Spatial and Temporal Geovisualisation and Data Mining of Road Traffic Accidents in Christchurch, New Zealand

Clive E. SABEL and Phil BARTIE

Abstract

This paper outlines the development of a method for using Kernel Estimation cluster analysis techniques to automatically identify road traffic accident 'black spots' and 'black areas'. A Novel data-mining approach has been developed – adding to the generic exploratory spatial analysis toolkit. Christchurch, New Zealand, was selected as the study area and data from the LTNZ crash database was used to trial the technique. A GIS and Python scripting was used to implement the solution, combining spatial data for average traffic flows with the recorded accident locations. Kernel Estimation was able to identify the accident clusters, and when used in conjunction with Monte Carlo simulation techniques, was able to identify statistically significant clusters.

Keywords and phrases: Kernel Estimation, spatial and temporal patterns, road traffic accidents, Monte Carlo simulation, spatial data mining.

1 Introduction

Improvements in car safety, road engineering, driver education and wider road safety policy have brought about a decrease in the rate of car accidents globally. Nevertheless, accidents still occur, appearing disproportionately within, and to, deprived communities.

Incident reporting systems are increasingly being recognised as offering analysts the capability to data mine for trends which may be subtly related, in the hope that reoccurrence of incidents can be reduced (CASSIDY et al. 2003). Current practices in the spatial analysis of road traffic accidents mostly rely on a visual examination, and are therefore highly subjective depending upon the observer (CRESSIE 1991).

It is not desirable to reduce the dimensionality of data through aggregation as statistical trends can be obscured, or even reversed, a problem formally known as Simpson's paradox (TUNARU 2001).). It is also possible that attempting to join too many disparate datasets may introduce errors (PELED 1993).

2 Research Methodology

Vehicle crash records hold both temporal and spatial trends. For this research we have focussed at the local level to identify geographical variations, but also have utilised the strength in our dataset by exploring temporal trends over 25 years.

Jekel, T., Car, A., Strobl, J. & Griesebner, G. (Eds.) (2012): GI_Forum 2012: Geovizualisation, Society and Learning. © Herbert Wichmann Verlag, VDE VERLAG GMBH, Berlin/Offenbach. ISBN 978-3-87907-521-8.

2.1 Study Area

Christchurch, New Zealand, was chosen as the study area, where both crash data and average traffic flow data were available. The data recorded in the Land Transport New Zealand (LTNZ) accident system included all 28,645 reported minor, serious, and fatal accidents in Christchurch from 1980 to 2004.

2.2 Pattern Detection

We adopted an Exploratory Spatial Data Analysis (ESDA) approach. ESDA can be broadly defined as the collection of techniques to describe and visualise spatial distributions, identify atypical locations, or spatial outliers, and discover patterns of spatial association, clusters or hotspots and suggest spatial regimes or other forms of spatial heterogeneity (ANSELIN, 1999) with a view to develop hypotheses.

A number of spatial tools have been developed in recent decades that help in understanding the geography, and changing geographies, of point-patterns. For our purposes the most promising of these is Kernel Estimation (KE), whereby a distribution of discrete point 'events' is transformed into a continuous raster surface (BAILEY & GATRELL 1995). Extensions to this method accommodating the time dimension are also available (SABEL et al. 2000). Other pattern detection methods such as Kulldorff's Spatial Scan statistic or Moran's I are not as useful for our point pattern analysis which attempts to accurately geographical define the spatial patterns at a local neighbourhood level.

2.3 Simulated Risk Input Surface – Model Input

Point analysis using KE is integrated functionality in modern GIS packages, such as ESRI's ArcGIS, however determining the statistical significance is not. We coded this using Python scripting. Road segment flow and junction data were used to predict the expected accident distribution.

2.4 Monte Carlo Simulation

Monte Carlo simulation was used to establish the statistical significance of clusters found using KE. Conceptually, we modified the methodology of KELSALL & DIGGLE (1995) to generate new synthetic data in a random allocation manner. If this process is then repeated m times, in a form of Monte-Carlo simulation, upper and lower simulation envelopes can be established and thereby an estimate of how unusual the observed pattern is obtained. If the observed pattern lies outside the simulation envelope, one can begin to speak of areas of significantly elevated, or reduced, risk.

The results of the simulations can be graphically displayed by constructing a 'p-value surface' which, for each x (location), gives the proportion of values $\hat{s}_I(x), I = 1,\ldots,m$ which are less than the original estimate $\hat{s}_0(x)$. That is, it gives the proportion of simulated cells which are less than each observed cell, for each grid cell in the matrix. The 2.5% and 97.5% contours of this surface are effectively tolerance and they can then be draped over the original image of estimated risk, to highlight regions which correspond to significantly high or low risk.

3 Results

Our results demonstrated in Figure 1 predictably identified raised incidence in the CBD (in the centre of the image), but also some major intersections outside the centre. Deprived neighbourhoods to the immediate east of the city centre are particularly impacted.

The risk of road travel changes at different temporal scales (time of day, season, school term time and annually) and the modelling carried out in this research assumes the dataset has remained constant throughout the entire study period. Despite this, the initial results are promising in terms of identifying areas which can become focal points for more detailed study. These identified areas are already attracting interest from public authorities. In Figure 2, we present density scatterplots, constructed using KE methods, comparing temporal changes over decade scales. Comparing differences between vehicular and cyclist accidents is revealing. Vehicle accidents have declined over the 25 year period of analysis, whereas accidents involving cyclists have not. Clear policy initiatives such as the introduction of seat belts and stricter drive-driving legislation clearly appear c. 1985 and towards the end of the 1990s.

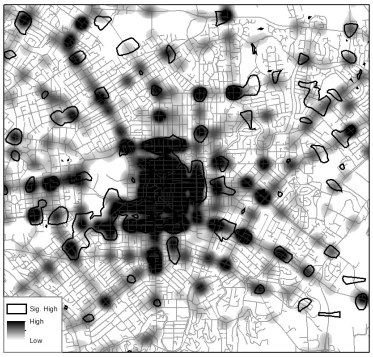

Fig. 1:
Significant Accident Clusters (top 2.5%) where Observed is higher than Expected – Christchurch, NZ

4 Discussion

Road traffic analysis is a very complex topic; the Police have almost 600 different cause codes which can be assigned to explain the origins of the accident in their report. It would be unrealistic to believe that our two input variables (flow and distance to junction) for a Simulated Input Risk Surface would form a complete answer to where expected accidents should be placed in the Monte Carlo simulation. Our results tend to focus on the statistically significant accident points and areas around junctions, which may be a characteristic of KE, being more suited to locating area patterns, than locating linear clusters (‘black’ routes). We propose that adjustment to the KE bandwidth allows the technique to be adapted for use from locating accident ‘black’ spots to ‘black’ zones. We acknowledge that the choice of KE as a technique which models a probability surface across all space might be problematic when considering road accidents, which by their nature occur only on the road network, but we draw readers attention to the broader zones or neighbourhoods of raised accidents rates that our analysis has highlighted that a linear based analysis would struggle to reveal.

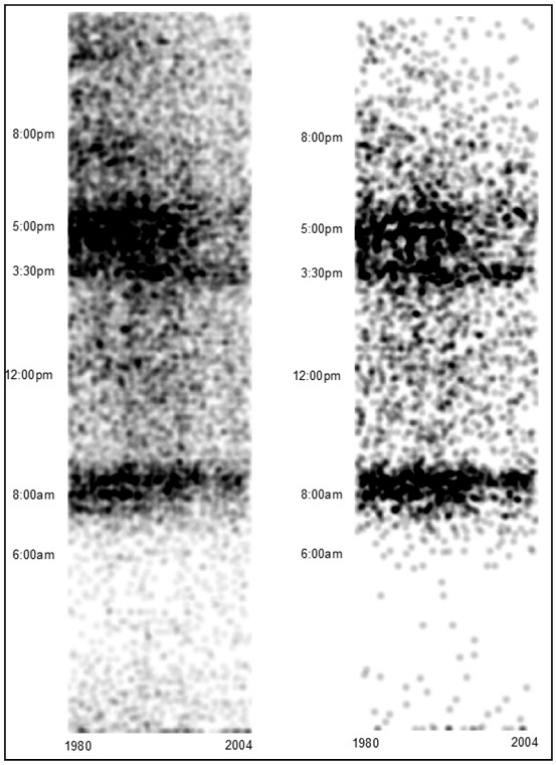

Fig. 2:
Temporal analysis of accidents involving all vehicles (left) and cyclists (right). Black density represents high incidence.

References

ANSELIN, L. (1999), Interactive techniques and exploratory spatial data analysis. In: Geographical Information Systems: principles, techniques, management and applications. Ed. by LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J. & RHIND, D. W. John Wiley, New York, 253-266.

BAILEY, T. C. & GATRELL, A. C. (1995), Interactive Spatial Data Analysis. Longman. Harlow.

CASSIDY, D., CARTHY, J., DRUMMOND, A., DUNNIOU, J. & SHEPPARD, J. (2003), The Use Of Data Mining In The Design And Implementation Of An Incident Report Retrieval System. In: Proc of the 2003 IEEE Systems and Information Engineering Design Symposium, 24-25 April 2003 Univ. of Virginia, USA.

CRESSIE, N. A. C. (1991), Statistics for Spatial Data. Wiley, New York.

KELSALL, J. E. & DIGGLE, P. J. (1995), Kernel Estim. of Relative Risk, Bernoulli, 1, 3-16.

PELED, A. & HAKKERT, A. S. (1993), A PC-oriented GIS application for road safety analysis and management. Traffic Engineering and Control, 34 (7/8), 355-361.

SABEL, C. E., GATRELL, A. C. et al. (2000), Modelling exposure opportunities: estimating relative risk for Motor Neurone Disease in Finland. Social Science & Medicine, 50 (7+8), 1121-1137.

TUNARU, R. (2001), Models of Association versus causal models for contingency tables. The Statistician, 50 (3), 257-269.