une valeur significative seront donc désormais dénommés *hotzones* s'ils sont contigus alors que les points isolés porteront le nom de *hotspots*.

Cette étape de chaînage des éléments utilise un algorithme de reconnaissance des composantes du réseau formé par l'ensemble des points de mesures avec une valeur significative. Les composantes de minimum 2 éléments significatifs définissent les *hotzones* alors que les éléments significatifs isolés sont les *hotspots*.

IV. RÉSULTATS

Dans cette analyse au niveau belge sur les routes numérotées, 79.182 AccRLC de 2006 à 2009 correctement géolocalisés sont utilisés. Cette période de 4 ans est conseillée par Elvik (2008) pour obtenir une image fiable des concentrations d'accidents. Après allocation au point le plus proche, le minimum observé par point de mesure est de 0 alors que le maximum est de 47 accidents avec lésions corporelles. La distribution du nombre d'accidents par point de mesure suit approximativement une distribution de Poisson avec une densité concentrée sur les effectifs faibles proches de 0.

Les résultats présentés en Figures 7 et 8 sont calculés avec un noyau gaussien de 300 m de fenêtre et un calcul de significativité des valeurs de K estimées sur base de 1.000 simulations de Monte-Carlo. Au niveau national, 10.768 points de mesure ressortent comme significatifs au seuil $\alpha < 0{,}001$ (voir Tableau 2). Selon nos définitions, ces points significatifs se répartissent en 1.315 *hotzones* et 2.172 points de mesures significatifs isolés. Les 1.315 *hotzones* sont donc composées de 8.596 points de mesures, couvrent 532 km du réseau routier et contiennent 21.810 accidents. En terme relatif, cela représente près de 27,54% des accidents pour seulement 3,18% du réseau routier des routes numé-

Signif.	Points significatifs	Hotzones						Hotspots				
				Longueur couverte		AccRLC			Longueur couverte		AccRLC	
[α]	[n]	[n]	[n points]	[km]	[% Total]	[n]	[% Total]	[n]	[km]	[% Total]	[n]	[% Total]
<0,001	10.768	1.315	8.596	532	3,18	21.810	27,54	2.172	134	0,80	8.336	10,53
0,001	12.680	1.500	10.328	645	3,86	24.581	31,04	2.352	148	0,89	8.587	10,84
0,002	13.862	1.641	11.359	713	4,26	26.168	33,05	2.503	168	0,96	8.818	11,14
0,003	14.731	1.742	12.140	765	4,58	27.296	34,47	2.591	167	1,00	8.877	11,21
0,004	15.486	1.811	12.801	810	4,85	28.235	35,66	2.685	175	1,05	8.989	11,35
0,005	16.143	1.887	13.338	846	5,06	28.977	36,60	2.805	184	1,10	9.194	11,61
0,006	16.659	1.948	13.774	877	5,25	29.587	37,37	2.885	190	1,13	9.279	11,72
0,007	17.185	1.998	14.248	911	5,45	30.218	38,16	2.937	194	1,16	9.342	11,80
0,008	17.637	2.042	14.636	938	5,61	30.721	38,80	3.001	200	1,19	9.399	11,87
0,009	18.072	2.091	15.021	964	5,77	31.222	39,43	3.051	205	1,22	9.401	11,87
0,010	18.452	2.148	15.369	988	5,91	31.687	40,02	3.083	208	1,24	9.382	11,85

Tableau 2. Seuil α de significativité et classification des points significatifs en hotzones et hotspots

rotées. A propos des *hotspots*, ce sont 8.336 AccRLC (soit 10,53%) qui sont concernés et 134 km (soit 0,8%) du réseau routier numéroté. Le Tableau 2 décrit l'évolution de ces différents paramètres pour des valeurs de significativité α comprises entre 0 et 0,01. Conformément aux attentes, ces chiffres illustrent bien la concentration spatiale des AccRLC. Un effet Pareto apparaît même pour les *hotzones* avec 3,18% du réseau routier qui concentrent 27,54% des accidents. Ensuite, le nombre de zones noires calculées est en relation directe avec le seuil de significativité sur l'intervalle [0-0,01]. Plus celui-ci augmente, plus le nombre de *hotspots* et de *hotzones* est important.

L'interprétation globale de la Figure 7 est conforme aux grandes tendances avancées par l'Institut Belge pour la Sécurité Routière (IBSR) (Casteels *et al*., 2010) avec:

- des concentrations (*hotzones* et *hostpots*) plus nombreuses à l'entrée des grandes agglomérations belges et dans les zones urbanisées qu'en rase campagne;
- une différence nette entre Wallonie et Flandre traduisant le risque supérieur d'accident de la Flandre bien que la gravité soit supérieure en Wallonie;
- une majorité des *hotzones* et *hostpots* prennent place sur les routes nationales et régionales;

À l'échelle locale de la Figure 8, un rapport est rédigé pour chaque *hotzone* et *hotspot* avec une analyse détaillée des données des accidents. Une fiche descriptive est ainsi produite avec le nombre de victimes par catégorie (tués, blessés graves, blessés légers), l'âge, le sexe, et le type de chacun des usagers impliqués, la longueur de la zone, les conditions particulières observées lors des accidents (lumières, qualités de la chaussée, etc.) et une carte de l'environnement immédiat. Ces fiches constituent l'aboutissement de notre méthodologie et se présentent alors comme un outil pratique pour le choix et la mise en oeuvre de mesures correctrices.

V. CONCLUSIONS

La sécurité routière est un des points majeurs d'attention en Belgique tout comme en Europe. Les moyens budgétaires limités actuels imposent des choix de la part des décideurs et des gestionnaires du réseau. Il convient d'identifier les portions du réseau routier les plus dangereuses et problématiques nommées communément points noirs. Cependant, la notion même de point noir reste sujette à de vives discussions entre les experts du domaine bien que des définitions fonctionnelles existent dans plusieurs pays (Elvik, 2008).

Plusieurs outils utilisés par les autorités en Belgique négligent plusieurs dimensions importantes des données d'accidents de la route, notamment l'aspect réseau dont elles proviennent, l'aspect stochastique des accidents, les éventuelles erreurs de localisation de ceux-ci, et rendent impossible la définition de zones noires d'accidents de la route. Des évolutions et de nouvelles méthodes apparaissent régulièrement dans la littérature mais aucune n'a été appliquée, à notre connaissance, à l'échelle d'un pays pour une fenêtre temporelle de 4 ans. Ainsi, Elvik (2008) propose de travailler avec des modèles bayésiens et décourage l'utilisation des fenêtres mobiles mais n'applique nullement ses idées et concepts à un cas pratique. Flahaut *et al*. (2003) et Manepalli *et al*. (2011) travaillent avec des mesures

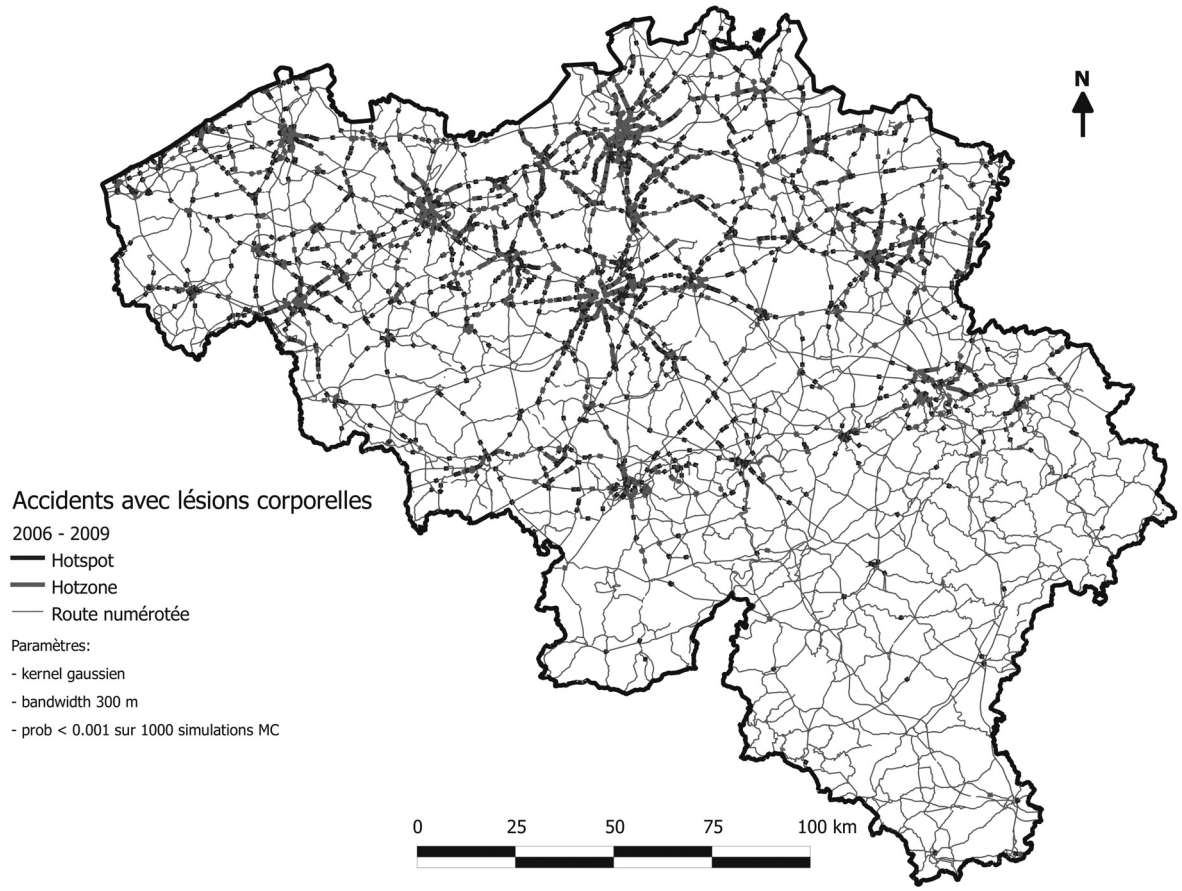


Figure 7. Aperçu global de la distribution des *hotspots* et *hotzones* sur les routes numérotées belges

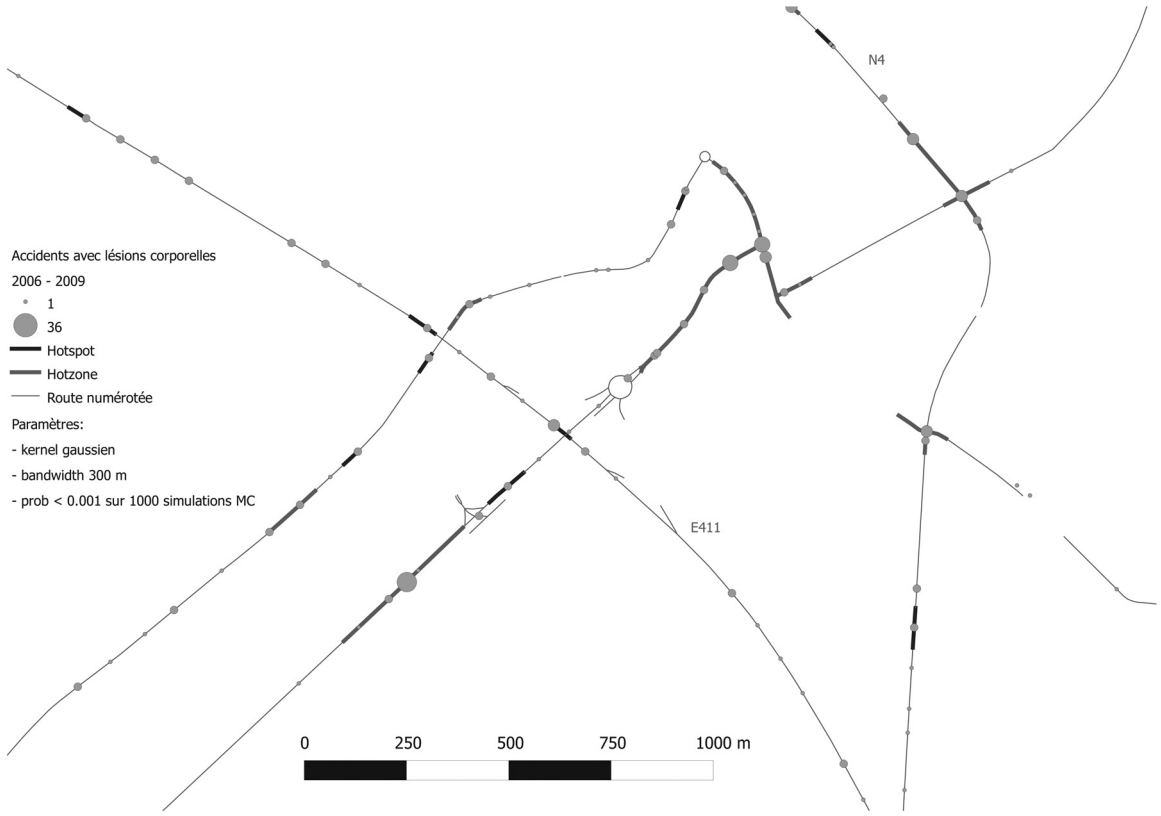


Figure 8. Détails de la distribution des *hotspots* et *hotzones* aux alentours du croisement de la E411 et la N4 sur la commune de Wavre

d'autocorrélation spatiale mais au niveau d'une seule route. Moons *et al*. (2009) étudient une province belge mais uniquement au niveau de 2 autoroutes. Finalement, Okabe *et al*. (2009), Xie et Yan (2008) et Steenberghen *et al*. (2010) analysent une seule ville. Ici, notre attention se porte sur le développement d'une solution permettant de considérer l'entièreté du réseau routier principal d'un pays. Pour plusieurs raisons motivées dans ce travail, nous proposons une méthodologie mixte en trois étapes: (i) la fonction de densité des événements ponctuels sur le réseau est évaluée par un KDE modifié pour tenir compte de l'aspect réseau des accidents de la route dont le choix des paramètres est justifié tant empiriquement que théoriquement, (ii) la significativité des valeurs lissées observées est calculée par des simulations MC et (iii) des outils d'analyse de réseaux permettent de classifier les éléments significatifs en *hotspots* et les *hotzones*.

La méthodologie est ensuite testée sur l'entièreté des routes numérotées belges pour la période 2006-2009. Les résultats se montrent conformes aux constatations globales des autres partenaires actifs dans le domaine de la sécurité routière en Belgique. La méthode est également opérationnalisée par la mise en place de fiches descriptives pour chaque point et zone noire.

Naturellement, plusieurs évolutions de la méthode sont possibles dont l'intégration d'information du trafic routier ou encore la considération de nouvelles fonctions par le KDE. De même, la considération d'une seconde période d'accidents (2010-2013) permettrait d'évaluer la stabilité des résultats et le pouvoir prédictif des *hotspots* et *hotzones* passés pour les configurations futures d'accidents de la route.

BIBLIOGRAPHIE

Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41 (3): 359–364.

Antoine, D. (2010). *Zones à risque et tronçons dangereux 2005 - 2009*. Rapport technique. Namur : Service Public de Wallonie.

Bailey, T. C. et Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Essex : Longman.

Casteels, Y., Martensen, H., Merckx, F., Nuyttens, N., Riguelle, F. et Thijs, R. (2010). *Satistiques de sécurité routière 2008*. Bruxelles : IBSR, Observatoire pour la sécurité routière.

Castellà, J. et Pérez, J. (2004). Sensitivity to punishment and sensitivity to reward and traffic violations. *Accident Analysis & Prevention*, 36(6): 947–952.

Chainey, S., Tompson, L. et Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21: 4–28.

Elvik, R. (2008). A survey of operational definitions of hazardous road locations in some european countries. *Accident Analysis & Prevention*, 40: 1830–1835.

Flahaut, B., Mouchart, M., Martin, E. S. et Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones a comparative approach. *Accident Analysis & Prevention*, 35: 991–1004.

Geurts, K. et Wets, G. (2003). *Black spot analysis methods: literature review.* Diepenbeek : Steunpunt Verkeersveiligheid bij stijgende mobiliteit.

Iversen, H. et Rundmo, T. (2002). Personality, risky driving and accident involvment among norwegian drivers. *Personality and Individual Differences*, 33: 1251–1263.

Manepalli, U. R. R., Bham, G. H. et Kandada, S. (2011). Evaluation of hotspots identification using kernel density estimation (k) and getis-ord (gi*) on i-360. *3rd International Conference on Road Safety and Simulation, Indianapolis, 14-16 September 2011.*

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58: 1246–1266.

Moons, E., Brijs, T. et Wets, G. (2009). Improving moran's index to identify hot spots in traffic safety. In Murgante, B., Borruso, G. et Lapucci, A., editors, *Geocomputation and Urban Planning*, volume 176 of *Studies in Computational Intelligence* (p. 117–132). Berlin - Heidelberg : Springer.

Okabe, A., Satoh, T. et Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a gis-based tool. *International Journal of Geographical Information Science*, 23 (1): 7–32.

O'Sullivan, D. et Unwin, D. J. (2002). *Geographic Information Analysis*. New Jersey : John Wiley.

O' Sullivan, D. et Wong, D. W. S. (2007). A Surface-Based Approach to Measuring Spatial Segregation. *Geographic Analysis*, 39 (2): 147–168.

Romano, U. (1997). *Atlas de l'insécurité routière 1990-1994*. Namur : Ministère Wallon de l'Equipement et des Transports (MET).

Romano, U. et Heuchenne, D. (1996). *Modèle mathématique d'évaluation de l'insécurité routière*. Namur : Ministère Wallon de l'Equipement et des Transports (MET).

Schabenberger, O. et Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton : Chapman & Hall/CRC.

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

Smith, S. C. et Bruce, C. W. (2008). *CrimeStat III Userbook*. Washington: The National Institute of Justice.

Steenberghen, T., Aerts, K. et Thomas, I. (2010). Spatial clustering of events on a network. *Journal of transport geography*, 18: 411–418.

Truong, L. T. et Somenahalli, S. V. C. (2011). Using gis to identify pedestrian-vehicle crash hot spots and unsafe bus stops. *Journal of Public Transportation*, 14(1): 99–114.

Vertet, M. et Giausserand, S. (2006). *Comprendre les principaux paramètres de conception géométrique des routes*. Rapport Sétra. Bagneux : République Française - Ministère des Transports, de l'Equipements, du Tourisme et de la Mer - Service d'Etudes techniques des routes et autoroutes (Sétra). Consultable sur http://www.setra.equipement.gouv.fr.

Xie, Z. et Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32 (5): 396–406.

Coordonnées des auteurs :
David DABIN, Christiane DICKENS
& Paul WOUTERS,
Analystes-stratégiques, Police Fédérale
- Direction de l'Information Policière Opérationnelle
- Service d'Analyse Stratégique.
Rue Fritz Toussaint 8, 1080 Ixelles,
david.dabin.5968@police.be.

BSGLg, 60, 2013, 33-43

SPATIOTEMPORAL ANALYSIS OF FORENSIC CASE DATA: A VISUALISATION APPROACH

Quentin ROSSY

Abstract
Whether for investigative or intelligence aims, crime analysts often face up the necessity to analyse the spatiotemporal distribution of crimes or traces left by suspects. This article presents a visualisation methodology supporting recurrent practical analytical tasks such as the detection of crime series or the analysis of traces left by digital devices like mobile phone or GPS devices. The proposed approach has led to the development of a dedicated tool that has proven its effectiveness in real inquiries and intelligence practices. It supports a more fluent visual analysis of the collected data and may provide critical clues to support police operations as exemplified by the presented case studies.

Keywords
crime analysis, forensic intelligence, spatiotemporal visualisation, visual data analysis.

Résumé
Que cela soit à des fins d'enquête ou de renseignement, les analystes criminels sont souvent confrontés à la nécessité d'analyser la répartition spatio-temporelle des crimes ou des traces laissées par des suspects. Cet article présente une méthode de visualisation soutenant des analyses récurrentes telles que la détection de séries ou l'analyse des traces laissées par des appareils numériques tels que des téléphones mobiles ou des GPS. L'approche proposée a conduit à l'élaboration d'un outil dédié qui a prouvé son efficacité dans de véritables enquêtes et à des fins de renseignement. Il permet une analyse visuelle et dynamique des données recueillies, facilitant ainsi la production de renseignements utiles à la définition d'opérations de police, comme le montrent les études de cas présentées.

Mots-clés:
analyse criminelle, renseignement forensique, visualisation spatiotemporelle, analyse visuelle de données

1. INTRODUCTION

Visualisation is a pillar of crime intelligence. Link diagrams between relevant entities (persons, objects), maps, quantitative representations (e.g. histograms) or timelines are always more frequently used in this context, and chosen depending on the problem to be analysed. Combined visualisations of those perspectives are challenging, and very few methodological support and computerised tools are available for this purpose.

The importance to develop frameworks and tools is particularly evident, when realising that spatiotemporal information are at the core of the study of crime: criminal behaviour, more often than not, follows patterns, that crime analysis tries to discern from collected data. Crime mapping techniques are now well established for representing the distribution of crimes, detecting crime concentrations or simply displaying set of events to be interpreted. Chronologies of events are represented generally on separate visualisations such as event charts or flow diagrams. When considering that crime occurs within highly specific situations, at a certain time, when the immediate social and physical environment offers opportunities, it clearly appears that space and time dimensions are closely related. Thus, visual possibilities of combining both perspectives in representing information are critical.

Examples where the spatiotemporal dynamic underlying specific problems must be analysed are manifold. For instance, hypothesis developed in the course of an investigation are frequently tested through the study of victim's and suspect's journey. This involves always more routinely the analysis of digital traces such as GSM or GPS records, supported by spatiotemporal visualisations. Such data also frequently help to link a

suspect activity with known offenses (Birrer and Terrettaz-Zufferey, 2008) or to locate him.

Other frequent analytical tasks concern the detection and understanding of crime repetitions. They support the development of investigative hypotheses (e.g. geographical profiling) and, occasionally, the development of the series can be predicted through the pattern detected, allowing most relevant measures, preventive and repressive, to be taken. The study of when certain types of premises located at certain places are repeatedly victimised also form the basis for strategic intelligence products.

Spatiotemporal visualisation techniques are available for dealing with such a variety of complex situations. However, a lot of difficulties have to be faced, already when time and space dimensions are represented separately: how to represent time imprecision and uncertainties (when did the crime occur?), how to deal with the overlapping of symbols representing different events, how to avoid overwhelming the reader with quantities of symbols, how to choose appropriate visual forms and levels of aggregations in order to avoid biasing the judgement of the reader? Combining spatiotemporal perspectives amplify those difficulties and make the representation sometimes intractable for its user.

The first part of the paper summarizes general concepts about multidimensional visualisation, and adapted to crime analysis. Within this framework, several elements from previous researches have been selected, focusing on static two-dimensional representations that manage spatial overlapping. On this basis, a dedicated methodology and tool will be presented. They combine spatiotemporal representation of set of relevant traces coming from series of events. The solution traditionally used to deal with this problem suggests superimposing a temporal dimension onto the spatial representation (i.e. a map). An effective alternative consists of proceeding the other way round, by dividing the temporal views with an appropriate use of colours corresponding to defined geographical areas shown on a map. On this basis a computerised prototype has been developed. It has demonstrated to be an effective tool for analysing crime data. It was used in a variety of situations for analysing actual crimes series, as well as for displaying several types of numerical traces resulting from the use of electronic devices (GPS, GSM) by offenders during their journeys. We illustrate the approach with two recurrent crime analysis tasks: mobile phone records analysis and crime series analysis.

II. MULTIDIMENSIONAL ANALYSIS AND VISUALISATION

Spatiotemporal analysis covers several recurrent crime analysis tasks, such as understanding past series of events, and predicting future occurrences, by clustering and pattern discovery (Boba, 2009; Helms, 2009; Laxman and Sastry, 2006). If the detection of temporal and geographic patterns of crime occurrences is important for intelligence purpose, more specific questions arise when specific crime activity or a set of traces are analysed for investigative purposes (for example left by an electronic device like a GSM or a GPS). Such recurrent questions are: where was a particular person on a defined time frame? Can we infer the home location of an offender from related crime events? Can we link GPS or GSM data collected from a suspect to known crime events? Moreover, many questions may also involve other information than time and space for both intelligence and investigative purposes: what are the relationships between crimes forming a particular repetition? What are the phone numbers in contact with a particular offender? How many crimes have been perpetrated by a specific group of offenders?

Approaching such a variety of questions, and imagining how to visualise a situation, requires an adequate methodology. It can be based on the grouping of typical situations in four dimensions: temporal, spatial, quantitative and relational. This multiple dimension approach follows a similar paradigm of the multidimensional data cube used in data warehouses and OLAP systems (Kimball and Ross, 2002). The classification of a specific problem by identifying its main dimensions, al-

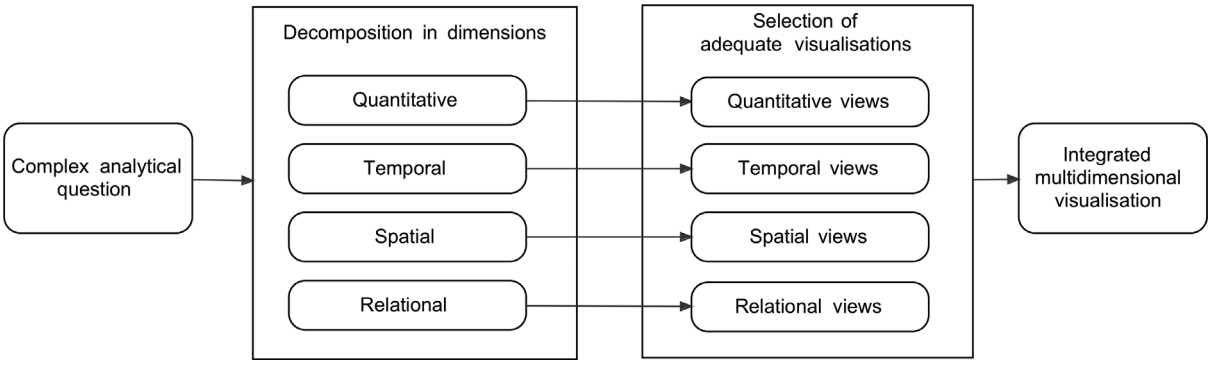


Figure 1. Visualisation selection process based on a multidimensional decomposition of crime analysis questions

lows a most appropriate and effective visual form to be chosen. Indeed for each of this dimensions, a variety of visualisations are now routinely used, but fundamental difficulties remains and will be explained.

The temporal dimension covers questions where time is the main component: when? On what period of time? How often? Is there a temporal pattern? A sequence? Unlike other quantitative variable, the temporal dimension has a complex semantic structure. Indeed, time has a hierarchical structure and many possible aggregation levels with varying divisions: sixty minutes, twenty-four hours, seven days of the week, twelve months of the year, etc. Furthermore time is analysed linearly and by cycles that may be regular (e.g. day of weeks) or irregular like holydays (Aigner et al., 2007). This intrinsic complexity requires the usage of multiple and dedicated visual forms in order to detect temporal patterns. They consist mainly of timelines and cyclic views.

Moreover, analysing the temporal distribution of crime events requires dealing with uncertainties. Indeed, many crime data are stamped by a time period, which is often not directly related to the duration of the event. It rather results from a lack of knowledge about when it precisely occurred. For instance, the temporal imprecision of burglaries is generally defined by when the victims left their premises. The timeframe is bounded by the period of absence at the location: during the night for shops and industries or daytime for apartments for instance. Several approaches are used to handle this imprecision that may affect distribution analysis (e.g. when a set of crimes tend to occur) or defining queries when searching a database. The simplest way to deal with such time intervals is to arbitrary chose the starting date/time or ending date/time or to use more elaborated approaches like mean calculation or an approach called *aoristic* (for details about this method see (Ratcliffe, 2000)). Temporal uncertainties have also to be handled in visualisations. Similarly, one immediate solution, consist of displaying an event on a timeline at an arbitrary defined date/time (e.g. starting, ending or mean date and time). A second approach is to use a box to represent the period, like with popular Gantt charts. In temporal distributions views, the aoristic approach can be used.

Temporal overlapping is another difficulty. Due to temporal imprecision or to the amount of data, time intervals associated to events can overlap. Specific strategies are used to avoid or manage this problem. As the time dimension is often depicted with the horizontal axis of the plan, the vertical axis is commonly used to distinguish overlapping events. For example, if two datasets of telephone calls have to be compared, they can be plotted in parallels, one on top of the other (e.g. in parallel plots or stacked views, see figure 3 for an example). When the vertical axis cannot be used, a dedicated visual property like the colour or transparency of the symbols can be used (e.g. lines charts sharing the same vertical scale).

The spatial dimension covers questions where space is the main component of the question: Where did the crime occur? In what area? Which path was followed? The interest in dealing with the spatial dimension of crime resides in the not-random nature of crime occurrence, even if no consensus is reached on how to explain it (Canter, 2000). The first maps of crimes are attributed to the works of both André-Michel Guerry (1833) and Adolphe Quetelet (1842) (Friendly, 2008). The creation of these maps is connected to police reforms made at the time, when more structured processes for criminal data gathering and recording were developed. Since that time and with the development of computerisation, maps of crime have been progressively more systematically and widely used to detect and follow crime activities in an intelligence-led perspective (Anselin et al., 2000; Boba, 2009; Chainey and Ratcliffe, 2005).

Spatial data also suffer from uncertainty. Often, the exact location of the event is known (for example, in GPS data analysis or for crime events analysis from which the location is generally known), but some datasets may contain inaccurate or imprecise spatial information. For instance, the location of a particular mobile phone connected to a cell is often defined by the area the cell covers. This area can be small (micro-antenna in buildings) or wide (rural antenna). This impreciseness cause many visualisation problems. In particular, it may result in spatial overlapping. When events overlap in space, they cannot be visually distinguished (i.e. variations of symbols or colours are inefficient). One solution is to use spatial aggregates (one symbol sized by the number of occurrences) or small multiples (see below).

Other visualisation problems arise when location is encoded at various level of accuracy. For example, it is not possible to produce a density estimation map (i.e. a *hotspot* map) with a dataset that is geocoded at varying levels of accuracy such as an address, a street, an area or a city. The same problem occurs with mobile phone data since the accuracy of the location is variable in regards of the type of the cell. If a choropleth map can by used to standardize levels of accuracy they may lead to the well known ecological fallacy. Graduated symbols map is than the only remaining visual forms that can be used.

The *quantitative dimension* deals with recurring questions in crime analysis containing *how many of*…. Obviously, crime analysis benefits from visual forms that have been designed in history to cover quantitative analysis. William Playfair (1759-1823) is considered as the inventor of many of them e.g. line and cyclic graphs,

histograms, etc. (Friendly, 2008; Playfair et al., 2005). Quantitative visualisation techniques have been widely study as evidenced by the encyclopaedic list described in Harris (2000). However, the seminal work of Edward Tufte (Tufte, 2001) and the fundamentals provided by Jacques Bertin (Bertin, 2005, first edition in 1967) have significantly contributed to theorise and structure their modern use. Stephen Few add some useful distinctions by defining quantitative analysis as the study of relationships between values. Consequently, dedicated visualisations can be classified in regards of them: part-to-whole and rankings, deviations, distributions, correlations and multi-valuated patterns (Few, 2009).

Finally, *the relational dimension* deals with the most elementary analytical task: identifying relevant entities (e.g. events, persons, objects, traces) and their relationships. Specific visualisation methods are used, such as graphs, trees, diagrams or flow charts. These graph-like techniques are particularly useful for representing criminal networks, smuggling of goods, links between events, as well as telephone records and financial data. In this context, visualisations are used along many objectives, such as analysing traces and information gathered, evaluating a cold-case, helping along the categorization of a particular offense, facilitating the transmission and receipt of a case or supporting an argument at trial (Rossy and Ribaux, 2012). The visualisation of the relational dimension faces also many fundamental difficulties, but we will mostly remain focus here on spatiotemporal and quantitative dimensions.

A variety of problems faced in crime analysis can be handled along one of the four dimensions. However, this is a very strong limitation, as the study of real cases often necessitates the combined analysis of several dimensions. This is particularly true for spatiotemporal analysis. Using separate visualisation can thus necessitate jumping from a static map to a temporal view or vice-versa. This inevitably causes damageable ruptures in the reasoning process. Thus, supporting the interpretation of spatiotemporal information by displaying both variables on single charts, or by providing links between perspectives when a computerised system is available, are crucial for crime intelligence and the analysis of forensic case data in particular.

III. COMBINED SPATIOTEMPORAL VISUALISATION APPROACHES

The challenge of spatiotemporal visualisation is to produce representations of data that allow the exploration, analysis and communication of information in both dimensions at a glance. Several techniques and tools have been developed to support multidimensional visualisation (Andrienko et al., 2003; Brunsdon et al., 2007; Guo et al. 2006; Ratcliffe 2004). However, as Buetow et al. (2003) note there are still few techniques that let examine a single dataset from multiple perspectives. They propose a multiple views tool made of a timeline, a periodic data visualisation and a map (Buetow et al., 2003). Beyond this kind of work, there is still a clear need to search for the most effective way of combining representations in function of the situations to be visualised. Indeed, recent approaches proposed to visualise spatiotemporal datasets are based on 3D visualisations, in particular with space-time cubes (for examples, see: Wolf and Asche, 2009; Nakaya and Yano, 2010). Other studies focus on the display of crime displacements and journey to crime (for a recent discussion upon theses techniques, see: Wheeler, 2013).

Our proposal starts by limiting the focus on two-dimensional and unanimated representations of data. Indeed, many of intelligence products are delivered through static and two-dimensional supports. Although 3D visualisations offer numerous opportunities to represent multivariate data, they bring additional challenges on their own (Card et al., 1999). For example, the data exploration and analysis made with a three dimensional visualisation needs a dynamic environment (Lodha and Verma, 2000) and information is often hidden by the projection in the planar space (occlusion problem). The overall dataset cannot be seen at once with a 3D visual abstraction (MacEachren, 2004). Modern systems also integrate facilities for animating spatiotemporal visualisations (Brunsdon et al., 2007). Animation is intuitive and can associate a proper time with the data event. But the human ability to remember and process the pertinent aspects in case of long and complex animations is a potential problem. In addition, animation has to be interactively controlled and needs a particular support to be communicated. In crime analysis, most of the products have to be static, mainly because they have to be joined to a written report. This is the main reason why we will not integrate animations at this stage. Even if the proposed visual approach was designed to produce static and two-dimensional supports, the analytical process requires a dynamic environment to go along with reasoning performed by analysts. Thus the developed tool allows the dynamic design of static end products. Having limited the scope to static and 2D spatiotemporal representations, several ideas have been brought together by discussing weaknesses and strengths from the practice of existing approaches.

A. Two-dimensional and static visualisation of spatiotemporal data

One usual way to visualize spatiotemporal data is to integrate the time dimension in the spatial view. For examples, grey scale encoding of time can be applied for each point in the map or arrows can be used to show movement. The main drawback of these approaches is of course the spatial overlapping of points. In two-di-

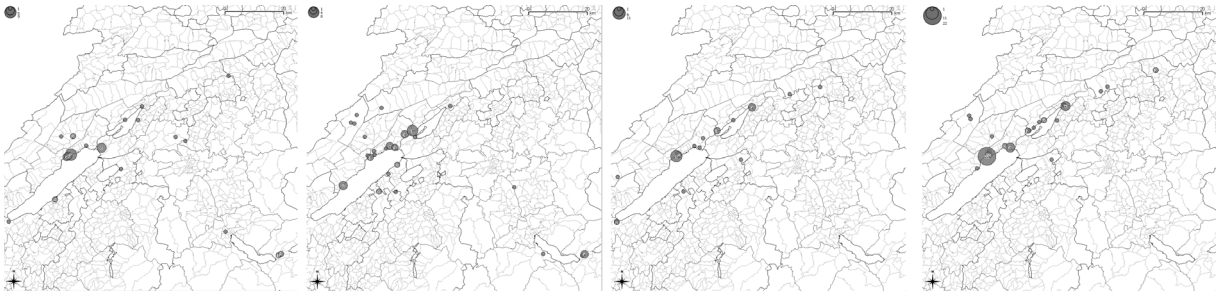


Figure 2. Comaps representing the spatial distribution all the communications made with a mobile phone during four consecutive months. All communication made during each month are aggregated on a dedicated map.

mensional static visualisations, two solutions remain to manage the problem of geographical overlapping: map iterations and linked plots (Andrienko et al., 2003). Both approaches have been used in our developments.

B. The map iteration approach (comap, small multiples)

A *comap* is the juxtaposition of several maps (see figure 2) where each iteration represents a subset of the data, for instance several time frames (Keim et al., 2005). Tufte (2001) defines this concept as *small multiples* of diagrams. He notes that the information slices have to be positioned within the eye-span so that the viewer can make comparison at a glance. According to Tufte comaps are the best representation solution for a wide range of comparison problems. The map iteration approach can allow emphasizing or revealing patterns and multivariate interactions from a period of interest. If the number of iterations increases, the resulting visualisation can be wide. Then, the comparison process can be time consuming and complex. Some scientists of Pennsylvania University reject the use of small multiples to explore problems of multivariate analysis because they judge the comparison too difficult and imprecise (MacEachren, 2004). In our approach comaps are used for the time views (see below), but was not adopted to separate the spatial view, which shows the complete geographic distribution on a single map.

C. Linked plots - focusing, linking and arranging views

The linked plots approach used multiples views to explore a dataset. To be integrated into a coherent visualisation, all the views have to be linked with each other. Each view depicts one or two-dimensional representation like maps, scatter plots, timelines or cyclic views. The way views are linked depends on whether they are displayed in sequence over time, or simultaneously in parallel (Fredrikson et al., 1999). The main method for linking parallel views is to use same colours in the different representations to encode a particular attribute of the dataset (Chen and Yu, 2000). Linked plots can show patterns from the whole time period and allow a great degree of interaction – for a review of software packages see (Brunsdon et al., 2007). The key idea of linked plots is to create multiples perspectives on data rather than try to find a single optimal view (MacEachren, 2004).

One of the main difficulties that occur when trying to link several views to produce a coherent representation of data is the choice of the linking parameter. One common way for linking views is by using a selection process also called *brushing* (Keim et al., 2005). For instance, the user selects a particular time frame within the timeline and the others views are updated and show only the selected items. Similar operations are performed by selecting a particular zone on the spatial views. The main drawback of these approaches is the loss of the overall view. Such global outlook is required to answers questions like *what is the time distribution of the data for each region of interest?* or *what is the spatial distribution for each period?* One solution is to highlight selected items in the other views, which keep the overview of the whole dataset in the view. Another approach is to use colours to represent a specific attribute. The defined colour is then transposed in each view (for an example see (Guo et al., 2006). Linking subviews with colours is generally done on a categorical attribute (like crime types, etc.). Such approach is not efficient for spatiotemporal analysis because of both spatial and temporal overlapping.

IV. PROPOSED APPROACH: GEOGRAPHICALLY LINKED PLOTS

The proposed visualisation rests on linked-plots, separating temporal views and maps, with colours on temporal views pointing to geographical area. This is a key aspect, as traditionally, time is integrated onto maps in the other way round. The dataset is divided into geographically separated groups that are assigned the dedicated colour used on the temporal views.

The colours (reprint in grayscale) in temporal views

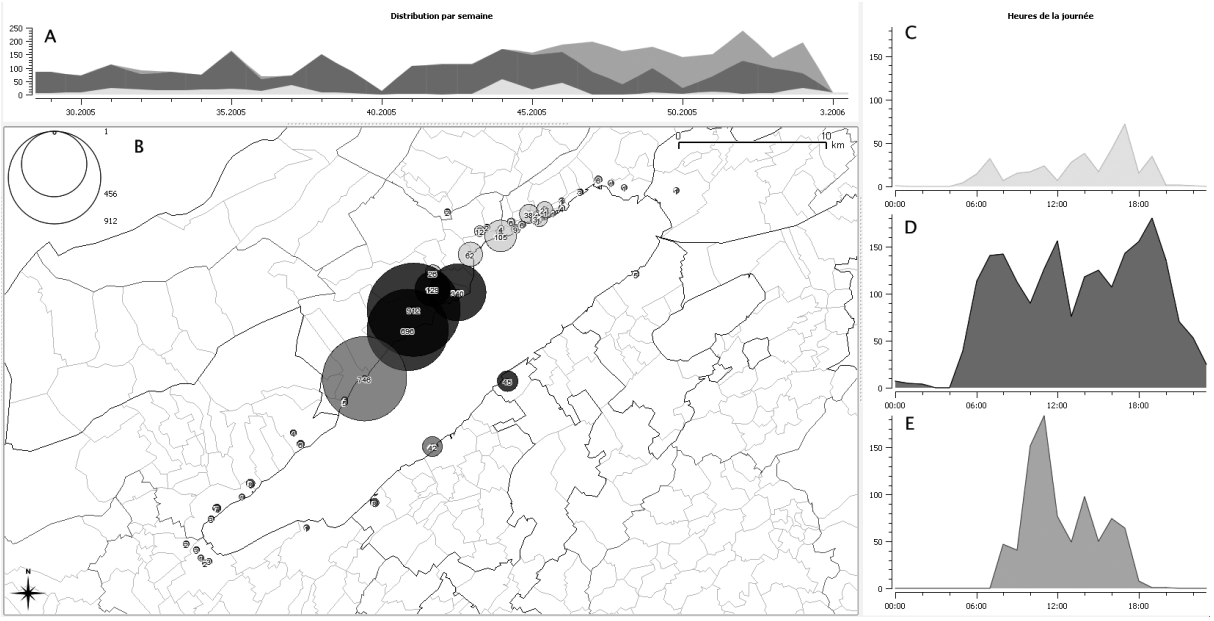


Figure 3. Geographically linked plots: a colour/grey level is assigned to spatially defined subsets of the data and applied in the temporal views (timeline on top (A) and hour of the day area charts on the right (C, D, E))

thus depict specific geographic regions. In the method, spatial groups can be either arbitrary defined or by using parametric clustering algorithm (K-mean clustering on both geographic dimensions). Temporal views integrate both linear and cyclic structures at any level of aggregation (e.g. hours of day, days of week, month of year, etc.). Temporal uncertainty is handled by the possibility to choose both start or end date/time or the aoristic calculation that is implemented as suggested by Rattclife (2000). Temporal distributions can be visualised on a single view or on separated small-multiples. Several renders have been integrated such as line graph, histogram or area graph.

Based on these basic principles, a computerised system has been implemented. It is thus hoped to support spatiotemporal inferences with the greatest fluidity. It has been developed as a python plugin of Quantum GIS 1.8 (http://www.qgis.org, last access 05.02.2013) and is available online with installation and usage instructions on http://www.analysecriminelle.org/visualist/ (last access 05.02.2013). It allows dynamic updating of views, at different levels of aggregation, automatic calculations of geographical clusters. Many other facilities have been implemented, but their descriptions fall beyond the scope of this paper. All Figures presented in this article are screenshots of the developed tool.

V. EFFECTIVENESS OF THE PROPOSED SOLUTION TO REPRESENT RECURRENT SPATIOTEMPORAL ANALYSIS PROBLEMS

It is assumed that the suggested dynamic methodology supports a more fluent visual spatiotemporal exploration of data and supports several forms of reasoning processes during the analysis of traces for both investigative and intelligence purposes. This has still to be demonstrated. In this section, the system is tested on two recurrent forms of spatiotemporal analysis and exemplified by real cases: the analysis of mobile phone billing records and the detection crime series by the combination of traces and spatiotemporal pattern detection.

A. Mobile phone billing records analysis

One of the most common spatiotemporal dataset in crime investigation (except crime itself) is telephone-billing record. A telephone call is indeed a particularly interesting item to analyse in all dimensions. Each call is composed of a time description (date, time and duration), the spatial position of the cell mast and by definition it describes a particular relationship between two phones. Reasoning with telephone calls data requires the combined use of visual abstractions along all these crime analysis dimensions.

The example presented above (Figure 3) represents all the communications made by a mobile phone during six months. It illustrates the effectiveness of the proposed approach to explore the data by emphasizing interesting patterns useful to support investigation. One recurrent useful inference drawn during an inquiry concerns the location of a suspect. The spatiotemporal analysis of mobile phone data may support this reasoning for instance by selecting calls at specific time periods (e.g. early in the morning) and plotting their spatial positions. Figure 3 shows another approach that consists first of selecting particular spatial areas. The temporal distribu-

tion of the black region then reveals a common pattern: more activity early in the morning, at noon and in the evening (see graphic D on figure 3). Moreover, the timeline view reveals that calls are done during the whole period (graphic A). Such pattern allows developing hypothesis about the home location of the suspect.

The multidimensional views may also give other insights. For instance the light grey and grey regions traces daytime activities (graphic C and E) and the timeline (graphic A) reveals several spatiotemporal changes, which indicate ruptures in the use of the cellphone, and, in turn, in the occupational activities of its user. The detection of such patterns through this visualisation process allow analytical hypothesis to be developed. The actual explanation was that the (single) user of the phone was living in the black region and was employed by a society located in the light grey region (which contains the main town of the region). After a period of holydays he was transferred to another branch office (in the grey region). It might be that these patterns would have been detected by other forms of visualisation, data mining technologies or even by sorting a spread-sheet. But the combined spatiotemporal visualisation allows detecting the underlying patterns quickly, dynamically and without complex spatiotemporal modelling knowledge. The method goes fluently along with crime analysis inference structures.

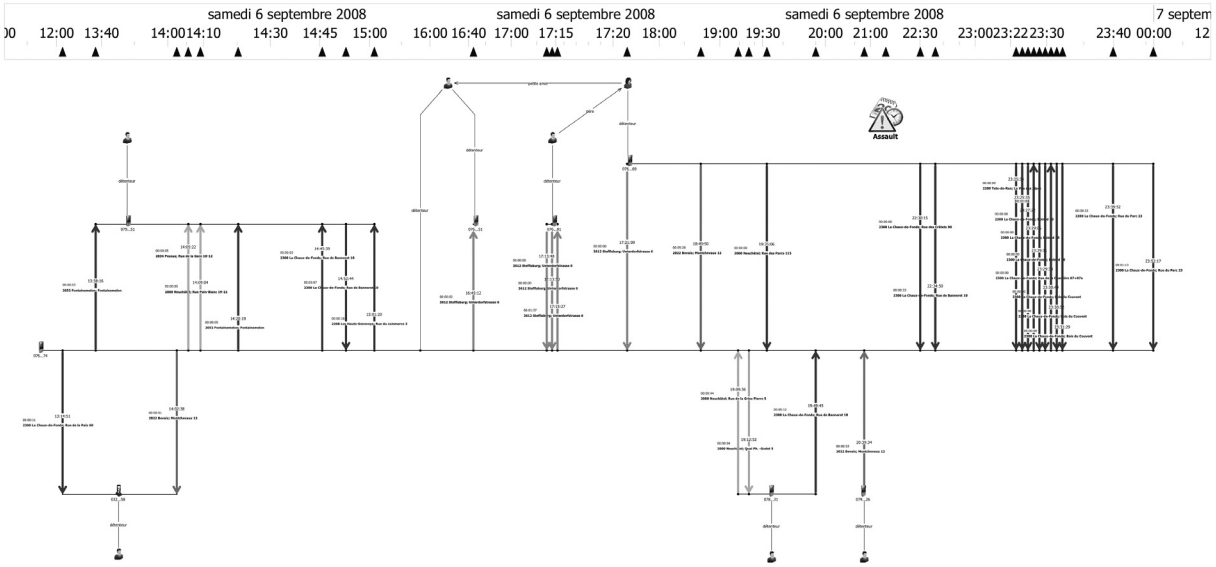


Figure 4a. Linked plots with a temporal flow chart created with Analyst's Notebook®. Each coloured arrows represent a communication (outgoing and incoming calls) horizontally fixed along the timeline.

Figure 4b. A dedicated colour is assigned to each spatial region and transposed on the flow chart (Figure 4a)

Figure 4 extents the use of geographic transposition of coloured clusters into time views by adding the relational dimension into the workspace of the crime analyst's. The relational timeline view, called temporal flow chart, was created with Analyst's Notebook® software from IBM®. Each theme line (horizontally) represents one involved phone (represented by telephone icons linked to know owners) and all calls are visualized by coloured and arrowed links horizontally aligned horizontally in time. The direction of each arrow depicts the direction of the communication (an outgoing and incoming call). This combination of temporal, spatial and relational dimensions is illustrated by an example of visualisations drawn during an assault's investigation.

The selected dataset includes all the calls made by a suspect the day of the aggression. During the first interview, the suspect denied being involved, and even having visited the red area (i.e. the location of the crime). After he has admitted that he was the only one who used this phone, investigators show him this spatiotemporal representation of his mobile phone activity (Figure 4). Exposed to the traces of his activity, the suspect explains his journey and finally confesses the crime. The detected spatiotemporal pattern brings to light the multiple displacements of the suspect from his home (in the blue area) to the home location of the victim (red area). The suspect even tried to find the victim in the orange location where he gets information the victim might be. He was in fact hunting his victim and the traces were the sign of the premeditation.

Even if the detected pattern is very specific to the case; the overall methodology to perform the spatiotemporal analysis of the traces can be generalised and used for analysing many cases. Moreover, it was illustrative of the need to integrate forensic information, crime analysis methodology and police interviews strategies to solve cases.

B. Traces and spatiotemporal analysis to detect crime series

The last example concerns the early detection of crime repetitions. This is a very common task in crime analysis. One strategy is to search for concentrations of burglaries where marks or forensic case data with similar characteristics have been collected. This process can be supported through the proposed visual approach.

The detection methodology is explained in previous researches (Ribaux et al., 2003; Ribaux et al., 2006). It is based on the exploitation of shoe-mark's traces collected from crime scenes. A simple classification system of shoe-mark's patterns is used. The analysis consists of producing spatiotemporal views, displaying the occurrence of a selected shoe pattern. If the time structure shows ruptures (temporal hotspot) or the spatial distribution of cases displays a pattern (see 54), this may indicate the activity of a single perpetrator using the same shoes. This hypothesis must be still obviously confirmed by the systematic comparisons of all the information available on each case.

The cumulative curve (on top of Figure 5) shows the occurrences of cases over time. The aim is to reveal particular increases depicting temporal hotspots. The example presented shows a slightly more complex inference structure using this spatiotemporal visualisation. The same shoe-mark's pattern was collected during three short periods spaced each time by almost one

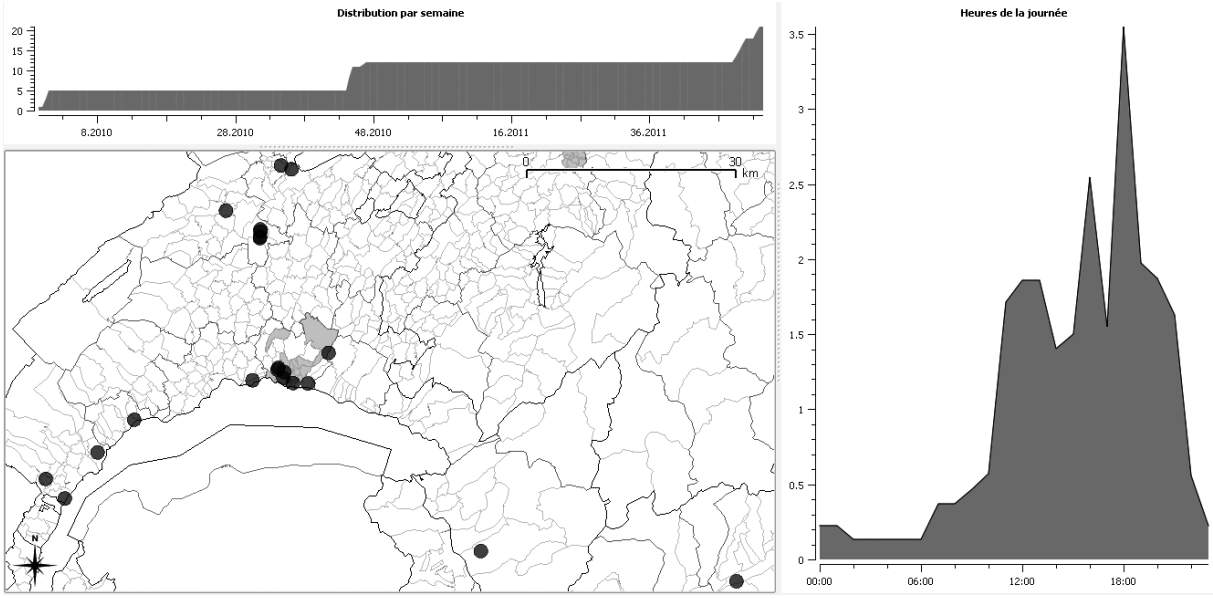


Figure 5. Spatiotemporal visualisation of crime events where traces matching a common shoe-mark pattern were collected. The temporal view on the top is a cumulative curve of crime occurrences.

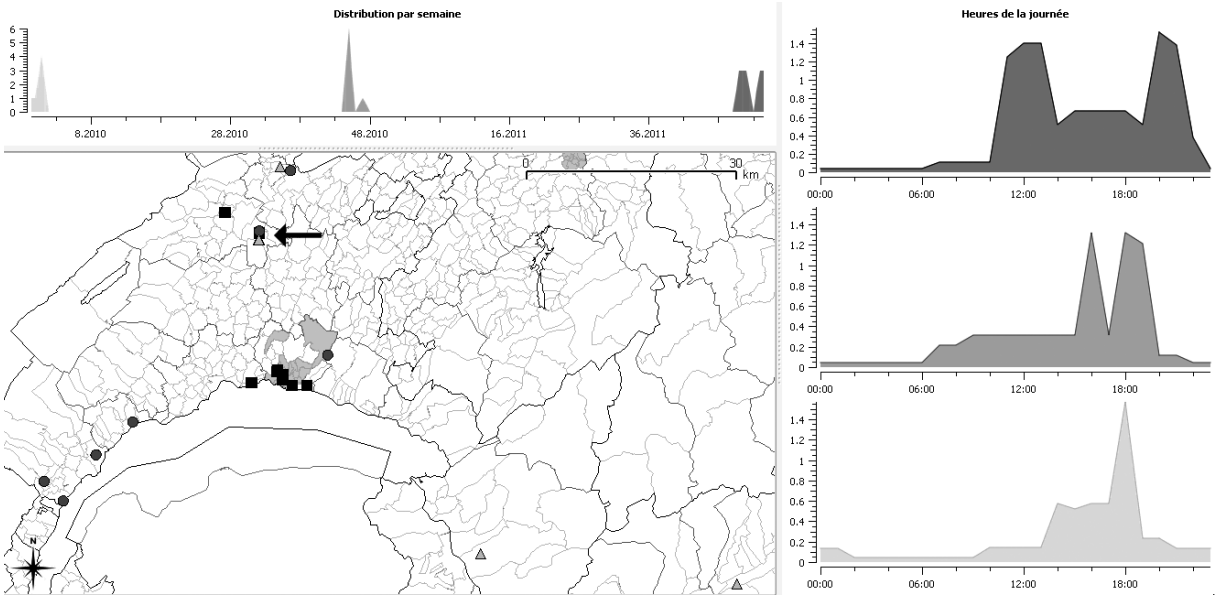


Figure 6. Comparison of the spatial distribution of the three temporal hotspots. The time frame is divided in three periods (from light grey to dark grey). The spatial distribution of the crimes for each period is represented on the map with the same colours and dedicated symbol (respectively: triangles, circles and squares).

year. Such a long period may suggest the activities of separate offenders and support the hypotheses of three distinct series. However a deeper look on each temporal hotspot and a cross-comparison of each spatial distribution reveal a new pattern (see Figure 6).

This appears by assigning a dedicated colour (reprint in grayscale) to each temporal hotpot. It reverses the way to link views. Spatial patterns and spatial hotspots are then revealed for each time frame. One unique region where cases are committed during the three periods appears clearly by these operations (pointed-out by the arrow). This region is a small town called C. Interestingly almost every cases were committed during the evening (see temporal views on the right) except three of the four cases occurring at C which occur near midday. The hypotheses that all cases are linked and that one offender may lives or has a particular anchor point near C can then be developed and lead to investigative recommendations. For instance checking police databases for already known offenders living near the detected town might be a relevant suggestion.
This example shows how complex inference structures, alternating detection of patterns and specific operations for testing hypotheses drawn (e.g. targeted comparisons of forensic case data), can be supported with fluidity by the spatiotemporal methodology and its derived tool (Ribaux et al., 2006).

VI. CONCLUSION

Traditional spatiotemporal visualisation methodologies used superimposition of the temporal dimension onto the spatial representations. An effective alternative consists of proceeding the other way round, by dividing the temporal views with an appropriate use of colours corresponding to defined geographical areas shown on a map. On this basis a computerised prototype has been developed.

The proposed combined spatiotemporal visualisation methodology has shown great potential for the analysis of all sorts of crime data, in particular forensic case data. It allows approaching a broad spectrum of situations through the visualisation and detection of complex spatiotemporal patterns and well support inferences drawing. It has been exemplified with two recurring crime analysis problems: the analysis of mobile phone billing records in crime investigation and the detection of crime series by the combination of traces and spatiotemporal pattern detection.

The case studies presented show how the approach may be used to support reasoning in investigation or more broadly in crime intelligence. In practice, several Swiss police forces currently use the developed tool for the analysis of traces left by digital devices (e.g. GPS or GSM) and for both operational and strategic analysis of crime repetitions. Indeed, the developed methodology also well supports more general analysis of spatiotemporal trends of crime phenomena.

ACKNOWLEDGEMENTS

The author would like to thank Benedict Heidl a former student of the École des Sciences Criminelles at

the University of Lausanne who have worked on the topic of spatiotemporal visualisation of forensic case data and helped to the development of the presented methodology and tool.

REFERENCES

Aigner, W., Bertone, A., Miksch, S., Tominski, C. and Schumann, H. (2007). Towards a conceptual framework for visual analytics of time and time-oriented data. In Henderson, S.G., Biller, B., Hsieh, M.H., Shortle, J., Tew, J.D. and Barton, R.R. (eds). *Proceedings of the 39th Winter Simulation Conference* (pp. 721-729). Piscataway (NJ): IEEE Press.

Andrienko, N., Andrienko, G. and Gatalsky, P. (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing,* 14(6), 503–541.

Anselin, L., Cohen, J., Cook, D., Gorr, W. and Tita, G. (2000). Spatial analyses of crime. *Criminal Justice,* 4, 213–262.

Bertin, J. (2005). *Sémiologie graphique: les diagrammes - les réseaux - les cartes*. 4ème éd. Paris: Les ré-impressions des Editions de l'Ecole des Hautes Etudes en Sciences Sociales.

Birrer, S., and Terrettaz-Zufferey, A.-L. (2008). Croisement spatial et temporel de données issues d'activités délictueuses et d'appareils permettant une géolocalisation. L'apport de la théorie des graphes dans l'automatisation de la comparaison entre des données personnelles issues de téléphones mobiles. *Revue internationale de criminologie et de police technique et scientifique,* LXI(4), 481–500.

Boba, R. (2009). *Crime analysis and crime mapping*. 2nd ed. Thousand Oaks (CA): Sage Publications, Inc.

Brunsdon, C., Corcoran, J. and Higgs, G. (2007). Visualising space and time in crime patterns: a comparison of methods. *Computers, Environment and Urban Systems,* 31(1), 52–75.

Buetow, T., Chaboya, L., O'Toole, C., Cushna, T., Daspit, D., Petersen, T., Atabakhsh, H. and Chen, H. (2003). A spatio temporal visualizer for law enforcement. In *Proceedings of the 1st NSF/NIJ conference on Intelligence and security informatics* (pp. 181–194). Tucson, AZ: Springer-Verlag.

Canter, P. (2000). Using a geographic information system for tactical crime analysis. In Goldsmith, V., McGuire, P.G., Mollenkopf, J.H. and Ross, T.A. (eds). *Analyzing crime patterns: frontiers of practice* (pp. 3-11). Thousand Oaks (CA): Sage Publications, Inc.

Card, S. K., Mackinlay, J. and Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco (CA): Morgan Kaufmann.

Chainey, S., and Ratcliffe, J. (2005). *GIS and crime mapping*. London: John Wiley and Sons.

Chen, C., and Yu, Y. (2000). Empirical studies of information visualization: a meta-analysis. *International Journal of Human-Computer Studies,* 53(5), 851–866.

Few, S. (2009). *Now you see it: simple visualization techniques for quantitative analysis*. Oakland (CA): Analytics Press.

Fredrikson, A., & North, C. (1999). Temporal, geographical and categorical aggregations viewed through coordinated displays: a case study with highway incident data. In *Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management* (pp. 26–34). Kansas City, MI: ACM press.

Friendly, M. 2008. A brief history of data visualization. In Chen, C., Härdle, W.K. and Unwin A. (eds.) *Handbook of Computational Statistics: Data Visualization* (pp. 15-56). Berlin-Heidelberg: Springer.

Guo, D., Chen, J., MacEachren, A.M. and Liao, K. (2006). A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics,* 12(6), 1461–1474.

Harris, R. L. (2000). *Information graphics: a comprehensive illustrated reference*. New York (NY): Oxford University Press.

Helms, D. (1999). The Use of Dynamic Spatio-Temporal Analytical Techniques to Resolve Emergent Crime Series. In *Third Annual Crime Mapping Research Center confrence*. Orlando, FL. Available at http://www.iaca.net/resources/articles.html.Helms, D. (2009). Temporal analysis. In Gwinn, S., Bruce, C., Cooper, J. and Hick S. (eds.) *Exploring crime analysis* (pp. 214–257). Overland Park (KS): BookSurge Publishing.

Kimball, R., & Ross, M. (2002). T*he Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd ed.). New York, NY, USA: John Wiley & Sons, Inc.

Keim, D. A., Panse, C. and Sips, M. (2005). Information visualization: scope, techniques and opportunities for geovisualization. In Dykes, A., MacEachren, M. and Kraak, M.-J. (eds). *Exploring Geovisualisation* (pp. 23–52). Amsterdam: Elsevier.

Laxman, S., and Sastry P.S. (2006). A survey of temporal data mining. *Sadhana,* 31(2), 173–198.

Lodha, S. K., and Verma., A.K. (2000). Spatio-temporal visualization of urban crimes on a GIS grid. *Proceedings of GIS'00, 8th ACM international symposium on Advances in geographic information systems* (pp. 174-179). New York (NY): ACM.

MacEachren, A. M. (2004). *How maps work: representation, visualization, and design*. New York (NY): The Guilford Press.

Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *Transactions in GIS*, 14(3), 223–239.

Playfair, W., Wainer H. and Spence I. (2005). *Commercial and political atlas and statistical breviary*. New York (NY): Cambridge University Press.

Ratcliffe, J. H. (2000). Aoristic analysis: the spatial interpretation of unspecific temporal events. *International Journal of Geographical Information Science,* 14(7), 669–679.

Ratcliffe, J. H. (2004). The hotspot Matrix: A Framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research,* 5(1), 05–23.

Ribaux, O., Girod, A., Walsh, S.J., Margot, P., Mizrahi, S. and Clivaz, V. (2003). Forensic intelligence and crime analysis. *Law, Probability and Risk,* 2(1), 47–60.

Ribaux, O., Walsh, S.J.and Margot, P. (2006). The contribution of forensic science to crime analysis and investigation: forensic intelligence. *Forensic Science International,* 156(2-3), 171–181.

Rossy, Q, and Ribaux, O. (2012). La conception de schémas relationnels en analyse criminelle: au-delà de la maîtrise des outils. *Revue internationale de criminologie et de police technique et scientifique,* 3, 345–362.

Tufte, E. R. (2001). *The visual display of quantitative information*. 2nd ed. Cheshire (CT): Graphics Press.

Wheeler, A.P. (2013). *Visualization Techniques for Journey to Crime Flow Data*. Available at SSRN: http://ssrn.com/abstract=2275379 or http://dx.doi.org/10.2139/ssrn.2275379.

Wolff, M., and Asche, H. (2009). Geovisualization Approaches for Spatio-temporal Crime Scene Analysis – Towards 4D Crime Mapping. In Z. M. H. Geradts, K. Franke, & C. Veenman (Eds.), *Computational Forensics SE - 8* (Vol. 5718, pp. 78–89). Springer Berlin Heidelberg.

Coordonnées de l'auteur :

Quentin ROSSY
Université de Lausanne,
Institut de Police Scientifique, Batochime,
1015 Lausanne-Dorigny, Switzerland.
quentin.rossy@unil.ch

BSGLg, 60, 2013, 45-57

OPERATIONALITY OF GEOGRAPHIC PROFILING THROUGH A HYPOTHETICO-DEDUCTIVE METHOD. A REVIEW OF CONSTRAINTS AND FACTORS

Marie TROTTA, André LEMAÎTRE & Jean-Paul DONNAY

Abstract
This paper is dedicated to the identification of the constraints and factors enabling the computation of an effective geographic profile, with the specificity of focusing only on the elements that could be available during an investigation. It aimed at filling the gap between the inductive demarche of environmental criminology and the deductive, operational procedure followed by geographic profilers. It reviews successively the relationship between the premeditation, the seriousness of the facts, the nature of the offences and the spatio-temporal pattern of the crimes with the criteria required to build effective likelihood surfaces in geographic profiling. A decision tree is provided as a tool for evaluating the risks of an ineffective geographic profile with regard to the non-respect of the different conditions.

Keywords
environmental criminology, decision making, inference, serial offenders, geographic(al) profiling, spatio-temporal analysis

Résumé
Cet article est dédié à l'identification des contraintes et facteurs permettant la construction d'un profil géographique efficace, avec la particularité de s'intéresser aux seuls élements disponibles durant une enquête criminelle. L'article cherche à combler le vide méthodologique entre la démarche inductive de la criminologie environnementale et la procédure déductive et opérationnelle suivie par les profileurs géographiques. Il examine successivement la relation entre la préméditation, la gravité des faits, la nature de l'infraction et la configuration spatio-temporelle des crimes avec les critères nécessaires à la construction de surfaces de vraisemblance efficaces dans le profilage géographique. En synthèse, un arbre de décision permet d'évaluer les risques d'un profile erroné en fonction du non-respect des différentes conditions.

Mots-clés:
criminologie environnementale, prise de décision, inférence, auteurs en série, profilage géographique, analyse spatio-temporelle

I. INTRODUCTION: TOWARD MORE OPERATIONAL CONCERNS IN CRIME MAPPING

Pin Mapping is a common practice carried out by analysts such as geographers, statisticians and practitioners in various disciplines (epidemiology, marketing, criminology). If such approach is not recent in criminology (work of Quetelet and Guery in the 19th.), the study of the geographical environment in order to understand the offender's spatial decision process is a more recent approach developed by the environmental criminology (Brantingham and Brantingham, 1981a). Several studies in this field concern the journey-to-crime, with the elementary but fundamental finding that there is a friction with the distance (the distance decay effect) between the crime location and the offender's home or anchor point (Phillips, 1980, Rhodes and Conly 1981, Rengert et al., 1999). According to the crime pattern theory, offenders generally are less likely to commit their crimes far from their activity nodes (Brantingham and Brantingham, 1990).

Since the nineties, Geographic Profiling (GP) presents several issues distinct from the environmental criminology (Rossmo, 1997). The principle of distance decay is maintained and makes possible to build a likelihood surface under certain conditions. However, unlike the largely inductive approach developed in environmental criminology, GP seeks the residence or anchor for an unresolved series of crimes. It corresponds to a hypothetico-deductive process. Therefore, it must rely on a model and on assumptions limiting or favouring the GP application in order to make it consistent with such a

deductive approach.

Among the first constraints that allow the application of GP, we find the classical distinction between marauder and commuter offenders (Canter and Larkin, 1993). The first commit their crimes in their home range while the latter are travelling outside it. Therefore, GP applies preferentially to marauders. Rossmo (1997) classification between poachers and hunters is quite similar, and only the actions of the former are candidates for a GP analysis. Rossmo adds yet an important criterion to which this paper returns later: premeditation. As only the marauders can be studied by GP, several analyses have sought to distinguish them from commuters. Literature suggests simple geometric but hardly discriminating theories: the circle or convex polygon theory, the nearest neighbour index, etc. Laukkanen and Santtila (2006) try to connect the distance travelled by the offender to characteristics of crime to facilitate the distinction between commuters and marauder. On the other hand Paulsen (2007) adds a valuable time criterion: longer cool-off period between offences tends to reflect marauder behaviour. However he distinguished both behaviours on the idealized geometrical criterion of the circle theory.

Behind those categories characterized by a difficult distinction, Rossmo (1997) lists other conditions to allow the creation of a likelihood surface:

- *The profile must be based on several crime scenes. It can be locations of different crimes or several places associated to the same crime.*

- *The crime scenes must have been attributed to a same offender*. Rossmo and Velarde (2008:36) redefined this postulate as: *the linkage analysis for the crime series is accurate and reasonably complete (i.e. there are not a significant number of unlinked crimes that should be part of the series).*

- *The offender's residence or anchor point and the area of criminal activity must not be separated by a too long journey.*

- *The targets are distributed more or less throughout space.*

- *The offender must not change his anchor point or operates from several different anchor points during his crime series.*

However, these conditions are questionable. There is no consensus on the minimum number of crimes to constitute a series (5 according to Rossmo, least or more according to other authors). The latter condition cannot be validated in the case of an on-going investigation, without any knowledge of the offender or a suspect. Besides, the metric used to measure distances should vary according to the organization of the road network. Generally, the environment, as perceived and described in North American cities, is different from that of the old European cities (street network, density, etc.) what can lead to quite distinct distribution of targets and crimes (Brantingham and Brantingham, 1990) as well as different micro-spatial offender's behaviours (Alston, 1994 and 2001).

The use of GP in an operational situation is only possible under favourable conditions, which can only be inferred from the characteristics of the recorded crimes. "*This will include everything that the police may know before the offender is identified"* (Canter, 2011: p6) However, what is the relationship between the crime characteristics and the GP applicability ? Recent literature in geographic profiling (e.g. Knabe-Nicol and Alison, 2011, Canter, 2005) tries to answer to this question. It recognises that such a deductive approach requires a theoretical framework as well as hypotheses strictly defined on both the offender's decision-making behaviour and on the way to deal with available spatio-temporal information.

In order to improve the operability of geographic profiling, this paper has for objective to identify the constraints and factors enabling the computation of an effective geographic profile, with the specificity of focusing only on the elements that could be available during an investigation. It aimed at filling the gap between the inductive demarche of environmental criminology and the deductive, operational procedure followed by geographic profilers. Inspired by Capone and Nichols (1975),(i) the kind of offence, (ii) elements from the crime scene and (iii) temporal aspects will be analysed through their relationship with the spatial dimension.

II. CONSTRAINTS FOR THE APPLICATION OF GEOGRAPHIC PROFILING

What are the circumstances for which an application of geographic profiling may not be effective? The determination of the constraints aims at answering to this question. A constraint is a binary, excluding criterion that implies two possible situations: the respect or not of the constraint. If this one is not respected, the implementation should be rejected.
Constraints will be discussed successively through several characteristics that generally maybe inferred from the crime scene: the seriousness of the facts, the premeditation and the relationship between the victim and the offender. The reader may consult (Douglas et al., 1992) for explanation on the relationships between the crime scenes elements and those characteristics.

A. The seriousness of criminal activities

The seriousness of the crimes influences the resources spent for the investigation and the information collected by the police force.

A priori both petty and violent crimes could benefit from a geographic profiling.. However, in practise, petty crimes are too numerous and treated by diverse municipalities so that the spatial and temporal information collected for such facts is too often incomplete to implement GP analyses. Besides, victim's expectations for such offences are lower so that fewer resources are spent for solving them.

It is worth noting that, with time, mapping and GIS-like softwares (eg: CrimeSTAT, RIGEL ,Dragnet respectively developed on the research works of Levine, Rossmo and Canter) have reduced the time and the qualifications required to apply GP techniques. Today literature proposes examples of GP applications to less serious offences such as vehicle theft (Tonkin et al., 2010) or burglaries (Laukkanen et al., 2008).

Nevertheless these last studies neglect the way the offences could be linked together before the solving of the investigation. Crime series is defined as several crimes attributed to the same single or group of suspect(s). Tracing techniques such as DNA, ballistic, etc. are the best and often the only way to demonstrate or at least suppose the belonging to a series. Failing that, a comparison of the modus operandi is at least required to link offences. For these many reasons, violent crimes involving thorough investigations are often better connected to each other.

Moreover GP is always based on some offender's spatial or spatio-temporal behaviour. In order to build this hypothesis, it may be necessary to resort to psychological profile which is mainly studied for violent crimes. The VICAP (Violent Criminal Apprehension Program -United States) and the VICLAS (Violent Crimes Linkage Analysis System - Canada, United Kingdom, Belgium, etc.) allow a systematic tracking of the recidivist offenders by identifying the similar behaviours and modus operandi. And these systems are again focused on violent crimes and attempts (sexual offence, murder, kidnapper, child luring).

B. The premeditation

The distinction between premeditated and not premeditated acts is fundamental to apprehend the spatio-temporal hypotheses on the offender's behaviour. For series of violent crimes, premeditation presupposes one or more choices on the part of the offender: victim (specific or specific type), place and time, and possibly the route to follow to reach or leave the crime location from or towards a single anchor point.

In premeditated acts, offender is driven by the crime, what Elffers (2004) calls a crime-led journey. This concept is crucial as it reflects a logical, calculated behaviour that suits well to the rational choice theory. Premeditation may involve a prior identification of crime sites and journeys, which is not without consequences in the analyse of the offender's travels. In contrast, the opportunistic crimes are more influenced by external choices without a specific choice of victim. In those situations, the influence of environmental factors can affect the consistency of the offender's behavioural (Alison et al., 2002). This, in turn, makes difficult the connection with the other crimes of the series.
The major difference between opportunistic and premeditated decisions is that the journey followed to commit the offence in premeditated acts is driven by the rational choice. The minimisation of distances in the journeys-to-crime is mainly based on this theory, considering that travelling far from the anchor point has a cost (Beauregard et al. ,2007,Brantingham and Brantingham 1990).

In opportunistic situations, places and moments are randomly chosen. At best, it can be assumed that opportunistic crimes are committed in a low risk area what corresponds to an offender's known spatial environment. It could be not far from his residence, but equally not far from any well-known location such as relative's residence, work place, shopping place, etc. In such situations, the anchor point is probably not constant in the series, making event more difficult the implementation of GP. These observations explain why GP should be restricted to premeditated behaviours.

Premeditation also underlies the categories of "organized" offenders (vs. disorganized) identified by Ressler et al.(1988) and "ritualistic" offenders (vs. impulsive) recognized by Hazelwood and Warren (2000). Paradoxically, greater mobility is observed in premeditated acts (organised and ritualistic offenders) what might seem weaken the assumption of a distance decay effect. But the principle of cost minimization, often cited to explain this assumption, only makes sense if the offender adopts a rational behaviour to reduce simultaneously his travel expenditures (time, cost) and crime risks. In Rossmo's typology (2000), geographic profiling is only applied to the hunter who leaves his residence with the purpose to commit an offence and looks for interesting target from this place. This category is explicitly linked to premeditation.

The distinction between expressive vs instrumental crimes may also be related to the criterion of premeditation. In terms of distances Fritzon (2001) mentions several studies according to which an offender travel

greater distances to commit an instrumental crime. This can be compared to the observations for premeditated acts. The meaning of instrumental may, however, differ from one author to another, implying different relationships with the concept of premeditation. In Santtila et al. (2008:346), an offence is *« instrumental when the offender attempts to achieve goals that serve some ulterior purpose which is external to the actual offence* ». According to this definition, even if premeditated, the instrumental crime will not be the final purpose of the offender. This implies that there may be inter-dependence between the journeys or spatial choices made to commit both the instrumental and the final crimes. In such a situation, instrumental crimes should not be directly introduced in a geographic profile.

According to Wortley (2008), an instrumental violence is defined as " *a planned attack with a clearly formulated purpose while an expressive violence is an impulsive reaction to events carried out in the heat of the moment* ". With regard to this definition for which the premeditation criterion is explicit, murders or sexual assault can, according to the circumstances, be classified as instrumental or expressive offences. In this case, it is clear that GP profile will be much more effective for instrumental offences.

This section shows that distinguishing premeditated and opportunistic events before applying GP is very important. For this reason, the premeditation has to been estimated, before any geographic profile, by agents specialized in behavioural analyses. They may evaluate it thanks to elements from the crime scenes such as the use of weapons, similar modus operandi for several offences, etc. A large literature exists on the subject but is not the purpose of this article.

C. The nature of criminal activities

The nature of criminal activities plays a key role in the creation of a geographic profile. Typologies based on the offender's characteristics are often restricted to a single type of activities because of the different spatial or behavioural hypotheses they imply. Typologies based on offender's characteristics are often built for a specific crime type as behavioural hypotheses will be different for burglars or sexual offender for example. They are not governed by the same motivation; they do not have the same constraints, etc.

As this paper considers the situations favouring the application of GP, we can only focus on the types of crimes that do not contradict criteria relevant for the use of spatial and temporal data. Geographic profiling will then preferentially be applied to the most serious offences (mainly the violent crimes) due to their better documentation and the opportunity to build a series for which a rational behaviour is conceivable. Those violent crimes are homicides, rapes and sexual assaults, and arsons according to the FBI classification (Douglas et al., 1992). Burglaries may be added in the specific situation where offences have previously been linked by the comparison of modus operandi or ballistic / DNA analyses.

With regard to rapes and sexual assaults, geographical data are useless for domestic acts. Specific situations of multiple offenders have also to be excluded as it multiplies the risk of several anchor points. Apart from those situations, rapes and sexual assaults are probably the best category of crimes that can benefit from the geographical profiling. Series of offences can often be tracked with precision (DNA) while the victim's testimony facilitates the precise determination of place and time as well as the offender(s)'description (mode of transportation, for example).

The ViCLAS system records the homicides,i.e. the deaths of a human person caused by another one. If they fall obviously into voluntary acts, only a part of them are premeditated and very few can be connected with a series. The expression "serial murder" refers to" *the unlawful killing of two or more victims by the same offender(s), in separate events"* (ViCAP, 2008). The presence of a cooling-off period distinguishes the serial murder from the mass murder who commits all crimes simultaneously (Lundrigan and Canter, 2001).

The serial murder can select his victims in specific categories of the society (according to the social status, the race, the religion, etc.). The more specialised is this selection, the weaker will be the relationship between the crime locations and its anchor point as the offender typically travels to very precise places where those victims are located.

The mobility is also varying according to the nature of the offences. Property offenders travel on average further than rapists (e.g. White, 1932, Rhodes and Conly, 1981 cited in Beauregard et al., 2005). For serial murders, Holmes and De Burger (1988) distinguished geographically stable and transient ones. While the geographically stable murders live, kill and dispose of the bodies in the same or nearby area for some time, the geographically transient murders travel continuously from one area to the next and dispose of bodies in far-flung places (Lundrigan and Canter, 2001). However, this last study demonstrates that the home location has a strong centralizing influence on the spatial patterns of disposal locations. The highest risk of failure for a geographic profile is then the possible change of residence between crimes for murders presenting a transient behaviour.

Finally homicides motivated by fanaticism and terrorism should be considered in isolation. Bennell and

Corey (2007) have studied the applicability of geographic profiling for terrorism. . One difficulty lies in the multitude of anchor locations as in the terrorism, criminality is organised around a widespread network of membership with multiple offenders who can live in different areas. Besides, It requires a very good knowledge of the organisation and their objective as those influence the choice of targets." *terrorists with a specific target-selection strategy will be more likely to exhibit commuting behaviour.*"(Bennell and Corey, 2007:194)

III. FACTORS FAVOURING THE APPLICATION OF GEOGRAPHIC PROFILING

The previous sections described non-spatial elements of criminal investigation that deeply constrain the implementation of geographic profiling. Even if not spatial, those elements already provide essential information about the possible relationship between the crime locations and the offender anchor point. This is essential to assess the validity of the spatial hypotheses underlying GP methodologies.

Complementary to the constraints, some elements may be more favourable to the implementation of geographic profiling without constituting excluding criteria. Those elements correspond to the factors which are quantitative (or ordered) criteria strengthening or reducing the relevance of an alternative (here the effectiveness of the geographic profile). The next section is dedicated to the discussion of three categories of such criteria: the offender's characteristics, the spatial factors and the temporal properties imbricated with the spatial dimension.

A. Offender's characteristics

Among the multiple social and economic factors describing the offender, three of them have a major influence on the geographical data characterizing a criminal investigation: (i) the offender's age, (ii) the socio-economic environment where he grew up and where he lives, and (iii) its mode(s) of transportation. Those characteristics may be estimated thanks to the information provided by the victim or some witnesses.

The age and the socio-economic environment indirectly interfere on the geography of crimes via the experience and knowledge of the environment. It is accepted that the length of the journey-to-crime increases with age, younger offender's having a more limited knowledge (Brantingham and Brantingham, 1981b). This is supported by several studies in different countries such as Canada, England, the Netherlands, and the United States but also for different types of crime such as burglary, rape, arson, robbery murder (many references in Snook, 2004).

Mutatis mutandis, the living environment would have the same kind of influence. A rich environment is associated with a wider offender's area of activity. This tendency is partially explained by the opportunity to travel with car as lower socioeconomic status or young age might make it more challenging to use a car and move about when committing crimes. (Snook, 2004, Laukkanen et al., 2008)

The mode of transportation is a key factor in GP analyses. Snook showed that distances to crime vary in function of the mode of transportation (Snook et al., 2005). Private vehicle, public transport and footpath are the three modes to distinguish. Different travelling speeds are associated to each of them and therefore different time distances, itineraries and even specific schedules. The private vehicle is the less constraining mode and, used alone or with another mode of transportation, it considerably widens the geographical area for the investigation while keeping limited to street network and traffic plans. The mode of transportation also influences the spatial pattern of crime locations. Public transportation implies a pattern concentrated around access locations. The more poorly the mode of transportation deserves the region, the more concentrate the pattern will be.

In the deductive process of GP, the assumption concerning the mode of transportation will then greatly influence the location and shape of the prior search area. An assumption of pedestrian behaviour will suited to the classical application of the journey-to-crimes function with Euclidean distances. Car-driver ones would require the development of other methods based on road-network distances where the distance-decay effect has less influence. Levine and Block (2011:226), among others, note that: "*more research is needed on integrating additional information to narrow the likely origin location of the offender, such as land use information and actual travel networks (e.g., roads, transit).*"

B. Spatial factors

As this paper aims at developing a procedure for evaluating the effectiveness of geographic profiling, the spatial properties of the crime locations, directly available for the investigation, have to be deeply taken into account. The principle according to which the characteristics of crime locations are connected to the offender's spatial behaviour is fundamental in any methodology of geographic profiling. Geographers are familiar with the analysis of such components and this section proposes to investigate three geographical concepts: the place attractiveness, the density and the proximity.

1. The place attractiveness impacts the distance decay from the crime location

The concept of attractiveness has been deeply studied from a spatial perspective. It forms the basis of the central place theory (Christaller, 1966) according to which the urban centres with the higher levels of services are characterized by larger hinterlands than those with lower levels.

This reasoning can be transposed in the field of the environmental criminology: places offering the most criminal opportunities attract offenders who are willing to travel longer distances. Correlatively, such places are those where the proportion of local offenders is relatively smaller.

Brantingham and Brantingham (1981a) make an interesting distinction between the crime generators and attractors. The crime generators are places attracting a large number of people but for reasons unrelated with criminal motivations. By contrast, the crime attractors are particular areas well known by offenders for their criminal opportunities. Offender may travel quite long distances to reach those locations. In addition to these two categories, the same authors define the crime neutral areas as place that do not create particular offending opportunities and do not particularly attract people. When an offence occurs in such a place, there is a high probability that it was committed by a local insider. Distance decay or simple pathway models seem appropriate to model the offender's journey-to-crime for these specific places (Brantingham and Brantingham, 2008).

Does this mean that the distance decay effect does not have any influence on offenders travelling to the crime generators or attractors? Probably not and it would be more correct to postulate that the slope reflecting the decay varies with the place attractiveness. The more attractive is a location, the slighter is the slope. In the figure 1, the probability to commit a crime in a crime generator or attractor is first higher than in a neutral place as more crimes occur there by definition. Besides, the slop is very steep for neutral places as they are mainly the place of activity or awareness of local offenders.

Finally, the attractiveness is also influenced by the individual's perception and experience, which results in different spatial mobility (Beauregard et al., 2005). Some attempts exist in the literature for evaluating the "absolute" place attractiveness with a matrix of origin-destination of known offenders on the analysed territory (e.g. Levine and Lee, 2009). But this approach neglects individual's preferences. However, an evaluation of the "relative" place attractiveness may be more useful. It consists in comparing the crime locations of the series and in evaluating their respective neutrality. Sharp distance decay should be in priority applied to the neutral locations.

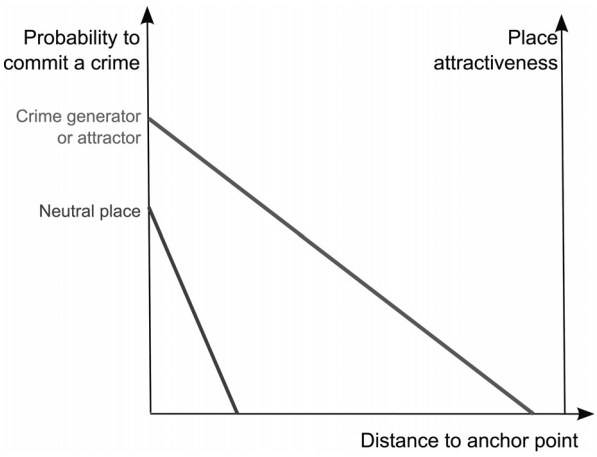


Figure 1. The relationship between place attractiveness and the length of the journey-to-crime.
The slope of the distance decay should vary according to the place attractiveness.

2. The potential index: influence on possible victims and on the efficiency of searching methodologies

Density is the ratio between a specified population and a unit area. This concept is extremely dependant of this chosen unit area. In geographic profiling, we will prefer to focus on the potential defined as the ratio between the population and the distance which is directly in line with the application of distance decay.

At small scale, the potential is connected to the length of the journey-to-crime. In areas with high potential, the threshold to access a sufficient number of possible targets is reached at a shorter distance from the offender's residence. Of course, this distance is still function of the kind of targets that the offender is looking for. Population potential will be, for example, replaced by industrial building potential if a burglar is focusing on such properties.

At a larger scale, the potential influences the size of the buffer area (area close from an offender's anchor point and considered by him as too risky to commit a crime). It has chances to be reduced in zones with a high population potential. This must be linked to the capability of guardianship. First in such areas, people do not know well their neighbours According to social disorganisation theory, residents living in areas characterized, among other factors , by high building density are less able to perform guardianship activities.(Sampson, 1983). Secondly in the same areas, police officers have to manage a lot of potential offenders and targets what reduce their control capacities

3. Proximity: a central concept in crime linkage analysis

The concept of proximity is central in the linkage of criminal activities. If the first meaning of proximity is the nearness in space or time, it must be analysed in combination with its second definition: nearness or closeness in a series. Does proximity in space and/or time help to link criminal events and to which degree of effectiveness?

Firstly, proximity makes easier comparisons between cases and their modus operandi. If all the offences occur in the same police area, the investigators are generally aware of the other similar cases. By contrast, communication can be very limited between different police services especially for the most common offences. This comparison is also easier with temporal proximity. The reason is not directly connected with communication but more with memory, the capacity to remember similar events in the past. It will be easier to remember a similar modus operandi if the previous offence occurs only some days or weeks ago.

However, crime officers can have a more proactive behaviour by monitoring possible linkages for typical offences. In the context of an exploratory research, the spatial and temporal proximities are two dimensions of the hyperspace of information (Turton et al., 2000) with respect to which data mining can search for nearby events. By blocking a range of space and time, the police may have a systematic comparison of the modus operandi of nearby events.

In the cases described above, the linkage is based on similar modus operandi while the spatio-temporal proximity facilitates their comparison. However, the temporal and spatial proximities of crimes can be themselves important criteria for the linkage.

The impacts of geographical and temporal proximities in linkage have been studied for offences such as serial burglaries, serial car thefts or serial sex offences (Bennell and Canter, 2002; Bennell and Jones, 2005; Goodwill and Alison, 2006; Tonkin et al., 2008; Grubin et al., 2001 cited *in* Markson et al., 2010). The analysis of pairs to link crimes showed that there was greater consistency in the spatial and temporal similarities than those in the modus operandi (Goodwill and Alison, 2006; Grubin et al., 2001). In order to estimate geographical proximity, studies mainly used the mean inter-crime distance. Studying temporal proximity, Goodwill and Alison (2006) showed that the day interval was a better indicator of crime linkage than the time of crimes (hours).

4. Unit area issue and its impacts on factors

The problem of the modifiable unit area (MAUP) is well-known in geography where it is considered as a source of bias in multiple spatial analyses (Openshaw, 1984; Cressie, 1996; Unwin, 1996). Geographic profiling obviously is not exception to this issue. As soon as point data – occurrences or measures - are aggregated by postcode areas or police precincts for instance, the selected boundaries affect the meaning and the significance of the figures. This section does not develop deeply all the impacts of the MAUP but highlights the influence it has on already discussed factors.

The most frequent inconsistency is a scale issue: the unit areas are too small or too large related to the phenomenon under investigation. For example, police crime statistics aggregated at the municipal level make impossible the identification of crime attractors or generators at the street or city-block levels. The population potential computed for each unit may hide great variations inside the area. In the same vein, measuring the overall attractiveness of a city can hide significant spatial variations between its different neighbourhoods. Recent studies advised to choose a small spatial unit (Weisburd et al., 2009), especially for studying the attractiveness of crime places (Bernasco, 2010).

However, they do generally not propose to consider simultaneously several levels to evaluate place attractiveness. They neglect the fact that a place may have different attractive influences in function of the level of observation. A city with potential attractive functions, at a national level for example, has also ordinary functions for which the attractiveness is reduced to local vicinity. For example, the common bar institutions, located outside the most frequented neighbourhoods probably do not have a greater attractiveness than those located in smaller, medium sized cities. A first advice is then to favour multi-scale analysis for the evaluation of this factor.

Secondly, the spatial partition forced / imposed by the use of unit areas may conflict with a phenomenon which is intrinsically spatially continuous. The distance decay model used in geographic profiling is precisely a spatially continuous model (in an isotropic or anisotropic space) which can hardly be correlated with figures aggregated in arbitrary areal units. For example, the influence of nearby units may be important (Bernasco and Block, 2011) when the units are small in comparison with the length covered by the distance decay (smooth slope). In addition, arbitrary space discretization generates a segmentation of information. Thus, a police officer investigating a crime committed on the edge of a spatial entity and who is not aware of similar crimes in a neighboring entity, will tend to move the assumed anchor point towards the centre of the entity under investigation (Rengert and Lockwood, 2009).

There is also concern about the impact of unit size on the calculation of rasterized surfaces. Thus, the accuracy of the likelihood surfaces is affected by the over-or underestimation of distances depending on the position

of crime sites in cells superimposed on the study area. A second advice should be to pay attention to the possible influence of close units when small unit areas are chosen. This influence should be taken into account in the modelling with a kernel density estimator (Levine, 2004).

C. Integration of the temporal properties to complete spatial factors

The previous section revealed that the temporal dimension is extremely linked to the spatial factors conditioning an effective geographic profile. Capone and Nichols (1975) already observed that time impacts the average distance to crime. Crime generators during shop opening hours are different from those during the night. Spatial proximity is useless without the temporal one.

Introducing temporal information presents several issues. First, the information is often missing, police officers failing to record the temporal elements (Brantingham and Brantingham, 2003). Besides, when recorded, it is often imprecise as the victim is not capable to report an accurate time of when the crime occurs because of his absence (burglary) or because of the choc (rapes). Thereby, the precision about the moment of crime varies, among others, in function of the offence type with better precision for robbery, assault or street offences than burglaries (Ratcliffe, 2002).

However, even when it is imprecise, temporal information may still be useful. The temporal dimension is multi-scale: from the years to the seconds, with several dichotomies organising people activities: night/day, working/vacation days, or even categories such as the days of the week, the hours, etc. Chainey and Ratcliffe (2005) describe several temporal categories: the moments, the duration, the structured time (hours), time as a distance and time span, and all those categories can be used in an investigation. The following sections analyses successively several temporal categories that may bring new insight on the offender's spatial behaviour. It details the relationship between the offender's crime locations and anchor point thanks to the introduction of the temporal dimension.

1. The Moment of crimes: relation with offender's activities, potential targets and absence of guardianship

The moment when crime occurs is influencing our understanding of the offender's decision process (evaluation of risks, costs and benefits according to the rational choice). From the temporal point of view, place properties are not static. They evolves over long time periods but also according to the hourly rhythm of human activities. Time influences the three components of the crime described by the routine activity theory: the presence/absence of offender, potential target and capable guardian (Cohen and Felson, 1979). As" *the relevant actors – victims, offenders, guardians, and place managers – adjust their relative densities over time and around specific places, the opportunities for crime shift and coagulate* "(Ratcliffe, 2010:15).

From the perspective of profiling, the timing of the crimes informs police investigators about the offender's activities or constraints. Indeed, the facts taking place as well during the day as during the night will rather be liaised with someone unemployed, with a rambling lifestyle. By contrast, the facts occurring on very small time slots could mean more constraints for the offender such as work with a regular schedule, a family to whom he must justify his absences, etc.

Potential victims are also changing through time. The concepts of crime attractor and generators are particularly closely related to the temporal dimension (Brantingham and Brantingham, 2008). Depending on day or night, a location can be either a crime generator or a crime attractor or a neutral place for the same crime type. Supermarkets or commercial locations are crime generators only during the day for pickpockets. A very busy place during the day may become completely deserted at night, for example in business districts, and vice versa for residential areas. This results in differences in the presence of guards during the day.
The integration of temporal dimension is then crucial when analysing the pattern of crime locations in order to evaluate if they can be considered as neutral so that the distance decay could have some positive impact.

2. Spatio-temporal clustering

Among conditions enumerated by Rossmo (2000), the presence of a stable anchor point is generally required for any GP methodology. This condition, if difficult to validate, can be better estimated with the integration of the temporal dimension.

Indeed, the analysis of spatio-temporal clusters is a first indication of multiple anchor points or different modes of transportation. The offender or victim's mode of transportation can vary according to the days of week. As an example, a young offender is studying in one city, leaving there during the week, travelling mainly with public transportations. During the weekend, he comes back to his parent's house in the countryside. He travels there only as a pedestrian. By dividing the series into two sub-series (week and weekend), the two spatial patterns around his two anchor points will be clearly identified. Besides, the mean inter-crime distances for the weekend pattern will probably be shorter than the week one. Besides, the clustering can be explained by the choice of victims such as a rapist operating in the vicinity of bus stops in the morning hours but never during the weekend given the lack of victims.

3. Chronology

The chronology of events allows to precise another of the Rossmo's assumption: the presence of several crime locations. As it was described above, Lundrigan and Canter (2001) studying serial murders, insisted on an interesting temporal aspect: the presence of an emotional "cooling-off period" between each crime as opposed to the mass or spree type, in which all crimes occur more or less simultaneously. If all the crimes occur in a few hours, there is a high probability to face one or multiple offenders commuting to the areas. The several locations are only steps in the same journey-to-crime. Several crimes locations are then not enough; they must also be separated in time to build a likelihood surface based on multiple journeys-to-crime.

This cooling-off period can be days, weeks or months. A smaller number of days between crimes seems to be correlated with a commuter behaviour (Paulsen, 2007). By contrast, if crimes occur during several months or years at locations not far from each other, the conditions of stable anchor point and small journey-to-crimes will often be satisfied. However, longer time periods increase the probability of a moving anchor point.

If an offender doesn't go back to his residence after a crime, conducting a circuit path, the chronology indicates the direction/sense of this path. It provides then information about his provenance.

The relationship between distances travelled by the offender and the chronology of events is reviewed by Snook et al. (2005) for serial murders. Both a decrease and an increase in distances from home base to crime locations have been observed (Godwin and Canter, 1997, Rossmo, 2000) and find a logical explanation. An increase in travelled distances could be explained by the fear to be recognised, leading the offender to travel further from his anchor point. A decrease would be the result of a confidence-building, the offender taking more and more risks.

"Closely associated with series chronology is the belief that serial offenders live in closer proximity to their first crime location than their subsequent crime locations (Canter, 1994, Canter and Larkin, 1993, Warren et al., 1995)" in (Snook et al., 2005: 150). A more steady distance decay to home should then be observed from this crime location. But even this believed is not always observed. While Rossmo (2000) found that 41% of serial murders commit their first crime at the nearest location from home, the closest offence location corresponds to the first one for only 18% of the serial rapists studied by Warren et al. (1995)

Besides, in an operational perspective, police investigators can never be sure that the first offence recorded in the series was the first committed by the offender. Some victims may have not complained their injury. Or even, the offender was maybe already involved in other criminal activities before which had some influence on its knowledge of crime locations.

IV. DISCUSSION: A DECISION TREE SYNTHETIZING CONSTRAINTS AND FACTORS.

A decision tree synthetises the relationships between the constraints and factors discussed in the paper and the usual conditions required to build an effective likelihood surface in GP. The tree is a tool for investigators confronted to a new investigation for which it should be decided to build or not the geographic profile.

It should be read from top to bottom. At each step, the investigator can check which conditions (italic) depend on the constraint or factor. The constraints of the premeditation and the seriousness of the offence have a binary reading while the factors have a more gradual impact on the application of GP. Premeditation is linked to a rational choice for which a single anchor point and short distances are more frequent than in opportunistic situations. Violent crimes are preferred to the other offences for the resources spent in linking such crimes and to meet the higher expectancies of the victims.

At the third level of the three, decisions must be taken with regard to different components: the type of crime, the offender's properties and the spatio-temporal pattern of the crime locations. For example, serial rapes and sexual assaults are the crime for which an application of GP generally presents the best conditions: small journeys-to-crime, easy determination of multiple offenders by contrast to arsonists who are often acting in groups. The application to serial murders is also possible if some categories such as the extremists or terrorism are rejected. For the spatio-temporal pattern, neutral places are crucial for applying distance decays functions but place attractiveness should always be evaluated for a specific moment.

V. CONCLUSION

This paper had for objective to identify the constraints and factors enabling the computation of an effective geographic profile, with the specificity of focusing only on the elements that could be available during an investigation. It aimed at filling the gap between the inductive demarche of environmental criminology and the studies on the journey-to-crime and the hypothetico-deductive, operational procedure followed by geographic profilers.

The article was based on the conditions, mainly defined by Rossmo (2000), necessary for the implementation of likelihood surfaces with a particular attention to the respect of the distance decay and uniform distribution

of potential targets around the anchor point, two conditions closely linked to the geography of the crimes.

With regard to the constrains, premeditated violent crimes committed by a single offender are required to build a geographic profile on the hypotheses of a rational choice, unique anchor point with a sufficient number of linked events. Concerning the factors, the spatio-temporal pattern of the crime series plays has a major influence on all the conditions defined by Rossmo. Geographic concepts such as place attractiveness, population potential, spatio-temporal proximity have to be evaluated for each crime location in order to estimate which condition risks to be rejected. Especially, neutral places associated, by definition, with lower attractiveness should be distinguished from crime generators and attractors in the implementation of likelihood surfaces. A distance decay function with a steepest slope should be applied on the neutral places. Temporal aspects are extremely connected to the spatial behaviour. The moment of crimes informs the investigators on the offender's constraints and helps to qualify places as attractive or not. Spatio-temporal clusters may be associated with sub-patterns around different anchor points. The chronology is an indicator of the stability of the anchor point.

This decision tree provided as a synthesis should be considered as a tool for evaluating the risks of an ineffective geographic profile. Investigators may still develop an alternative approach to the classical likelihood surfaces that does not require the unmet condition. The uniformity may not be required for Bayesian approaches that include origin-destination matrices (Levine and Lee, 2009). The distance decay effect can be replaced by a minimization of travelled distances or departure time (Trotta, 2012, Trotta et al., 2011). The investigators will have to choose for the appropriate spatial hypothesis and develop a corresponding suitable research methodology. Until now, such ways of modelling that do not rely on domocentricity have been less studied (Canter and Youngs, 2008) and still need to be tested on large sample of data.

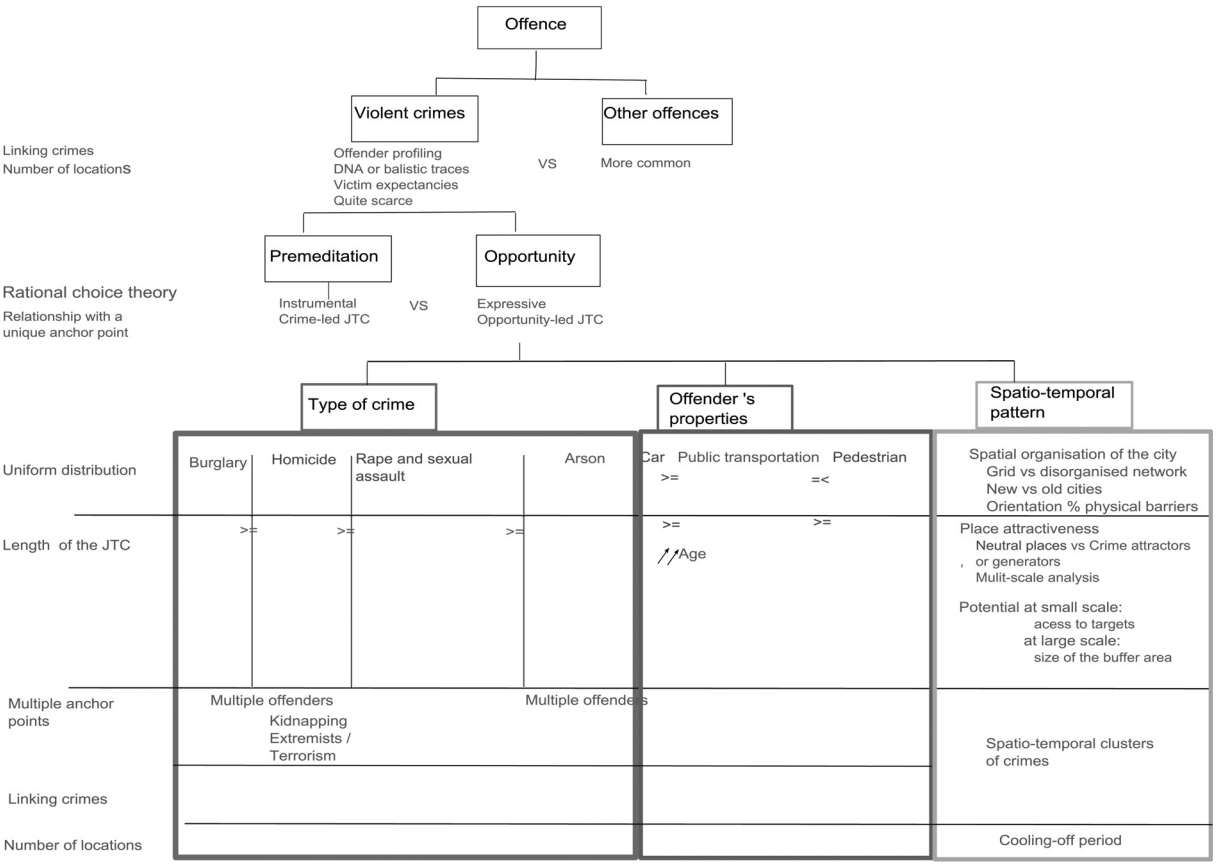


Figure 2: A decision tree to determine the optimal conditions for building a geographic profile.

REFERENCES

Alison, L., Bennell, C., Mokros, A., & Ormerod, D. (2002). The personality paradox in offender profiling: A theoretical review of the processes involved in deriving background characteristics from crime scene actions. *Psychology, Public Policy, and Law*, 8(1), 115-135.

Alston, J. (1994). The Serial Rapist's Spatial Pattern of Target Selection. Master Thesis Simon Fraser University, Barnaby.

Alston, J. (2001). The Serial Rapist's Spatial Pattern of victim selection. . In M. G. Godwin (Ed.), *Criminal Psychology and Forensic Technology: A Collaborative Approach to Effective Profiling* (pp. 231-249). Boca Raton: CRC Press.

Beauregard, E., Proulx, J., & Rossmo, D. K. (2005). Spatial patterns of sex offenders: Theoretical, empirical, and practical issues. *Aggression and Violent Behavior*, 10(5), 579-603.

Beauregard, E., Rossmo, D., & Proulx, J. (2007). A Descriptive Model of the Hunting Process of Serial Sex Offenders: A Rational Choice Perspective. *Journal of Family Violence*, 22(6), 449-463. doi: 10.1007/s10896-007-9101-3

Bennell, C., & Canter, D. V. (2002). Linking commercial burglaries by modus operandi: tests using regression and ROC analysis. *Science & Justice*, 42, 153-164.

Bennell, C., & Corey, S. (2007). Geographic Profiling of Terrorist Attacks. In R. N. Kocsis (Ed.), *Criminal Profiling* (pp. 189-203): Humana Press.

Bennell, C., & Jones, N. J. (2005). Between a ROC and a hard place: a method for linking serial burglaries by modus operandi. *Journal of Investigative Psychology and Offender Profiling*, 2, 23-41.

Bernasco, W. (2010). Modeling Micro-Level Crime Location Choice: Application of the Discrete Choice Framework to Crime at Places. *Journal of Quantitative Criminology*, 26(1), 113-138. doi: 10.1007/s10940-009-9086-6

Bernasco, W., & Block, R. (2011). Robberies in Chicago: A Block-Level Analysis of the Influence of Crime Generators, Crime Attractors, and Offender Anchor Points. *Journal of Research in Crime and Delinquency*, 48(1), 33-57. doi: 10.1177/0022427810384135

Brantingham, P.J, & Brantingham, P. L (1981a*). Environmental Criminology*: Waveland Press.

Brantingham, P. L., & Brantingham, P. J. (1981b). Notes on the Geometry of Crime. In B. P.L. & B. P.J. (Eds.), *Environmental Criminology* (pp. 27-54). Beverly Hills: Sage.

Brantingham, P. J., & Brantingham, P. L. (1990). *Environmental Criminology*, . Long Grove, Illinois,: Waveland Press.

Brantingham, P. J., & Brantingham, P. L. (2003). Anticipating the displacement of crime using the principles of environmental criminology. In M. J. Smith & D. B. Cornish (Eds.), *Theory for Practice in Situational Crime Prevention* (Vol. 16, pp. 119-148). Devon, UK: Willan Publishing.

Brantingham, P.L & Brantingham, P.J (2008). Crime pattern theory. In R. Wortley & L. Mazerolle (Eds.), *Environmental Criminology and Crime Analysis* (pp. 78-93). Cullompton, Devon Willan Publishing.

Canter, D. (1994). *Criminal shadows: Inside the mind of the serial killer*. London: Harper Collins.

Canter, D. (2005). Confusing operational predicaments and cognitive explorations: comments on Rossmo and Snook et al. *Applied Cognitive Psychology*, 19(5), 663-668.

Canter, D.. (2011). Resolving the offender "profiling equations" and the emergence of an investigative psychology. *Current Directions in Psychological Science*, 20(1), 5-10.

Canter, D., & Larkin, P. (1993). The Environmental Range of Serial Rapists. *Journal of Environmental Psychology*, 13, 63-69.

Canter, D., & Youngs, D. (2008). Geographical Offender Profiling: Origins and Principles In D. Canter & D. Youngs (Eds.), *Principles of Geographical Offender Profiling* (pp. 1-18). Burlington, USA: Ashgate Publishing Company.

Capone, D., & Nichols, W. (1975). Crime and distance: An analysis of offender behaviour in space. *Proceedings of the Association of American Geographers*, 45-49.

Chainey, S., & Ratcliffe, J. (2005). *GIS and crime mapping*. West Sussex: John Wiley & Sons

Christaller, W. (1966). *Central Places in Southern Germany*. Translate by C. Baskin. *Die zentralen Orte in Süddeutschland. Eine ökonomisch-geographische Untersuchung über die Gesetzmäbigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen Jena*; Fischer Verlag ; Dissertation ; 1933.: Prentice Hall.

Cohen, L., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44, 588-608.

Cressie, N. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems,* 3, 159-180.

Douglas, J. E. (1992). *Crime classification manual*. New York, Toronto: Lexington Books , Maxwell Macmillan Canada, Maxwell Macmillan International.

Douglas, J. E., Burgess, A. W., Burgess, A. G., & Ressler, R. K. (1992*). Pocket Guide to the Crime Classification Manual . A Standard System for Investigating and Classifying Violent Crimes*. New York Lexington Books

Elffers, H. (2004). Decision models underlying the journey to crime. In G. Bruinsma, H. Elffers & J. W. de Keijser (Eds.), *Punishment, Places and Perpetrators. Developments in Criminology and Criminal Justice Research* (pp. 182–197). Cullompton: Willan

Publishing.
Fritzon, K. (2001). An examination of the relationship between distance travelled and motivational aspects of firesetting behaviour [doi: DOI: 10.1006/jevp.2000.0197]. *Journal of Environmental Psychology*, 21(1), 45-60.
Godwin, M., & Canter, D. (1997). Encounter and Death - The Spatial Behavior of US Serial Killers. Policing: *An International Journal of Police Strategies & Management*, 20, 24-24.
Goodwill, A. M., & Alison, L. J. (2006). The development of a filter model for prioritising suspects in burglary offences. [doi: 10.1080/10683160500056945]. *Psychology, Crime & Law*, 12(4), 395-416. doi: 10.1080/10683160500056945
Grubin, D., Kelly, P., & Brundson, C. (2001). *Linking serious sexual assaults through behaviour* (Vol. 215). London: Home Office Research Study.
Hazelwood, R. R., & Warren, J. (2000). The sexually violent offender - Impulsive or ritualistic? [doi:10.1016/S1359-1789(99)00002-6]. Aggression and Violent Behavior, 5, 267-279.
Holmes, R., & De Burger, J. (1988). *Serial murder: studies in crime law and justice*. Newbury Park, CA.
Knabe-Nicol, S., & Alison, L. (2011). The cognitive expertise of Geographic Profilers. In L. Alison & S. L. Rainbow (Eds.), *Professionalizing Offender Profiling: Forensic and Investigative Psychology in Practice* (pp. 296). London & New York: Routledge. Taylor & Francis group.
Laukkanen, M., & Santtila, P. (2006). Predicting the residential location of a serial commercial robber. [doi: DOI: 10.1016/j.forsciint.2005.03.020]. *Forensic Science International*, 157(1), 71-82.
Laukkanen, M., Santtila, P., Jern, P., & Sandnabba, K. (2008). Predicting offender home location in urban burglary series. Forensic science international, 176(2-3), 224-235.
Levine, N. (2004). *CrimeStat III: A spatial statistics program for the analysis of crime incident locations. Version 3.0*. Houston, TX, and Washington, DC: Ned Levine and Associates and the National Institute of Justice.
Levine, N., & Block, R. (2011). Bayesian Journey-to-Crime Estimation: An Improvement in Geographic Profiling Methodology. [doi: 10.1080/00330124.2010.547152]. *The Professional Geographer,* 63(2). doi: 10.1080/00330124.2010.547152
Levine, N., & Lee, P. (2009). Bayesian journey-to-crime modelling of juvenile and adult offenders by gender in Manchester. [References]. *Journal of Investigative Psychology and Offender Profiling*, Vol.6(3), 237-251.
Lundrigan, S., & Canter, D. (2001). A multivariate analysis of serial murderers' disposal site location choice. [doi: DOI: 10.1006/jevp.2001.0231*]. Journal of Environmental Psychology*, 21(4), 423-432.
Markson, L., Woodhams, J., & Bond, J. W. (2010). Linking serial residential burglary: comparing the utility of modus operandi behaviours, geographical proximity, and temporal proximity. *Journal of Investigative Psychology and Offender Profiling*, 7(2), 91-107.
Openshaw, w. (1984). The modifiable areal unit problem *Concepts and Techniques in Modern Geography* (Vol. 28). Norwich: Geo Books.
Paulsen, D. (2007). Improving Geographic Profiling through Commuter/Marauder Prediction. *Police Practice and Research,* 8(4), 347 - 357.
Phillips, P. (1980). Characteristics and typology of the journey to crime. In D. E. Georges-Abeyie & K. D. Harries (Eds.), *Crime: A spatial perspective* (pp. 167-180). New York: Columbia University Press.
Ratcliffe, J. (2002). Aoristic Signatures and the Spatio-Temporal Analysis of High Volume Crime Patterns. *Journal of Quantitative Criminology*, 18(1), 23-43.
Ratcliffe, J. (2010). Crime Mapping: Spatial and Temporal Challenges. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology* (pp. 5-24): Springer New York.
Rengert, G. F., & Lockwood, B. (2009). Geographical Units of Analysis and the Analysis of Crime. In D. Weisburd, W. Bernasco & G. J. N. Bruinsma (Eds.), *Putting Crime in its Place* (pp. 109-122): Springer New York.
Rengert, G. F., Piquero, A. R., & Jones, P. R. (1999). Distance decay re-examined. *Criminology,* 37(2), 427-446. doi: 10.1111/j.1745-9125.1999.tb00492.x
Ressler, R. K., Burgess, A. W., & Douglas, J. E. (1988*). Sexual homicide: patterns and motives*. New York
Rhodes, W. M., & Conly , C. (1981). Crime and Mobility - An Empirical Study In P. J. Brantingham & P. L. Brantingham (Eds.), *Environmental Criminology* (pp. 167-188). Prospect Heights: Waveland Press Inc.
Rossmo, D. K. (1997). Geographic Profiling. In J. L. Jackson & D. A. Bekerian (Eds.), *Offender Profiling: Theory, Research and Practice*.: Wiley and Sons.
Rossmo, K. (2000). *Geographic profiling*. Boca Raton.: CRC Press.
Rossmo, K., & Velarde, L. (2008). Geographic Profiling Analysis: Principles, Methods and Applications. In L. T. Spencer Chainey (Ed.), *Crime Mapping Case Studies* (pp. 33-43).
Sampson, R. J. (1983). Structural Density and Criminal Victimization. Criminology, 21(2), 276-293. doi: 10.1111/j.1745-9125.1983.tb00262.x
Santtila, P., Laukkanen, M., Zappala, A., & Bosco, D. (2008). Distance travelled and offence characteristics in homicide, rape, and robbery against business. *Legal and Criminological Psychology*, Vol.13(12), Sep 2008, pp.
Snook, B. (2004). Individual differences in distance travelled by serial burglars. *Journal of Investigative Psychology and Offender Profiling*, 1(1), 53-66.
Snook, B., Cullen, R. M., Mokros, A., & Harbort, S.

(2005). Serial murderers' spatial decisions: factors that influence crime location choice. *Journal of Investigative Psychology and Offender Profiling*, 2(3), 147-164.

Tonkin, M., Grant, T., & Bond, J. W. (2008). To link or not to link: a test of the case linkage principles using serial car theft data. Journal of Investigative Psychology and Offender Profiling, 5(1-2), 59-77. doi: 10.1002/jip.74

Tonkin, M., Woodhams, J., Bond, J. W., & Loe, T. (2010). A theoretical and practical test of geographical profiling with serial vehicle theft in a U.K. context. *Behavioral Sciences & the Law*, 28(3), 442-460.

Trotta, M. (2012). New hypotheses on serial offender's spatial behaviour Paper presented at the 1 st AGILE Phd School 2012, Wernigerode.

Trotta, M., Bidaine, B., & Donnay, J.-P. (2011). Determining the Geographical Origin of a Serial Offender Considering the Temporal Uncertainty of the Recorded Crime Data. Paper presented at the GEOProcessing 2011 : The Third International Conference on Advanced Geographic Information Systems, Applications, and Services, Gosier.

Turton, I., Openshaw, S., Brunsdon, C., Turner, A., & Macgill, J. (2000). Testing space-time and more complex hyperspace geographical analysis tools. In P. Atkinson & D. Martin (Eds*.), GIS and Geocomputation* (pp. 87-102): Taylor and Francis.

Unwin, D. J. (1996). GIS, spatial analysis and spatial statistics. *Progress in Human Geography*, 20(4), 540-551.

National Center for the Analysis of Violent Crime. (2008). Serial Murder: Multi-Disciplinary Perspectives for Investigators. In R. J. Morton (Ed.).

Warren, J., Reboussin, R., & Hazelwood, R. R. (1995). *The geographic and temporal sequencing of serial rape.* Washington: DC: Government Printing Office.

Weisburd, D., Bruinsma, G. J. N., & Bernasco, W. (2009). Units of Analysis in Geographic Criminology: Historical Development, Critical Issues, and Open Questions. . In D. Weisburd, W. Bernasco & G. J. N. Bruinsma (Eds.), *Putting Crime in its Place* (pp. 3-31): Springer New York.

White, R. C. (1932). The Relation of Felonies to Environmental Factors in Indianapolis. *Social Forces*, 10(4), 498-509.

Wortley, R. (2008). Situational precipitators of crime In R. Wortley & L. Mazerolle (Eds.), *Environmental Criminology and Crime Analysis* (pp. 294). Devon: Willan Publishing.

Coordonnées des auteurs :

Marie TROTTA
Research Fellow F.R.S – FNRS
Unité de Géomatique, Université de Liège
Allée du 6 août 17, B5a, 4000 Liège, Belgium
Marie.Trotta @ulg.ac.be

André LEMAÎTRE
Institut des Sciences humaines et sociales / Criminologie, Université de Liège
Boulevard du Rectorat 7, B31, 4000 Liège, Belgium.
alemaitre@ulg.ac.be

Jean-Paul DONNAY
Unité de Géomatique, Université de Liège
Allée du 6 août 17, B5a, 4000 Liège, Belgium
jp.donnay@ulg.ac.be

BSGLg, 60, 2013, 59-72

ENHANCING THE DESIGN OF OBSERVATIONAL STUDIES OF COMMUNITY POLICING: USING GEOSPATIAL DATA MINING TO DESIGN NON-EXPERIMENTAL PROGRAM EVALUATIONS

Christian KREIS

Abstract
The current research is an application of geospatial data mining algorithms to enhance the validity of an observational study of community policing in Switzerland's major urban areas. Both unsupervised and supervised data mining algorithms are used to cluster high-dimensional data on neighbourhood-level crime rates, the socio-economic and demographic structure, and the built environment in order to identify matched comparison areas across the five cities for the subsequent impact evaluation. The resulting neighbourhood typology reduced the within-cluster variance of the contextual variables and accounted for a significant share of the between-cluster variance in the survey measures of community policing impact. This suggests that geo-computational methods help to balance the observed covariates and hence to reduce threats to the internal validity of a non-experimental research design. The assessment of the validity of the neighbourhood classification system for evaluation purposes and its geo-visualization for better communication with practitioners and intelligence-based decision making form an integral part of the study.

Keywords
neighbourhood profiling, impact evaluation, machine learning, geospatial data mining

Résumé
La présente étude est une application d'algorithmes d'exploration de données géo-spatiales afin d'améliorer la validité d'une étude observationnelle de la police de proximité dans les plus grands centres urbains de Suisse. Des algorithmes supervisés et non supervisés ont été utilisés sur les données à haute dimensionnalité, relatives à la criminalité à l'échelle des quartiers, à la structure socio-économique et démographique et au cadre bâti. Le but poursuivi est d'identifier des zones comparables à travers les cinq villes étudiées afin d'analyser les impacts de la police de proximité. La typologie de quartier développée a abouti à une réduction de la variance intra-groupe des variables contextuelles. Elle permet d'expliquer une partie significative de la variance intergroupe des indicateurs d'impacts de la police de proximité recueillis par le biais de sondages. Ceci semble suggérer que les méthodes de géo-informatique aident à équilibrer les co-variables observées et donc à réduire les menaces relatives à la validité interne d'un concept de recherche non-expérimental. L'analyse de la validité de la typologie des quartiers à des fins d'évaluation ainsi que sa géo-visualisation pour une meilleure communication des résultats aux praticiens et l'aide à la décision stratégique font partie intégrante de l'étude.

Mots-clés:
criminologie environnementale, prise de décision, inférence, auteurs en série, profilage géographique, analyse spatio-temporelle

I. INTRODUCTION:

A recurrent demand in recent years in the area of crime prevention has been that programs be "evidence-based", meaning that criminal justice policies should be subjected to scientific evaluation in order to identify best practices (Sherman et al., 2002). The methodological standards of scientific program evaluation in general have been cogently defined already during the 1960s and 1970s, and have been reaffirmed more recently with particular reference to criminological interventions (Cook & Campbell, 1979; Shadish et al., 2002; Farrington, 2003). This body of knowledge posits a clear hierarchy of the methodological quality of different research designs, with the randomized controlled trial (RCT) held as the "gold standard" of

scientific evaluation.
In the areas of crime prevention and policing, however, RCT designs have seldom been implemented because field experiments are deemed as politically risky, or ethically questionable, or both. For area-based criminological interventions targeted at specific places or entire jurisdiction, finding a sufficient number of treatment and control areas can be challenging and statistical power correspondingly low. As a result, several authors have bemoaned a dearth of methodo¬logically sound program evaluations (e.g. Weisburd & Eck, 2004; Welsh & Hoshi, 2002).
An observational study is the alternative empirical analysis of the effects of a treatment intervention in cases where an experimental design is either unethical or infeasible. A good observational study strives to emulate the key aspects of a RCT design in order to enhance the validity of its conclusions. Crucially, in a true experiment, the distribution of covariates is similar between treatment and control group as a result of the random assignment of the study objects to the treatment and control condition. An observational study seeks to achieve this by selecting a set of com¬parisons that resemble the treated objects on the observable covariates prior to the treatment interven¬tion. Matching techniques are then used to achieve a similar distribution of the observed covariates (though not the unobserved covariates) between the treated objects and the selected controls (Rosenbaum, 2010).
The current research forms part of an observational study of community policing in major Swiss urban areas (Kreis, 2012). Community policing is both a philosophy and an organizational strategy of the police that promotes a renewed partnership between the police agency and local communities to solve prob¬lems of crime and disorder. For the current study, the selection of suitable treatment and control areas for the planned evaluation was compounded by the fact that police forces in Swiss cities, beginning in the late 1990s, rapidly introduced community policing across their entire jurisdiction without making any provisions for later evaluation. In the current context this meant that any valid control group had to be found outside each urban area and baseline data for any pre-test/post-test comparisons had to come from existing data sources.
The exploratory data analysis, which had been under¬taken as a preliminary study (Kreis, 2009), revealed that the spatiotemporal patterns of the four theoretical constructs of community policing impact – crime, fear of crime, neighbourhood disorder and public attitudes towards the police – displayed some remarkable parallels across the five urban areas. In particular, the exploratory spatial analyses established that the patterns of crime rates and perceptions of disorder had remained rather stable over the short and medium run, whereas areas with elevated levels of fear had shifted from the urban centres to the city boundaries between the late 1980s and 2005. Moreover, whereas these observable response patterns were noticeably different between the Swiss German and Swiss French cities, responses within a given language region proved to be unexpectedly homogenous (Kreis, 2012).
These rather systematic spatiotemporal patterns of the outcome indicators implied that an impact evaluation of community policing over an extended study period that did not control for shifting neighbourhood charac¬teristics, risks being unreliable at best and positively misleading at worst. The striking parallels between the cities under study gave rise to the idea of developing a neighbourhood typology to match similar neighbour¬hoods across urban areas in order to study the impact of different community policing strategies in similar neighbourhood contexts. The objective is thus to develop a classification system in order to group neighbourhood areas into clusters of similar type based on a series of environmental, socio-economic and socio-demographic indicators. This approach is based on the premise that the spatial dynamic of the socio-economic processes unfolding in a city affect the crime and response patterns to a considerable extent and that these processes would repeat themselves from one city to another.

II. THEORETICAL CONSIDERATIONS

A. The rationale of matching neighbourhood areas for performance evaluation

The classification or profiling of neighbourhoods or bigger administrative areas in the field of law enforcement and policing has been tried and applied primarily in England and Wales in an effort to increase police accountability and to set benchmarks to measure and improve police performance. In the mid-1990s, the government police inspectorate thus created the *most similar force* group that assigned all 43 separate police forces into groupings of similar type (Ashby & Longley, 2005, 56f.). In the early 2000s, the British Home Office published similar groupings of smaller scale police administrative units, which classified the more than 300 Crime and Disorder Reduction Partnerships (CDRP) and Basic Command Units (BCU) across England and Wales into *families* of similar type (Sheldon et al., 2002). The rationale behind the clustering of policing units was to identify areas that faced similar *policing environments* and were thus suited for meaningful cross-sectional comparisons to evaluate performance (Ashby & Longley, 2005, 56f.).
The classification and matching of police administrative units for performance evaluation is based on the premise that different areas differ significantly in their responsiveness to different policing styles. It rests on the observation that even though crime and poverty are correlated, not all deprived areas are equally crime-ridden. The clustering to categorize the different poli-

cing areas therefore must include not only ecological characteristics of an area such as the socio-economic status (SES) or demographic composition but data on attitudes and lifestyles as well. The very importance of such *soft* attitudinal aspects has been underlined by analyses of the British Crime Survey (BCS) data from the 1990s, which showed that even though actual levels of victimization had been falling, two thirds of respondents were under the impression that crime had gone up. This apparent mismatch between falling crime levels and the widely held belief of an increase in crime has become known in the literature as the *reassurance gap* and has spawned *reassurance policing*, which aims both to rectify the public's perception and ultimately to provide safer neighbourhoods (Williamson et al., 2006, 191-4).

Linking British census and BCS data, Williamson et al. (2006) developed an indicator of an area's level of social capital and compared these values to actual victimization rates. Their results showed not only that areas with higher levels of social capital suffered comparatively lower levels of victimization but also that people's perception of crime, crime reporting, fear of crime and attitudes towards the police differed according to the composition of their neighbourhood. The authors thus concluded that such geo-demographic profiling serves as a useful tool to design reassurance policing or community policing strategies, which are more likely to be effective if targeted specifically at the needs of each type of neighbourhood area.

Ashby and Longley (2005, 427-32) proposed three kinds of geo-demographic analyses that may support police strategic decision-making and performance evaluation: area profiling, operational data profiling and crime survey profiling. Firstly, the basic profiling of police patrol beats or precincts into neighbourhoods of different types provides basic strategic intelligence. Mapping such a typology in a GIS provides an additional spatial dimension to this kind of information. Secondly, the profiling of crime events and police operational data allow police analysts to compute the propensities of specific crime events in different neighbourhood types. Such information makes it possible to identify areas with unexpectedly high or low levels of victimization or to assess the effects of targeted policing interventions. Finally, adding survey data to the analysis helps unearth likely variations in popular attitudes to disorder, fear of crime and the police across different neighbourhood types. If the place of residence of each respondent is known, survey data can be pooled by neighbourhood type to calculate area-level, regional, or even national scores, which may then be extrapolated for analysis at the local level.

The current study builds on and tests these theoretical arguments for a comparative evaluation of community policing across Switzerland's five biggest urban areas. In order to match similar neighbourhoods across the five cities, the current study aims to create a classification system of urban neighbourhoods, which in few dimensions aptly describes the spatiotemporal patterns observed in the high-dimensional input data. The idea is to develop a neighbourhood typology based on a series of demographic, socio-economic and environmental indicators as well as survey data in order to classify the urban neighbourhoods within the study area into clusters of similar type.

This process of dimensionality reduction and clustering of the high-dimensional attribute data serves to find matching pairs of treatment and control districts in order to enhance the validity of an observational study of a complex intervention across multiple sites. The clustering procedure to develop this neighbourhood typology thus has a double objective: on the one hand, the algorithm should reduce the within-cluster variance of the neighbourhood ecological variables, which may be correlated with the outcome variables and thus risk confounding inferences about program impact in a non-experimental research design. Put differently, neighbourhoods that resemble each other in terms of their demographic and socio-economic structure as well as the built environment must be grouped into clusters of similar type. On the other hand, the resulting neighbourhood typology should account for a maximum of the between-cluster variance in the outcome indicators prior to program implementation, i.e. the survey response patterns should be similar for residents of a given neighbourhood type across urban areas. In other words, the goal was to match the comparison areas not only on the observed covariates describing the neighbourhood context but on the outcome variables targeted by community policing such as fear of crime, neighbourhood disorder and satisfaction with police as well, i.e. the typology should capture a maximum of the *neighbourhood effects* of the different neighbourhood types. An observational study based on a neighbourhood typology that meets both these objectives allows an evaluator to dismiss a series of threats to the internal validity that otherwise beset a non-experimental research design.

B. The methodology of matching neighbourhood districts

The Home Office researchers who developed the BCU and CDRP families across England and Wales pre-selected 20 variables form the British census describing the demographic, socio-economic and the built environment characteristics of these areas based on their correlation with area levels of crime and disorder. As clustering algorithm they then employed k-means and self-organizing maps (SOM) in order to develop a typology of BCUs and CDRPs that minimized the variance in crime rates within a given family (Harper et al., 2002).

Outside criminology, two more recent studies used artificial neural networks and data mining procedures

to develop typologies of areal units of analysis. Li and Shanmuganathan (2007) used the SOM algorithm for a clustering of 90 demographic and socio-economic variables for a social area analysis to classify 163 census tracts in a small-sized city in western Japan. Spielman and Thill (2008) used self-organizing maps for a geodemographic classification of 2,217 census tracts in New York City, using a dataset of 79 attributes from the U.S. Census.

The current study uses both unsupervised and supervised data mining algorithms to develop the neighbourhood typology. During the unsupervised learning phase, self-organizing maps are being used to cluster a high-dimensional data set in order to classify the neighbourhood areas into clusters of similar type (Skupin & Agarwal, 2008; Vesanto & Alhoniemi, 2000). During the following supervised learning phase, the random forests algorithm (Breiman, 2001) is used to select the most important features in order to develop a parsimonious model that makes a minimum of classification errors. In addition, the random forests algorithm serves as a gauge of the performance of the clustering algorithm overall as well as the predictive power of individual variables in the training data set.
As a final step, the resulting neighbourhood typology is to be visualized in a GIS as a map in the original geographic space, indicating the location of areas of a given type that are thus suited for matching and comparing during the subsequent impact evaluation.

III. DATA AND METHODOLOGY

The data analysed in the evaluation of community policing across Swiss urban areas come from three main sources: (a) official police crime statistics on area level crime rates, (b) the 1990 and 2000 Swiss population and housing census on the demographic composition and socio-economic status of the resident population as well as the structure of the built environment; and (c) the Swiss Crime Survey (SCS), a large-scale longitudinal criminal victimization survey on fear of crime, perceptions of neighbourhood disorder and popular attitudes towards the police. The SCS sampled the five urban areas under study repeatedly between 1998 and 2005. All data were measured at the level of postal ZIP code or administrative districts within the five urban areas, which are the smallest spatial unit of analysis for which all three data categories are available.

A. Unsupervised learning – Self-organizing maps

In geospatial data mining, unsupervised learning algorithms serve as analytical and modelling tools to discover patterns or structures in the data in order to classify study objects with (dis-)similar features in attribute or in geographic space (Kanevski et al., 2009). During the unsupervised learning phase, the current study uses self-organizing maps (SOM; Kohonen, 1990, 2001) as modelling tools to identify the underlying spatiotemporal patterns in the neighbourhood ecological data.
As a dimensionality reduction algorithm, the SOM is analogous to a discrete non-linear Principal Components Analysis (PCA). A PCA fits a hyper-plane into the data cloud that minimizes the distance to the original data points in order to replace the original variables by a smaller number of uncorrelated principal components. In the SOM algorithm, a network or *lattice* of artificial neurons is introduced into the input space instead of a hyper-plane. The segments of the SOM lattice are flexible and highly elastic and thus fit easily over curve-linear or unevenly distributed data during the training phase, a process which has been likened to covering the cloud of original data points with an *elastic fishing net* (Lee & Verleysen, 2007, 136).
After the training is completed, the artificial neurons or prototype vectors of the SOM lattice are aggregated by hierarchical agglomerative clustering (Skupin & Agarwal, 2008; Vesanto & Alhoniemi, 2000). As explained below, the resulting dendrogram is then cut at the level that produces the partitioning of the prototype vectors that best meets the twin objectives of the clustering algorithm. Finally, the original data points representing the urban neighbourhoods are assigned to the class of the nearest prototype vector, to which they have been attached during the training of the SOM lattice (Figure 1).

B. Supervised learning – Random forests

Supervised data mining algorithms seek structures or patterns in the data that explain or predict a priori information on outcomes or classifications (Kanevski et al., 2009). During supervised learning, the neighbourhood classification system that resulted from the unsupervised training is used as a priori information or labels of the training data. In this phase, the random forests algorithm (Breiman, 2001) serves to develop a classifier that aims to predict each neighbourhood's class based on the same explanatory variables that were first used during the unsupervised learning.
In the current context, supervised learning can be likened to non-linear multivariate logistic regression with the neighbourhood classification being the dependent variable. As a logistic regression equation predicts a value or class of the dependent variable, the random forests classifier contains a decision rule that assigns each neighbourhood area to the cluster to which it belongs based on its values on the explanatory or training variables.
The purpose of the supervised learning is on the one hand to assess the validity of the neighbourhood typology by analysing the overall classification error rate of the decision rule. As a by-product of the training of the classifier, the random forests algorithm produces a

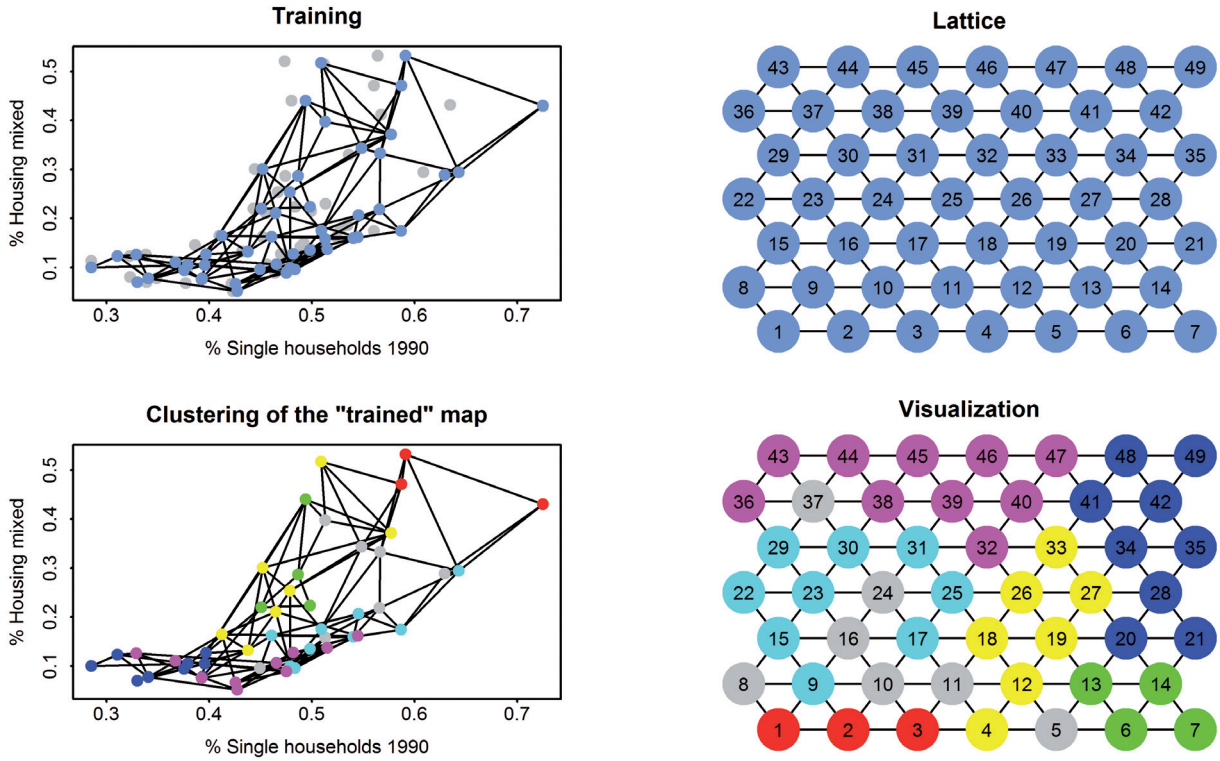


Figure 1. The SOM algorithm visually explained. During training, the network of prototype vectors (blue) is iteratively fitted over the original data points (top left). The structure of the SOM lattice in projected 2-D output space (top right). Following training, the prototype vectors are clustered and labelled according to their topological position inside the lattice (bottom left). Visualization of the partitioning of the lattice in output space (bottom right). NOTE: Training and clustering occur in high-dimensional input space; for simplicity, only 2 variables are plotted here.

proximity measure that indicates the probability that any two neighbourhoods will be assigned to the same cluster based on their explanatory variables and thus how closely they resemble each other in the original input space (Breiman, 2001; Liaw & Wiener, 2002, 18f.). As a general rule, the greater this distance is for neighbourhoods belonging to different classes, the more distinct the neighbourhood clusters really are and the fewer classification errors the random forests classifier will make.

On the other hand, the random forests algorithm serves to weed out noisy variables with no or little predictive power in order to come up with a parsimonious model that makes a minimum of classification errors. As a second analytical tool the random forests algorithm computes a variable importance measure, which indicates the predictive value of each explanatory variable. Not unlike the *p*-value of a coefficient in multivariate regression, the variable importance measure can be used for feature selection to remove noisy or unimportant explanatory variables from the training data set (Breiman, 2001, 23f.; Genuer et al., 2010, 2226, 2229). This procedure is quasi analogous to a stepwise regression procedure which recursively removes non-significant variables from the regression model until only the most pertinent predictors are retained.

C. The neighbourhood clustering procedure

The methodological approach of the current study to develop a typology of Swiss urban neighbourhood areas uses both unsupervised and supervised data mining in an iterative procedure. During unsupervised learning, self-organizing maps are used to detect patterns in the neighbourhood ecological data. After the training of the lattice is completed, hierarchical agglomerative clustering (HAC) serves to merge the SOM prototype vectors (Skupin & Agarwal, 2008; Vesanto & Alhoniemi, 2000). The dendrogram resulting from HAC is then cut at the level that produces a partitioning of the neighbourhood areas that best satisfies the twin optimization criteria of the clustering procedure.

Once the optimum number of clusters has been determined, the trained SOM lattice is divided into as many segments. The SOM prototype vectors are classified depending on their position inside the SOM lattice. The original data points representing the urban districts take on the label of the SOM prototype vector to which they have been attached during training.

During the supervised learning phase, the resulting neighbourhood classification system is used to label the training data set. The random forests algorithm serves to develop a classification rule that assigns each urban

district to a neighbourhood cluster based on the explanatory variables describing the ecological context. The random forest classifier serves both to assess the overall quality of the neighbourhood typology as well as to select the explanatory variables with high predictive value.

The original training data set to characterize the neighbourhood ecology contained 89 variables, which can be regrouped into five distinct categories: official crime rates, population demography, socio-economic status, population heterogeneity and residential stability and the built environment. In a first step, the SOM-random forests algorithms described above were run separately on four of the five categories of ecological variables in order to select the key features of each. The selection criterion for this initial clustering procedure was fairly straightforward: keep as few explanatory variables as necessary without unduly increasing the classification error rate of the random forests decision rule.

The following section first gives a brief account of the preliminary clustering procedures of each variable category, before describing the procedure and outcomes of the final model of the key 24 variables in greater detail.

1. Crime rates

For the four cities for which neighbourhood-level crime data were available for analysis – Bern, Geneva, Lausanne and Zurich – local police statistics have been harmonized and the standardized neighbourhood-level crime rates calculated for eight different types of criminal offenses: homicides, assaults, burglaries, motor vehicle thefts, robberies, vandalism, extortion and threats.

As these standardized neighbourhood-level crime rates of the eight criminal infractions turned out be highly collinear, a principal components analysis was run to replace them by their component scores on the emerging principal components. Only the first two principal components had eigenvalues greater than one and were thus retained for the final neighbourhood clustering analysis, labelled as "Crime PC1" and "Crime PC2".

2. Population composition

The second set of ecological variables describes the demographic composition of the neighbourhood population. It includes the percentage of the total population by age group in 18 categories of 5-year intervals, from "0 to 4 years old" to "85 years old and above" as well as the percentage of both single and family households in the area. Demographic data from both the 1990 and 2000 census were included in the analysis. After the initial clustering procedure, six out of the 40 variables were retained for the final analysis: the percentage of children aged "5 to 9 years old" and "10 to 14 years old" from the 1990 census as well as the percentage of "Single households" and "Families" from both the 1990 and 2000 census.

3. Socio-economic status

The third set of ecological variables describes an area's socio-economic status, measured as the percentage of the resident population by level of the highest educational achievement in seven categories ranging from "Mandatory schooling" to "University" (graduates) from the 1990 and 2000 census. The list of variables also included the percentages of the active working population residing in the area subdivided by eight professional categories of varying social prestige and remuneration, ranging from "Unskilled workers" to "Executives" from the 1990 and 2000 census. Of these 22 SES indicators, six were retained for the final analysis: the percentage of residents who had completed "Mandatory schooling" or an "Apprenticeship" from both the 1990 and 2000 census, as well as the percentage of university graduates and of residents employed in a "Middle management" position from the 2000 census.

4. Heterogeneity and residential stability

The fourth set of ecological variables included five variables measuring an area's degree of population heterogeneity and residential stability: the percentage of Swiss and foreign nationals among neighbourhood residents from the 1990 and 2000 census as well as the percentage of residents who in 2000 still lived at the same address as five years earlier. After the initial clustering, three variables were kept for the final analysis: the percentage of "Foreigners" among the resident population in 1990 and 2000 as well as the variable capturing the percentage of long-term residents in an area.

5. Built environment

The fifth set of ecological variables included in the analysis characterizes the built environment in a given area. A first set of nine variables measures the percentage of buildings of the total housing stock by height, ranging from "1 story" and "2 stories" up to "15 and more stories". A second set of eight variables captures the period of construction of the building units in the area, distinguishing eight different time periods from "before 1900", to "1900-1920" until the most recent period "1986-1990". Three more indicators were included in the analysis measuring the percentage of building units by their functional use, distinguishing between residential buildings, mixed housing complexes that include both apartments and offices or shops and non-residential buildings such as office complexes or commercial centres. All indicators describing the built environment were gathered from the 1990 census only. Seven variables proved important during the initial clustering run and were included in the final analysis: the three indicators on the functional use of buildings ("Residential", "Housing mixed" and "Non-housing"), three

indicators of building height ("2 stories", "6 stories" and "7-9 stories") as well as one variable indicating the construction period ("before 1900").

6. Final Clustering Procedure of the Key Variables

These initial clustering procedures served primarily to select the most important neighbourhood ecological features for each category of data. Based on their results, it was possible to reduce the number of variables in the training data set from 89 (87 ecological variables + 2 principal components of the crime data) to 24 (22+2) key variables, without an undue increase in the classification error rate of the resulting random forests classifier for each of the four data categories.

In a second phase, the identical clustering procedure - the SOM algorithm and hierarchical agglomerative clustering to classify the unlabelled data paired with the random forests algorithm to assess the quality of the clustering and to identify the most important features – was run on the 24 key variables retained for the final model. The diagnostic plots that resulted from this final clustering procedure are shown in Figure 2.

In order to determine the number of clusters that results in the optimum partitioning of the ZIP code or administrative districts, the dendrogram resulting from the hierarchical agglomerative clustering of the SOM prototype vectors of the final model is cut at different levels. For each possible number of neighbourhood clusters $k = 1,2\ldots K$, statistical tests are calculated to determine the optimum partitioning. As has been stated previously, the resulting neighbourhood typology has to reconcile a double objective: the typology should cluster the neighbourhoods that are most similar in terms of their ecological characteristics and simultaneously account for a significant share of the between-cluster variance in the survey measures of community policing impact, namely fear of crime.

Each of the two optimization criteria required a separate test. For the first problem, identifying the optimum number of clusters with regard to the neighbourhood ecological data, some kind of clustering validity index must be applied. In line with other studies (e.g. Vesanto & Alhoniemi, 2000), the current study used the Davies-Bouldin Index (DBI; Davies & Bouldin, 1979) to assess the quality of different partitions of the urban districts across the four cities.

For the second optimization problem, identifying the number of neighbourhood clusters that accounts for the biggest share of the between-cluster variance in the survey data, the Swiss Crime Survey respondents were regrouped according to each possible number of neighbourhood clusters $k = 1,2\ldots K$. The actual statistical test was a $\chi 2$-independence test to determine whether response patterns of the pooled survey sample varied significantly by neighbourhood clusters on the fear of crime survey item.

Figure 2(e) plots the test statistic of both the DBI (left scale) and the p-value of the $\chi 2$-independence test (right scale) as a function of the number of neighbourhood clusters into which the neighbourhood areas are divided across the four urban areas. This chart reveals that there is no unique solution that meets both optimization criteria simultaneously. Regarding the neighbourhood ecological data, a partitioning into merely two clusters would be ideal, as the DBI is at its global minimum for $k = 2$. However, if the neighbourhood areas are divided into just two clusters, the $\chi 2$-independence test

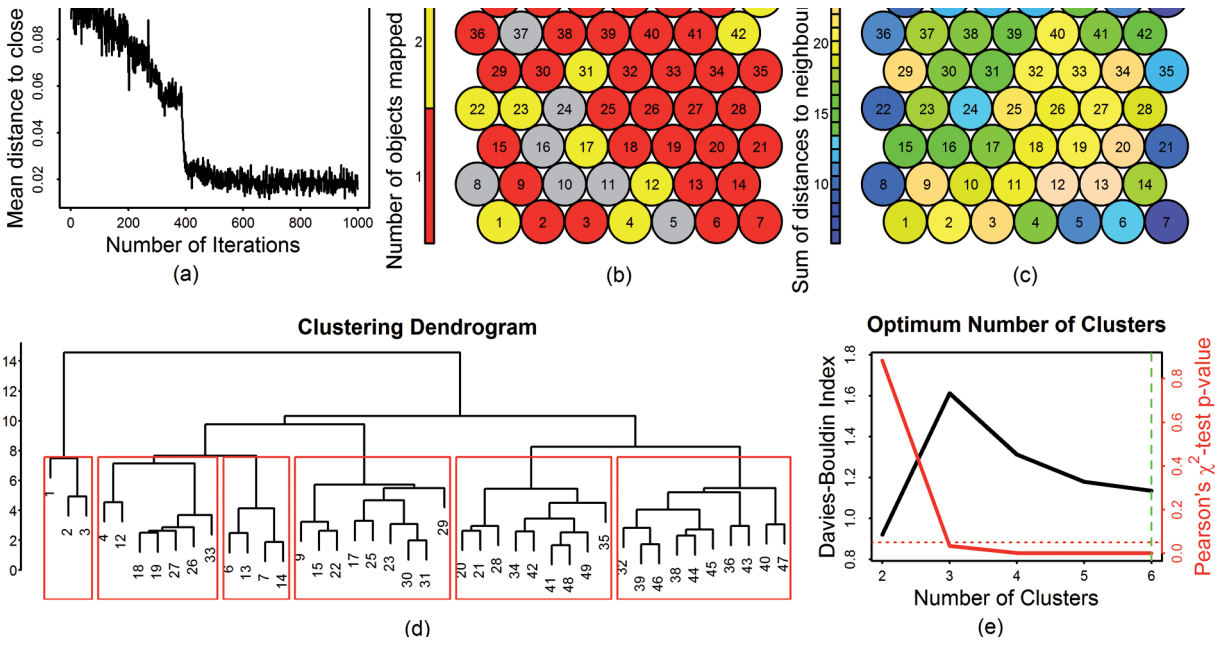


Figure 2. Training and clustering of the self-organizing map

on the survey data is not significant. A neighbourhood classification system that also accounts for a significant share of the between-cluster variance in the survey item requires at least three neighbourhood clusters. It turns out that the partitioning that best reconciles the twin optimization problems is six neighbourhood clusters: with $k = 6$, the $\chi 2$-independence test is significant and the test value of the DBI is at a local minimum.

Since the results of the SOM algorithm depend to some extent on the random initial values, the entire SOM training procedure and determination of the optimum number of clusters described above was replicated 50 times. From among the solutions with a minimum of four clusters, the number of neighbourhood clusters that the algorithm suggested most frequently was six. (A requirement of a minimum of clusters had to imposed since solutions with less than four clusters always pitted the urban centres against the rest of the urban areas, which failed to explain a significant share of the variance of the survey outcome measures). From among those replications that suggested k = 6 as the optimum number of clusters, the replication that resulted in the lowest value of the DBI was retained as the final clustering result and presented in Figure 2.

All computations were made using the R language for statistical computing (R Development Core Team 2012). The SOM and Random Forests algorithms were computed using the R packages 'kohonen' (Wehrens & Buydens, 2007) and 'randomForest' (Liaw & Wiener, 2002). Survey data were analyzed using the svymean and svytable functions of the package 'survey' (Lumley, 2010) and the Davies-Bouldin index was computed using the package 'clusterSim' (Walesiak & Dudek, 2012). Map illustrations and multi-panel plots were created using packages 'maptools' (Bivand et al., 2008) and 'lattice' (Sarkar, 2008), respectively.

IV. RESULTS

Once the optimum partitioning has been determined, the resulting neighbourhood typology can be visualized using different tools. Figure 2(d) plots the dendrogram of the hierarchical clustering of the SOM prototype vectors of the final model, cut at the optimum level $k = 6$. Since the leaves of the dendrogram represent the SOM prototype vectors, the result of the clustering procedure can also be visualized by means of the SOM lattice projected in 2D space with the different segments coloured according to the six neighbourhood types (Figure 3 top right panel). In addition, since each prototype vector represents one or more of the original urban areas, the ZIP code or administrative districts can be shaded according to the same colour scheme and the resulting neighbourhood typology visualized as maps of the five cities using GIS.

A. Geo-visualization of the neighbourhood typology

Figure 3 shows maps of the five urban areas. These maps display some striking parallels between the five cities. First of all, neighbourhood types 1 and 2 are the downtown areas located at or near the city centres. Neighbourhood types 3 and 4 form a first rim around the city centres, whereas neighbourhood types 5 and 6 are suburbs located on the outskirts of the five urban areas. Secondly, the five maps also reveal some striking differences, most notably between the French speaking cities of Lausanne and Geneva on the one hand and the Swiss German cities of Basel, Bern and Zurich on the other. Whereas in the French speaking areas, the suburban neighbourhoods are predominantly blue (type 5), the outskirts of the German speaking cities are either light blue or pink (type 4 and type 6; Kreis, 2012).

Moreover, the spatial pattern of the neighbourhood typology revealed some noteworthy characteristics both of the neighbourhood clusters themselves and the clustering algorithm used to identify them. First of all, in all five urban areas, the different neighbourhood types are neatly aligned on a centre-periphery axis. This pattern is all the more remarkable given that no geographic indicator was included as a variable in the training data used in the clustering algorithm. In other words, the variables included in the training data set on area-level crime rates, population composition, socio-economic status, heterogeneity and residential stability and the built environment display sufficient variation between the more central and the more peripheral areas for these neighbourhoods to cluster neatly into groups of similar type according to their geographic location (Kreis, 2012).

In a similar vein, the resulting spatial pattern of the neighbourhood typology also highlights the topology preserving quality of the SOM clustering algorithm. Topology preservation in a dimensionality-reduction algorithm means that two points that are close to each other in the high-dimensional original input space remain in close proximity of each other in the low-dimensional projected output space (Lee & Verleysen, 2007). This trait of the SOM algorithm is demonstrated neatly by the plot of the trained SOM lattice in the projected 2-D output space displayed in the top right panel of Figure 3. This chart shows the prototype vectors representing the urban centres or type 1 neighbourhoods at opposite ends of the SOM lattice from the suburban neighbourhoods of types 5 and 6, with the SOM prototype vectors representing the neighbourhood types 2, 3 and 4 lying in between. Again, it is noteworthy that this pattern resulted without imposing any geographic reference on the clustering algorithm. In other words, the trained SOM lattice accurately reflects the spatial logic following a centre-periphery axis inherent in the training data. This provides further evidence that the SOM algorithm is well suited for the clustering task at hand (Kreis, 2012).

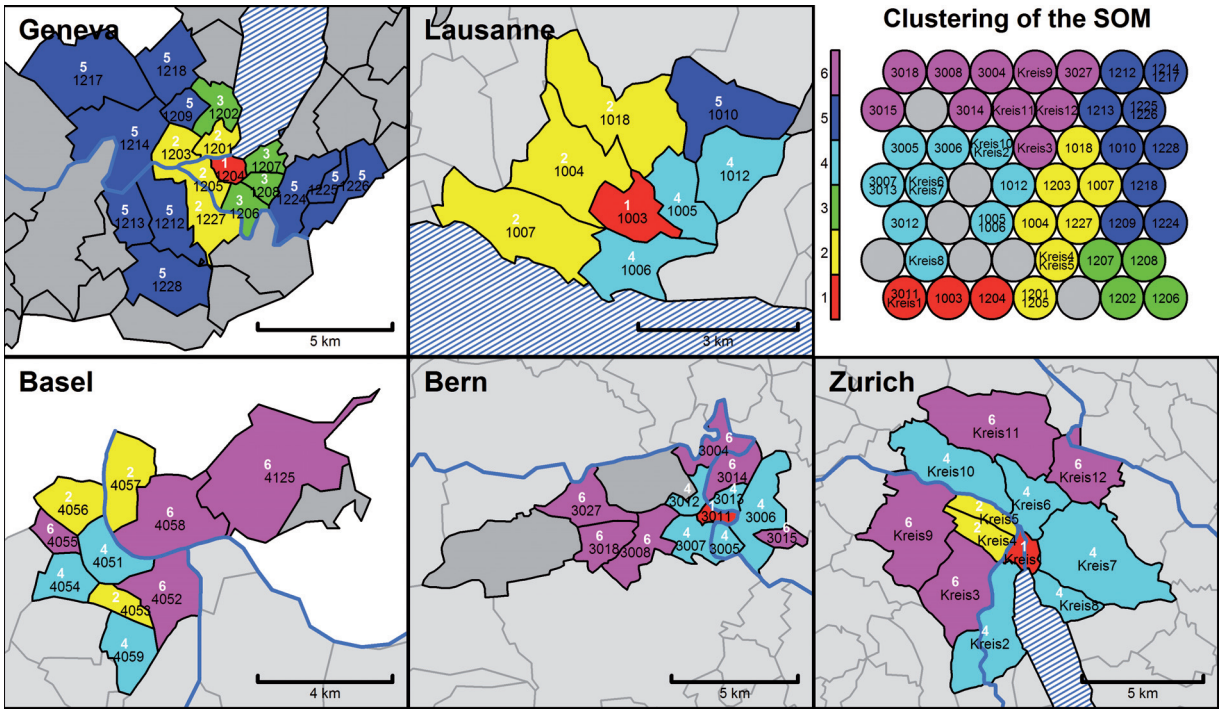


Figure 3. Visualization of the neighbourhood typology in geographic space. Maps of the five cities with neighbourhood areas coloured according to type. The top right panel displays the trained SOM lattice projected into 2-D output space with the segments coloured according to the best partitioning. Neighbourhoods in Basel were labelled based on the random forests classifier trained on the ecological data on the other four cities since for it no neighbourhood-level crime data were available.

B. Assessing the valdity of the neighbourhood typology for evaluation purposes

Once the clustering procedure is completed, the validity of the resulting neighbourhood typology for the planned evaluation of community policing still needs to be assessed. The first optimization criterion for the clustering algorithm was that the neighbourhood typology should minimize within-cluster variance in the neighbourhood contextual variables. In order to assess the clustering algorithm on this score, Figure 4 plots the attribute values of all the 51 neighbourhoods in the data sample. For each of the 24 variables retained in the final clustering algorithm to develop the neighbourhood typology, a panel in Figure 4 displays the original attribute values of the neighbourhoods as boxplots for each of the six neighbourhood types (Kreis, 2012).

The boxplots reveal that the clustering algorithm clearly reduced the within-cluster variance on each of the 24 key variables. To the extent that the boxplots by neighbourhood types are separated, Figure 4 highlights the underlying differences in the attribute between the different neighbourhoods. As was to be expected, not every variable is equally important in distinguishing the six neighbourhood types. Figure 4 thus also serves to identify the salient characteristics of each of the six neighbourhood types that most differentiate it from the other neighbourhood types (Kreis, 2012).

The second and arguably more important criterion of the validity of the neighbourhood typology is whether it accounts for a significant share of the between-cluster variance in the survey variables that serve as indicators of community policing impact. The survey item used as the optimization criterion in the clustering algorithm itself was the question on fear of crime, which thus varies significantly by neighbourhood type. It still has to be determined to what extent the neighbourhood typology also picks up some of the variance of the other survey outcome measures of community policing.

Figure 5 displays a chart for each of the six survey items that were used as the outcome indicators in the evaluation of community policing in Swiss urban areas: three items tapping into survey respondents' fear of crime, two indicators measuring physical and social disorder and one item capturing popular satisfaction with the police. For each of the survey items, the pooled SCS survey sample from all five urban areas was divided into six subgroups according to the neighbourhood type of a respondent's place of residence (type 1 to type 6). The six bar-plots represent the underlying contingency table of the neighbourhood cluster subgroup sample by the answering category to the survey item.

For each of the six survey items, the panels indicate the survey question with the corresponding answering categories. The bar-plots display the percentage of survey respondents by answering category. The single

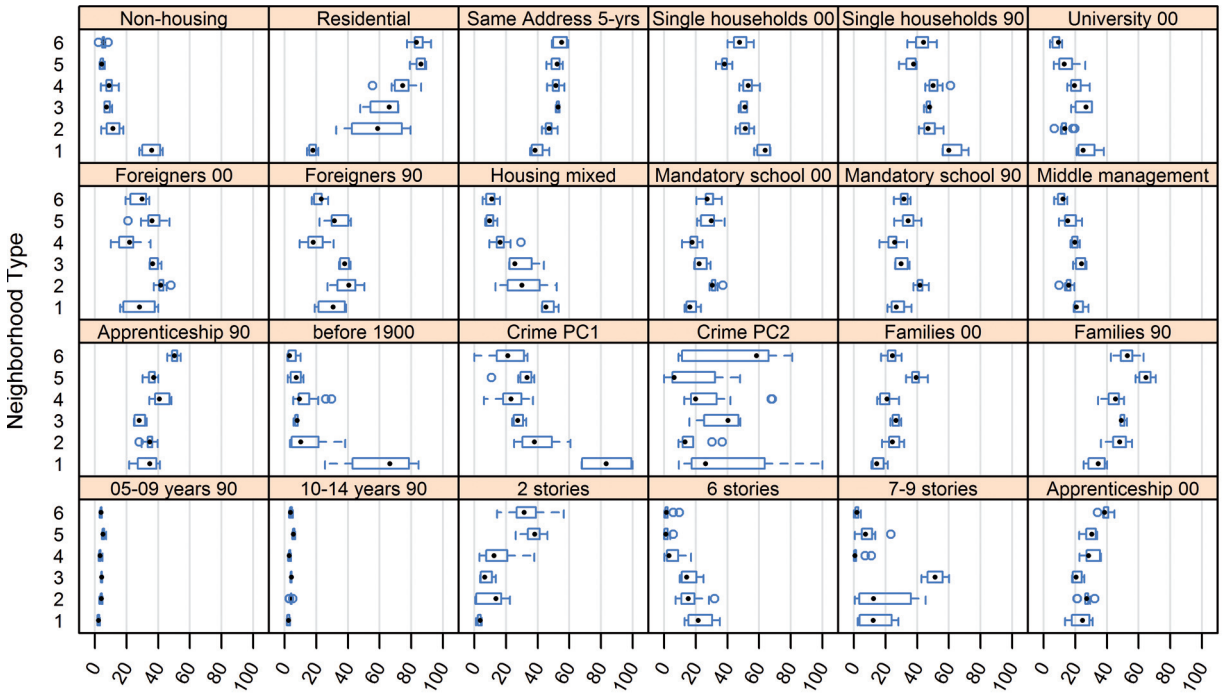


Figure 4. Defining characteristics of the neighbourhood typology. Original values of the 24 variables retained in the final clustering analysis to describe the neighbourhood ecology. All variables are percentages except for the crime principal components scores ("Crime PC1" and "Crime PC2"), whose true range was linearly transformed to a 0-100 scale.

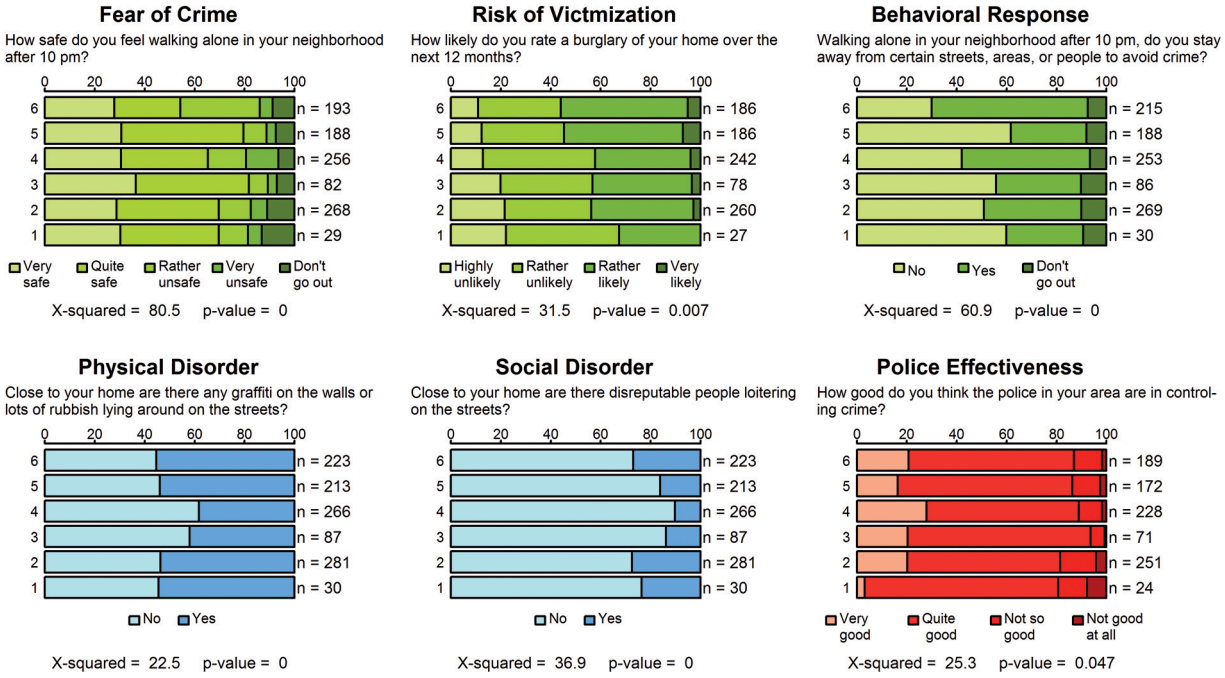


Figure 5. Survey response patterns by neighbourhood type. Percentage of respondents by answer category for the six survey items used to assess the impact of community policing on neighbourhood residents. Survey respondents were grouped together across cities by neighbourhood type, excluding respondents from Basel. The $\chi 2$-independence test statistics were calculated using Monte Carlo simulations. The total survey sample was weighted, stratified at the neighbourhood level, to correct for sampling bias in the age and gender distribution.

digits on the left of each bar-plot indicate the type of neighbourhood cluster, whereas the numbers on the right indicate the size of the subgroup sample of each neighbourhood cluster. The survey data were weighted as a stratified random sample at the ZIP code level to correct for sampling bias in the age and gender distribution to make the neighbourhood-level sub-samples representative of the local resident population. At the bottom of each panel are the test statistics of the $\chi 2$-independence test that was run to determine whether survey response patterns differ significantly by neighbourhood type. These test statistics were calculated by means of a Monte Carlo simulation to circumvent the problem of contingency table cells with an expected frequency of below five. As a matter of fact, the absolute number of survey respondents per answering category is at times rather low, especially for neighbourhood type 1, i.e. the urban centres. Monte Carlo simulations are implemented as a standard option in the chisq.test function in R and were computed on the basis of 2000 replications, which is the standard value R proposes for such simulations.

The contingency tables behind the boxplot charts and corresponding $\chi 2$-independence tests unearthed some very interesting spatial trends in the survey response patterns that deserve a closer inspection. As a matter of fact, the boxplots of the six different neighbourhood types are arranged in a spatial order inside each chart, with the more centrally located neighbourhood clusters (types 1, 2 and 3) being at the bottom and the more peripheral neighbourhood (types 4, 5 and 6) placed on top.

The top left panels shows the bar-plots of the survey item of fear of crime measured as the feeling of safety on a nightly stroll through one's own neighbourhood. As this chart indicates, the percentage of respondents who feel "very safe" or "quite safe" is generally higher in the more centrally located neighbourhoods than in the outskirts. The only notable exceptions to this spatial trend are neighbourhood clusters type 3 and 5, which are predominantly neighbourhoods located in the Geneva urban area (Kreis, 2012).

This general tendency of fear of crime to be more prevalent in the outer areas than in the urban centres is still more pronounced for the survey item asking about the perceived risk of a burglary of one's home. The percentage of respondents who rate the chances of a burglary of their home over the next twelve months as "likely" or "very likely" goes up systematically from neighbourhood type 1 to 6, i.e. as one moves out from the city centres to the outskirts of the five urban areas in the real world (Kreis, 2012).

For burglary, the spatial pattern of the perceived risk can be compared to actual victimization risk as captured by the official police crime statistics. An earlier study that mapped the neighbourhood-level burglary rates for four of the five urban areas under study here found that the relative risk of a residential burglary was more acute in the urban centres and tends to decrease the further one moves away from the downtown areas. The spatial pattern of the survey responses, however, moves in exactly opposite direction. In other words, neighbourhood residents collectively are rather poor at evaluating victimization risk: the risk of a burglary is underrated in the city centre and overrated in the peripheries (Kreis, 2012).

The general spatial trend of fear of crime to increase from centre to periphery also applies to the response pattern for the third indicator, actual behavioural changes to avoid crime. The percentage of respondents who resort to behavioural changes to avoid crime in their neighbourhood steadily increases from the centrally located neighbourhood clusters towards the urban peripheries. The only outlier in this general spatial trend is again neighbourhood cluster type 5, which are the suburban areas of Geneva and Lausanne (Kreis, 2012).

The first two panels of the bottom row of Figure 5 show the bar-plots for the survey item measuring physical and social disorder. For the disorder items, the spatial pattern is no longer a more or less linear trend that increases from the centre to the periphery as with fear of crime. The percentage of respondents who spotted signs of physical or social disorder, or both, is higher for the more central neighbourhood clusters type 1 and 2 as well as the peripheral areas type 5 and 6, but slightly lower for the in-between areas of type 3 and 4 (Kreis, 2012).

The sixth panel on the survey item measuring popular satisfaction with the police reveals a spatial pattern that is more in line with actual victimization risk. Asked whether local police was doing a satisfactory job in crime control, the percentage of respondents who rated the police as doing a "very good" or "quite good" job increases from centre to periphery, or from neighbourhood clusters type 1 to 6. This is in line with the spatial pattern detected in an earlier analysis of the neighbourhood-level burglary victimization rates in Swiss urban areas (Kreis, 2012).

The second criterion of the clustering algorithm to develop the neighbourhood typology – that it accounts for a significant share of the between cluster variance in the survey outcome measures – has thus been met. However, before the current neighbourhood typology may be used as intelligence basis to select matched treatment and control areas across urban areas to study the impact of different community policing strategies, a final check is still in order. This test must assess whether the current typology does indeed account for most of the variance in the outcome variables between residents of different neighbourhood types or if there is a significant amount of variance left at the higher aggregate city or regional levels.

In order to test this proposition, a second series of $\chi 2$-independence tests is run, this time to evaluate whether

the response patterns of the survey respondents of a given type of neighbourhood cluster vary significantly between individual cities. Table 1 displays all *p*-values of these tests for all six survey outcome indicators for each of the six neighbourhood types. The results are encouraging: the *p*-value of a majority of these $\chi 2$-independence test is not significant, suggesting that the response patterns of survey respondents residing in the same type of neighbourhood do not differ significantly between cities and that the null hypothesis of basic independence cannot be rejected (which is what the author was hoping for). However, several $\chi 2$-independence tests have a *p*-value that is significant. As a matter of fact, for all five neighbourhood clusters, which are present in more than one urban area, is the p-value below the conventional 0.05 significance level on at least one occasion. This implies that for those neighbourhood types and those survey items, city-level factors still impinge on survey response patterns (Kreis, 2012).

V. DISCUSSION

The current research employed both unsupervised and supervised data mining algorithms to develop a typology of neighbourhoods in order to group neighbourhood districts across the major Swiss urban areas into clusters of similar type. Since the objective of the clustering algorithm was to match suitable *treatment* and *control* areas across the five urban areas in order to enhance the internal validity of an observational study of community policing, the twin optimization criteria for the clustering algorithm were clear: on the one hand, the neighbourhood typology should minimize the between-cluster similarity in the contextual variables, which are potentially correlated with the outcome measures, and thus reduce the risk that these neighbourhood covariates confound any inferences about the impact of the treatment. On the other hand, the neighbourhood typology should account for a significant share of the between-cluster variance in the outcome measures used to evaluate community policing. This was meant to ensure that residents of the same neighbourhood type in different urban areas collectively expressed similar views prior to the onset of treatment and later observed differences in opinion are not due to pre-existing conditions at the onset of community policing implementation (Kreis, 2012).

As the diagnostic plots of the previous section revealed, these twin optimization criteria have largely been met: not only did the neighbourhood typology reduce between-cluster variance in the 24 key variables describing the neighbourhood ecological context. It also accounted for a significant share of the between-cluster variance in the survey response patterns that served as outcome indicators in the evaluation of community policing.

The neighbourhood classification system developed for the community policing evaluation thus goes a long way to reconcile these double optimization criteria notwithstanding the fact that the algorithm did not result in a single optimum number of clusters. Indeed, one conclusion from the clustering procedure was that no such single optimum number of neighbourhood clusters exists. The number of neighbourhood categories that has led to the most clear-cut separation of the clusters on the ecological data is smaller than the number of neighbourhood clusters that accounts for the biggest amount of variance in the outcome survey measures. This implies that in the present case the individual clusters of neighbourhoods of similar type are less distinct and tend to overlap in the original input space of the ecological data, so that individual neighbourhoods could be classified either way (Kreis, 2012). The apparent lack of a single optimum number of neighbourhood clusters compounds the variability of the results inherent in the SOM clustering algorithm. The shape to which the malleable SOM lattice converges during training depends to some extent on the randomly chosen initial values of the weights of the prototype vectors. The standard remedy to handle this aspect of the SOM algorithm is to replicate the clustering procedure multiple times and to compare the results of the individual runs before reaching any conclusions. For the current study, the complete clustering algorithm was replicated 50 times and the solution retained that best met the double optimization criteria set at the outset of the clustering procedure.

By contrast, the MC simulated $\chi 2$-independence tests as well as the random forests algorithm that were both based on the outcome of the SOM algorithm produced very stable results. Random forests were employed during the supervised learning phase in order to identify the key ecological variables among the indicators used

	Fear of Crime	Risk of Victimization	Behavioural Response	Physical Disorder	Social Disorder	Police Effectiveness
1	0.007	0.824	0.006	0.019	0.001	0.857
2	0.000	0.000	0.494	0.395	0.195	0.164
3						
4	0.001	0.130	0.137	0.000	0.007	0.034
5	0.817	0.107	0.130	0.037	0.262	0.367
6	0.301	0.000	0.590	0.734	0.060	0.052

Table 1. Survey response patterns by neighbourhood cluster. p-values of the $\chi 2$-independence tests of the indicators of community policing impact by city, computed separately for each of the six neighbourhood clusters (a value of $p < 0.05$ indicates that response patterns within a given neighbourhood cluster still vary significantly by city).

to describe the neighbourhood context. By selecting only the most important indicators it was possible to build a parsimonious final model with just 24 explanatory variables (out of the original 89) without unduly increasing the classification error rate of the model.

A second limitation of the neighbourhood typology is that it could not account for all or most of the variance in the survey measures used as outcome indicators at the higher aggregate levels of analysis. If survey respondents are grouped by individual neighbourhood clusters, there remains at times significant within-cluster variance in the response patterns at the city-level. This implies that city-level factors still influence survey response patterns, which risks undermining comparisons between neighbourhood residents across urban areas even within a given neighbourhood type. In an observational study design that compares the impact of the program between neighbourhoods of a similar type across urban areas, the current neighbourhood typology thus manages to reduce the threats to internal validity of selection and regression to the mean but cannot rule them out completely (Kreis, 2012).

VI. CONCLUSION

The current research employed geospatial data mining algorithms to classify Swiss urban neighbourhoods into clusters of similar type in order to find matching treatment and control areas for the evaluation of area-based crime prevention programs such as community policing. The clustering procedure made it possible to take high-dimensional data on the demographic and socioeconomic composition as well as the built environment of urban neighbourhoods into account. Not only were these data shown to impact survey response patterns, they may exert an influence on how neighbourhood residents perceive different community policing strategies as well. This approach thus attempted to blunt some of the criticism levelled against recent criminological research on crime prevention of being overly concerned with the question of *what works?* while neglecting the influence of contextual and environmental factors on the success of such initiatives (Williamson et al., 2006, 199f.).

Despite some shortcomings previously discussed, the resulting neighbourhood typology succeeded in achieving a considerable degree of within-cluster homogeneity regarding the ecological data while capturing a significant share of the between cluster variance in the survey items. The individual neighbourhoods within each cluster thus not only share similar ecological characteristics, which may confound inferences about the impact of treatment, but also display similar levels on the outcome measures prior to program implementation, which community policing is trying to influence. The neighbourhood typology thus helps to diminish some of the threats to internal validity of an observational research design and first make possible valid comparative analysis of the impact of area-based crime prevention programs across urban areas.

Besides diminishing the threats to internal validity of an observational study design, the neighbourhood typology unearthed some peculiar facts that are also interesting from a policy perspective. First of all, the typology suggests that neighbourhood residents collectively are not very perceptive when it comes to assessing victimization risk. Whereas actual (burglary) risk is highest in the city centres and tends to diminish towards the boundaries of the urban areas, the spatial pattern for perceived risk is exactly the reverse. The study thus produced evidence that also in Swiss urban areas there are neighbourhoods that show signs of the phenomenon that has become known as the *reassurance gap* (Tuffin et al., 2006), meaning a popular perception of an increase in crime when actual rates are low or falling. This is not to say that neighbourhood residents' perception are completely off the mark, however: when asked about whether the local police are doing a good job in controlling crime in the area, popular approval rates do co-vary spatially with actual levels of victimization (Kreis, 2012).

The second extra bit of information the neighbourhood classification system generated was to highlight the key characteristics of each neighbourhood type from the high-dimensional training data set of ecological indicators. The random forests algorithm proved highly informative in this regard, identifying the key characteristics of each of the six neighbourhood clusters. Needless to say that such information can be used to tailor policy interventions to the specific needs of these different areas.

ACKNOWLEDGEMENTS

This research was supported by funding from the Swiss National Science Foundation (Award No. PBLAP1-142787). The opinions, findings and conclusions or recommendations expressed herein are those of the author and do not necessarily reflect those of the aforementioned agency.

REFERENCES

Ashby, D.I. and Longley, P.A. (2005). Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS*, *9* (1), 53-72.

Bivand, R.S., Pebesma, E.J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. New York: Springer.

Breiman, Leo. (2001). "Random forests." *Machine Learning*, Vol. 45, No. 1, pp. 5-32.

Cook, T.D. and Campbell, D.T. (1979). *Quasi-experi-*

mentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally College Publishing Company.

Davies, D.L. and Bouldin, D.W. (1979)."A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1* (2), 224-227.

Farrington, D.P. (2003). Methodological quality standards for evaluation research. *The ANNALS of the American Academy of Political and Social Science*, 587 (1), 49-68.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31* (14), 2225-2236.

Harper, G., Williamson I., Clarke, G. and See, L. (2002). *Family Origins: Developing Groups of Crime and Disorder Reduction Partnerships and Police Basic Command Units for Comparative Purposes*. London: Home Office.

Kanevski, M., Pozdnoukhov, A. and Vadim T. (2009). *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*. Lausanne: EPFL Press.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78 (9), 1464-1480.

Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). Berlin: Springer.

Kreis, C. (2009). *The Spatio-Temporal Patterns of Fear of Crime in Major Swiss Cities: A quantitative analysis of the spatial distribution of fear of crime, disorder, and public satisfaction with the police among neighborhood residents in Switzerland's five metropolitan areas between 1987 and 2005* (Master thesis in études avancées en urbanisme durable, University of Lausanne, Lausanne).

Kreis, C. (2012). *Community Policing in Switzerland's Major Urban Areas: An Observational Study of the Implementation and Impact Using Geospatial Data Mining* (Ph.D. thesis, University of Lausanne, Lausanne).

Lee, John A. and Michel Verleysen. (2007). *Nonlinear Dimensionality Reduction*. New York: Springer.

Li, Y. and Shanmuganathan, S. (2007). *Social Area Analysis Using SOM and GIS: A Preliminary Research* (RCAPS Working Paper, Vol. 07-3). Beppu City, Japan: Ritsumeikan Asia Pacific University, Ritsumeikan Center for Asia Pacific Studies.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2* (3), 18-22.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Hoboken: Wiley.

R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum, P.R. (2010). *Design of Observational Studies*. New York: Springer.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Shadish, William R., Cook, Thomas D., & Campbell, Donald T. 2002. *Experimental and Quasiexperimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sheldon, G., Hall, R., Brunsdon, C., Charlton, Alvanides, M.S. and Mostratos, N. (2002). *Maintaining basic command unit and crime and disorder partnership families for comparative purposes* (RDS Online Report).

Sherman, L.W., Farrington, D.P., Welsh, B.C. and MacKenzie, D.L. (Eds.). (2002). *Evidence-Based Crime Prevention*. London: Routledge.

Skupin, A. and Agarwal, P. (2008). Introduction: What is a self-organizing map? In Agarwal, P. and Skupin, A. (Eds.), *Self-Organising Maps: Applications in Geographic Information Science* (p.1-20). Chichester: Wiley.

Spielman, Seth E. and Jean-Claude Thill. (2008). "Social area analysis, data mining, and GIS." *Computers, Environment and Urban Systems*, Vol. 32, No. 2, pp. 110-122.

Tuffin, R., Morris, J. and Poole, A. (2006). *An Evaluation of the Impact of the National Reassurance Policing Programme* (Home Office Research Study, Vol. 296). London: Home Office.

Vesanto, J. and Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11* (3), 586 - 600.

Walesiak, M., and Dudek, A. (2012). *clusterSim: Searching for optimal clustering procedure for a data set*. R package version 0.41-8.

Weisburd, D. and Eck, J.E. (2004). What can police do to reduce crime, disorder, and fear? *The ANNALS of the American Academy of Political and Social Science*, *593* (1), 42-65.

Wehrens, R., and Buydens, L. (2007). Self- and super-organizing maps in R: the kohonen package. *Journal of Statistical Software*, 21(5), 1–19.

Welsh, B.C. and Hoshi, A. (2002). Communities and crime prevention. In Sherman, L.W., Farrington, D.P., Welsh, B.C.and MacKenzie, D.L. (Eds.), *Evidence-Based Crime Prevention* (p 165-197). London: Routledge.

Williamson, T., Ashby, A.I. and Webber, R. (2006). Classifying neighbourhoods for reassurance policing. *Policing and Society*, *16* (2), 189-218.

Coordonnées de l'auteur :

Christian Kreis
Nederlands Studiecentrum Criminaliteit
en Rechtshandhaving
De Boelelaan 1077a
1081 HV Amsterdam
ckreis@nscr.nl