Cluster Detection in Point Event Data having Tendency Towards Spatially Repetitive Events

Allan J Brimicombe

Centre for Geo-Information Studies, University of East London,
University Way, London, E16 2RD, UK
Tel: +20 8223 2352
FAX: +20 8223 2918
Email: a.j.brimicombe@uel.ac.uk

Abstract

The analysis of point event patterns in geography, ecology and epidemiology have a long tradition. Of particular interest are patterns of clustering or 'hot spots' and such cluster detection lies at the heart of spatial data mining. Certain classes of point event patterns exhibit a tendency towards spatial repetitiveness (within the resolution of geo-positioning) although with a temporal separation. Examples are crime and traffic accidents. Spatial superimposition of point events challenges many existing approaches to cluster detection. In this paper a variable resolution approach, Geo-ProZones, is applied to residential burglary data exhibiting a high level of repeat victimisation. This is coupled with robust normalisation as a means of consistently defining and visualising the 'hot spots'.

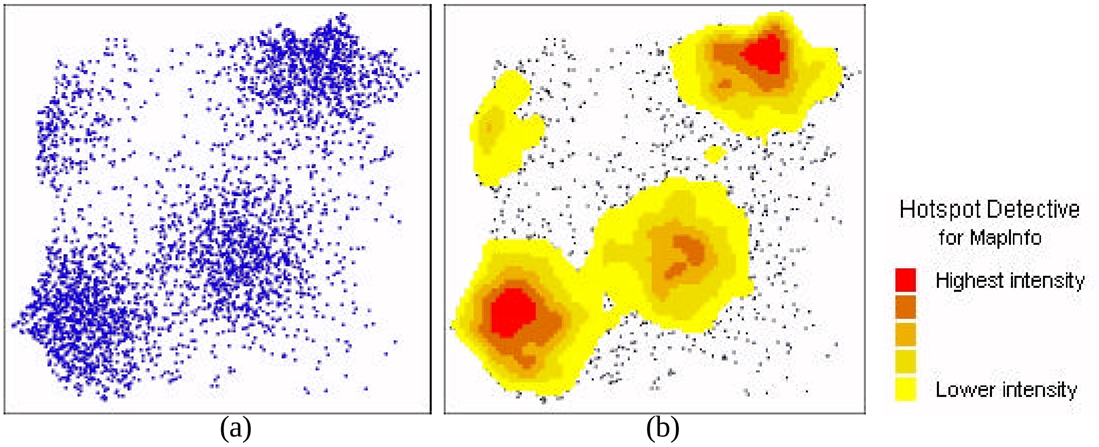
1. Introduction

The analysis of point event patterns in geography, ecology and epidemiology have a long tradition (e.g. Snow, 1855; Clark & Evans, 1954; Harvey, 1966; Mantel, 1967; Cliff & Ord, 1981). The patterns detected are usually broadly classified as random, uniform or clustered. Although a pattern of spatial randomness in data has traditionally been assumed to have no underlying process of interest, Phillips (1999) has nevertheless pointed out that such apparent randomness may be attributable to chaotic deterministic patterns and should therefore not be ignored out of hand. Where a point pattern exhibits spatial uniformity, a space-filling mutual exclusion process can be hypothesised. Clustered patterns, however, have generally raised the strongest interest and hypotheses for underlying processes. Thus cluster detection lies at the heart of spatial data mining (Murray & Estivill-Castro, 1998; Openshaw, 1998; Murray, 2000; Miller & Han, 2001).

Clustered point patterns can be visualised spatially as local concentrations of events in close proximity to one another with each cluster separated by intervening spaces characterised by empty, less dense or apparently random patterns of point events. However, certain classes of point event patterns have a significant proportion of their data having a tendency towards exact spatial repetitiveness (within the resolution of geo-positioning) although with a temporal separation between events. Typical examples would include: crimes recorded against a property address (e.g. residential burglary, shoplifting, intimate partner violence), traffic accidents recorded against a section of road or intersection, utility failures recorded against a node or discrete section of network and so on. The focus of analysis of such data sets is in defining 'hot spots' (e.g. for crime) or 'black spots' (e.g. for traffic accidents) where spatial clustering exists, but the occurrence of this spatial superimposition of point events challenges many existing approaches to cluster detection. In this paper a variable resolution approach, Geo-ProZone analysis, is applied to residential burglary data exhibiting a high level of repeat victimisation. This is coupled with robust normalisation as a means of consistently defining and visualising the highest densities or 'hot spots'.

2. Cluster detection of 'hot spots'

The literature on clustering of point event data can be broadly classified into two approaches. One set of approaches is allied to mainstream statistics emanating from the work of Sokal & Sneath (1963). Thus clustering is a means of classification or grouping where clusters can be defined as "groups of highly similar entities" (Aldenderfer & Blashfield, 1984, p7). Spatially, this approach to cluster analysis will seek to form a segmentation into regions which minimises within-cluster variation but maximises between-cluster variation. There is a general expectation that the clustering will be mutually exclusive in including all points and is therefore space-filling within the geographical extent of the data (see for example Murray & Estivill-Castro, 1998; Murray, 2000). Halls *et al*. (2001) and Estivill-Castro & Lee (2002) provide examples of the use of Dirichlet and Delaunay diagrams, respectively, to define spatial clusters. These algorithms, however, will fail where points occupy the exact same location. To delete duplicate points to overcome this problem is likely to lead to important data loss, whilst to shift points slightly into non-duplicate positions will introduce significant bias away from being able to detect such repeat events. The second broad set of approaches uses spatially exhaustive search to identify localised excesses of event occurrences. Typical of this approach is the Geographical Analysis Machine (GAM) and its descendants (Openshaw *et al*., 1987; Openshaw, 1998). Similar approaches are based around kernel density estimation (Silverman, 1986; Atkinson & Unwin, 1998; Brunsdon, 1995) in which the highest densities form 'hot spots' (e.g. Gatrell et al., 1996). This approach is particularly popular in crime analysis (Harries, 1999; Ratcliffe & McCullagh, 1999; McLafferty *et al*., 2000) with GIS functionality available, for example, in the Spatial Analyst extension to ArcView® and in Hotspot Detective for MapInfo® (Figure 1).



(a) (b)

Figure 1. Kernel density estimation for 'hot spot' detection: (a) burglary point event data set; (b) kernel density estimation using default parameters (superimposed on point pattern) - 'hot spots' are usually taken to be the highest intensity locations.

The popularity of the kernel density estimation (KDE) approach is clear from its ease of use and the striking visualisations it can produce. It is nevertheless an interpolation that transforms the point events into a more-or-less smoothed continuous surface and, with any such technique, parameters need to be set that are critical to the outcome. For KDE these are the underlying grid size and the kernel bandwidth. Reasonable values for parameters can be difficult to estimate and are often done so subjectively (Sabel *et al*., 2000). Fotheringham *et al.* (2000) suggest an optimum bandwidth calculated from the standard distance. For situations where there are contrasting densities across a study area (e.g. urban to rural), an adaptive bandwidth can be employed (Brunsdon, 1995). Best practice would suggest a form of sensitivity analysis to identify optimum parameter values (Brimicombe, 2003). Figure 3 shows such an approach for a fixed grid size (one hectare) and varying bandwidth. The maximum nearest neighbour distance (NND) between point events in Figure 1(a) is 574m or approximately 12 times the median NND of 47.5m; so as a simple sensitivity test the bandwidth has been bracketed at three, six and nine times the median NND. The effect is to produce increased size and severity of 'hot spots'. The software default settings producing Figure 1 (b) produces the greatest visual impact. But what then is an acceptable 'truth'? Although from a research perspective the sensitivity to grid size should also be tested, the pragmatics of the workplace usually means that analysts accept the default values for parameters suggested by the software as a matter of convenience.

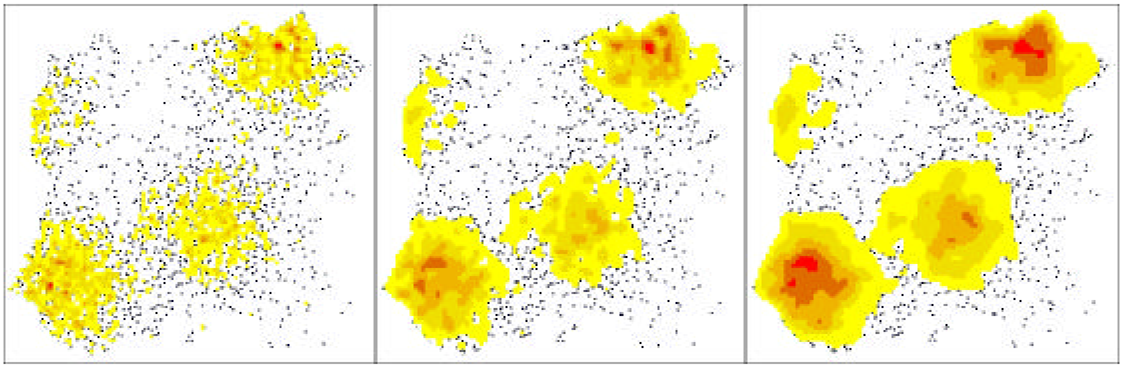


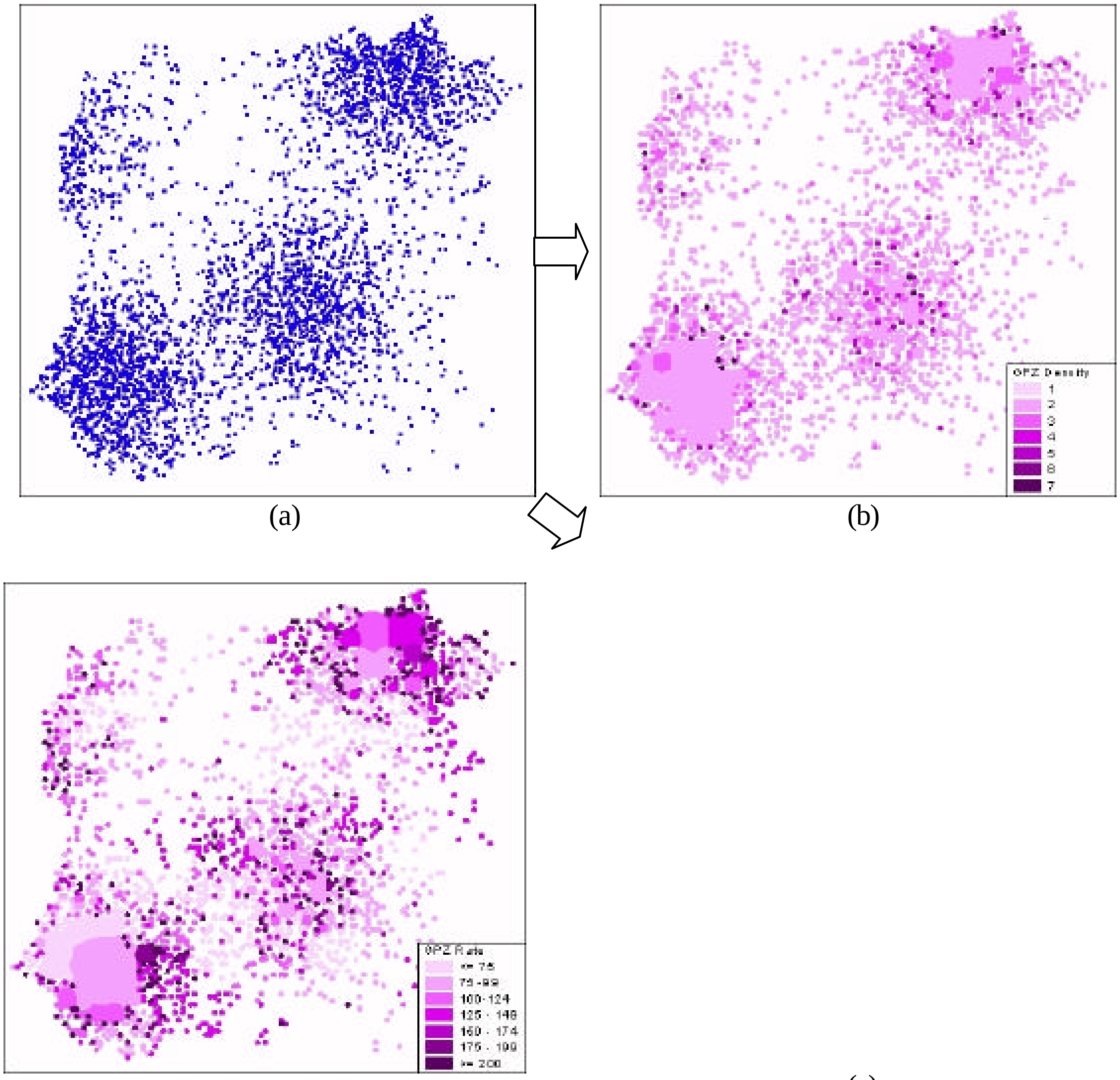
Figure 2. Searching for optimum bandwidth: (a) 3 times median NND; (b) 6 times median NND; (c) 9 times median NND;{for legend see Figure1}.

The burglary data set presented here has a large number of repeat victimisations giving superimposed point events. Theory would suggest that 'hot spots' would be quite localised. High crime areas are primarily so because they are areas of high repeat offending and high repeat victimisation (Trickett *et al*., 1992; Townsley *et al*., 2003). KDE, as used by many police analysts, smoothes over the very localised repeat victimisations in favour of the regional pattern with choice of end result driven more by the aesthetic qualities of the visualisation. Boundary effects around the edge of data sets are also a problem for density estimation and perhaps not surprisingly police analysts tend not to find 'hot spots' at the edge of their jurisdictions. KDE software in the public domain by Atkinson & Unwin (2002) for MapInfo® does offer a guarded buffer to avoid spurious values at boundaries but does not entirely overcome the problem of how to identify real 'hot spots' that exist at boundaries. Figures 1 and 2 focus attention on crime counts, that is, an elevated share of crime in a localised area. 'Hot spots' based on counts inform the deployment of resources in response to events. Less common in crime analysis (but more common, for example, in epidemiology) are 'hot spots' based on elevated rates. Such 'hot spots' reflect the level of risk and thus inform deployment of resources for mitigation. For the same distribution of point events, 'hot spots' based on counts are often different to those based on rates as the latter are not just a function of the distribution of point events but also of the underlying population at risk. Ideally both should be used. Sabel *et al*. (2000) report the use of KDE in association with an underlying population at risk to map relative risk of disease occurrence. Whilst readily implemented, it suffers from the added difficulty of parameter estimation for two KDE surfaces (disease occurrence and population at risk) that are then combined to produce a ratio surface.

3. The Geo-ProZone algorithm

The theory of adaptive recursive tessellations is given in Tsui & Brimicombe (1997a) with applications of their use for spatial analysis in Tsui & Brimicombe (1997b). Specific application to point pattern analysis can be found in Brimicombe & Tsui (2000) and Brimicombe (2003). At the heart of adaptive recursive tessellations is a variable resolution approach to space. No longer are scale and resolution treated as being uniform across an area but are allowed to vary locally in response to the point pattern. This is achieved through a recursive decomposition of space, similar to quadtrees, but allowing variable decomposition ratios (quadtrees only have 1:4 ratio) and rectangular cells (quadtrees are usually restricted to square cells). The algorithm makes no prior assumptions about the statistical or spatial distribution of points. Each point is treated as a binary occurrence of some phenomenon without further descriptive attributes. The starting point is a convex hull of all the point events. Maximum and minimum x and y values of the data set form the maximal cell. The decision to further decompose any one cell larger than the atomic cell size is based on the variance at the next level of decomposition and a heuristic on the number of empty cells that result. The atomic cell size (or smallest possible cell size) is mediated between the median nearest neighbour distance and average cell size per point, whichever is smallest. Any cells formed through decomposition that fall outside the convex hull are automatically deleted. Tests have shown the algorithm to be consistently effective in comparison with other approaches of point cluster detection (see Brimicombe & Tsui, 2000). The resulting clusters are termed Geo-ProZones (GPZ) as they represent zones of geographical proximity in the point pattern. As with kernel density estimation, the highest densities can be taken as 'hot spots'. However, GPZ are not an interpolation, but a segmentation into polygons having internal consistency in the distribution and density of the point events within them. Also, it does not suffer from boundary effects. GPZ for the burglary events in Figure 1 are given in Figure 3.

The pattern in Figure 3(b) reflects the pattern in Figure 1(b). The underlying speckle arises because all events are mapped without smoothing. The highest densities, or what would be interpreted as 'hot spots', occur as more localised concentrations of repeat victimisation. Since GPZ results in polygons, they can be readily overlaid on an underlying population at risk (such as from census data) and re-classified as rates. Figure 3(c) shows GPZ as rates per thousand households from the underlying census data. The pattern of 'hot spots' is quite different and identifies where citizens are at greatest risk. Some of these areas appear reasonably extensive, others are quite localised where repeat victimisation is occurring.



(a)

(b)

(c)

Figure 3. Geo-ProZone analysis: (a) burglary point events; (b) GPZ for density of burglaries per hectare; (c) GPZ for rate of burglaries per thousand households.

4. Robust normalisation for outlier detection and consistent visualisation

Whilst GPZ offers important methodological improvements in cluster detection where there is a tendency towards localised repetitive events, outstanding issues for this (and any other approach) relates to the well-known limitations of thematic mapping: number of class intervals, the fixing of class boundaries and what colours to use. There is the added issue of what constitutes the cut off for a 'hot spot'. In practice, decisions often lack consistency. One approach is through data normalisation. A new form - robust normalisation - (Formula 1, below) has been introduced (Brimicombe, 1999, 2000) as an alternative to the popular Z transformation where data are skewed and where a Z transformation of such data is likely to be biased.

$$\begin{aligned} RN &= \frac{(x - median)}{(median - lower\_quartile)} && for\ x < median \\ RN &= \frac{(x - median)}{(upper\_quartile - median)} && for\ x > median \\ RN &= 0 && for\ x = median \end{aligned} \qquad (1)$$

The term 'robust' refers to the use of the median and inter-quartile range from robust statistics (Hettmansperger & McKean, 1998). The outcome of robust normalisation (Figure 4) is a distribution of median = 0, lower quartile = -1 and upper quartile = +1. Values of <-3 and >+3 are considered extreme values and the transformation can be used consistently for detecting outliers. It also provides a means of defining consistent class intervals and cartographic representation where the ability to make visual comparisons is important. Robust normalisation is achieved using the algorithm in Formula 1 which is easily coded as a Microsoft® Excel macro. For very 'flat' data sets where the lower quartile equals the median or upper quartile equals the median (or all three equal each other), then robust normalisation is likely to fail (division by zero). The Excel macro therefore contains diagnostics to warn the user of such situations.

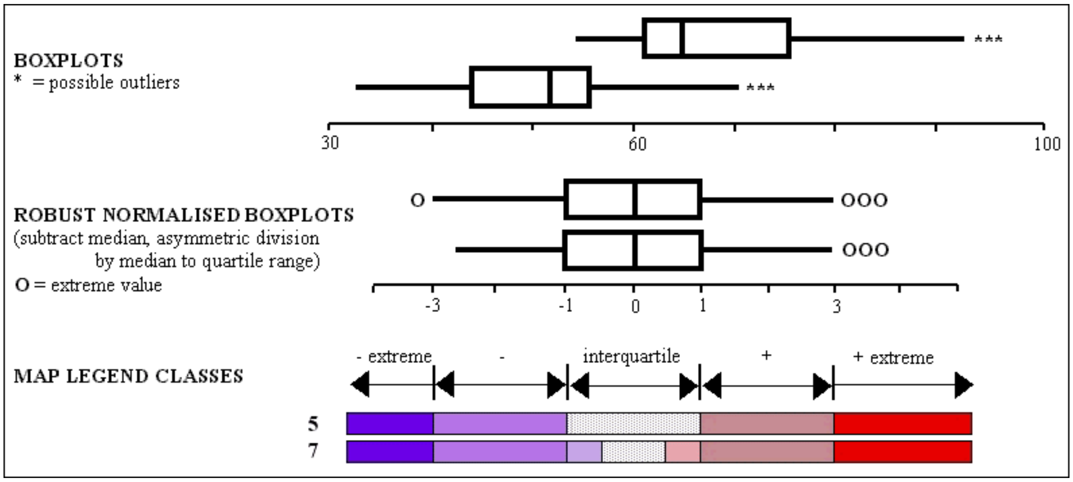
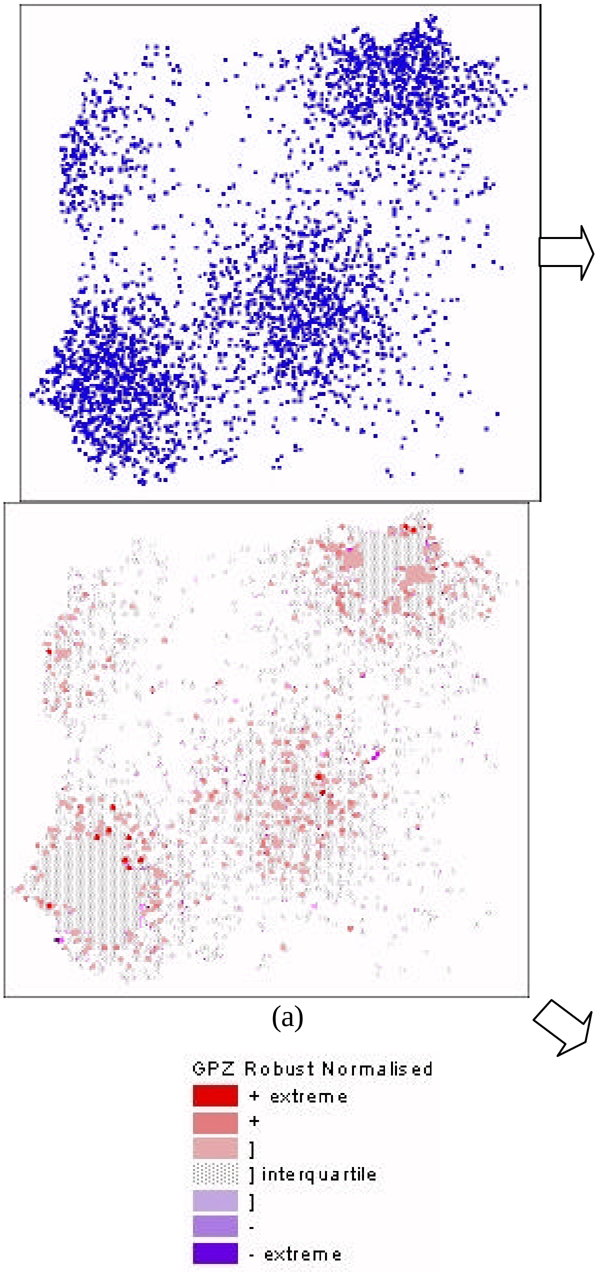


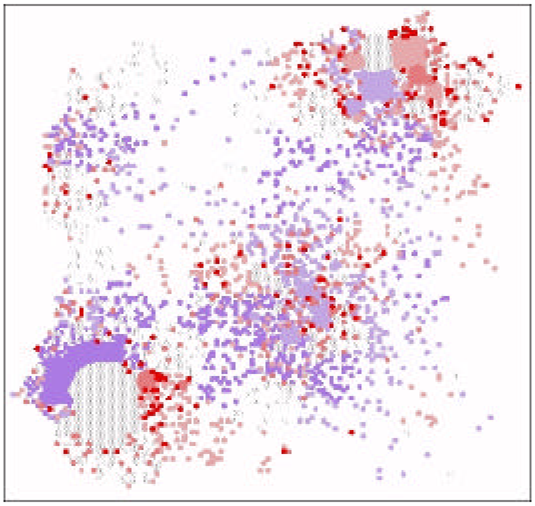
Figure 4. Robust normalisation of two dissimilar distributions to achieve the same interquartile range, identify outliers and achieve consistent map legend classes (5 or 7 intervals) for visualisation.

Robust normalisation can be applied both to area-based data and to density estimate interpolations. For 'hot spot' detection it is the extreme positive values (>+3) that are of most interest. The robust normalised distribution easily lends itself to five or seven class intervals with class boundaries at quartiles (in the seven class interval scheme the values immediately around the median are further separated, as in Figures 4 and 5) and can be used in a standardised way for all visualisations. This allows more objective comparisons between maps (either of different variables or of the same variable over time). Figure 5 shows robust normalisation applied to both GPZ densities and rates from Figure 3 overcoming problems of arbitrary numbers of classes and class intervals. Localised 'hot spots' are where there are extreme positive values. By analogy 'cool spots' would be where there are extreme negative values. Clearly picked out in Figure 5(b) are the localised excesses of counts that represent 'hot spots' of repeat victimisation. More striking though is Figure 5(c) which shows many more localised excesses of rates when counts are rela ted to the underlying population at risk. Importantly, in both cases the 'hot spot' distributions do not necessarily conform to initial subjective impressions of the point patterns (Figure 5(a)) as the 'hot spots' are in fact occurring where point events are superimposed and hence can neither be picked out by eye nor effectively by kernel density estimation.



(a)

(b)



(c)

Figure 5. Applying robust normalisation: (a) burglary point events; (b) GPZ for density of burglaries per hectare; (c) GPZ for rate of burglaries per thousand households; {legend applies to both (b) and (c)}.

5. Conclusions

A consistent approach to cluster detection and visualisation of 'hot spots' through the combined use of Geo-ProZones and robust normalisation has been presented. The Geo-ProZones algorithm overcomes problems raised when data sets exhibit a tendency towards spatially repetitive events and where 'hot spots' will be highly localised. It also overcomes the boundary issues associated with KDE. The algorithm is suited to producing both segmentations of point densities and rates/risk in relation to underlying populations at risk. Problems can arise, however, from the presence of spatial outliers distorting the initial convex hull created around the point events. Improvements to the algorithm are being investigated to reduce sensitivity to any such outliers. Robust normalisation provides consistency in defining class intervals with 'hot spots' as localised extremes. Visual map comparisons for decision making are rendered more objective. Applications of the approach have been successfully used in analyses of crime, health and pipe bursts in water reticulation systems.

6. References

Aldenderfer, M. S., and Blashfield, R. K., 1984, *Cluster Analysis* (California: Sage).

Atkinson, P.J., and Unwin, D.J., 1998, Comparisons and problems of applying density estimation techniques to the distribution of hepatitis. *Geographical Systems*, **5**, 301-312

Atkinson, P.J., and Unwin, D.J., 2002, Density and local attribute estimation of an infectious disease using MapInfo. *Computers and Geosciences*, **28**, 1095-1105.

Brimicombe, A.J., 1999, Small may be beautiful - but is simple sufficient? *Geographical and Environmental Modelling*, **3**, 9-33.

Brimicombe, A.J., 2000, Constructing and evaluating contextual indices using GIS: a case of primary school performance. *Environment & Planning A*, **32**, 1909-1933.

Brimicombe, A.J., 2003, *GIS, Environmental Modelling and Engineering*. (London: Taylor & Francis).

Brimicombe, A.J., 2003, A variable resolution approach to cluster discovery in spatial data mining. In Kumar *et al.* (eds.) *Computational Science and its Applications*. (Berlin, Springer-Verlag), vol. 3, pp. 1-11.

Brimicombe, A. J., and Tsui H. Y., 2000, A variable resolution, geocomputational approach to the analysis of point patterns. *Hydrological Processes*, **14**, 2143-2155.

Brunsdon, C., 1995, Estimating probability surfaces for geographical point data: an adaptive kernel algorithm. *Computers and Geosciences*, **21**, 877-894.

Clark, P. J., and Evans, F. C., 1954, Distance to nearest neighbour as a measure of spatial relations in populations. *Ecology*, **35**, 445-453.

Cliff, A. D., Ord, J. K., 1981, *Spatial Processes: Models and Applications*. (London: Pion).

Estivill-Castro, and V., Lee, I., 2002, Argument free clustering for large spatial point-data sets via boundary extraction from Delaunay Diagram. *Computers, Environment and Urban Systems*, **26**, 315-334.

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

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S., 1996, Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, **NS 21**, 256-274.

Halls, P.J., Bulling, M., White, P. C. L., Garland, L., and Harris S., 2001, Dirichlet neighbours: revisiting Dirichlet tessellation for neighbourhood analysis. *Computers, Environment and Urban Systems,* **25**, 105-117.

Harries, K., 1999, *Mapping Crime: Principle and Practice*. (Washington DC: National Institute of Justice).

Harvey, D. W., 1966, Geographical processes and point patterns: testing models of diffusion by quadrat sampling. *Transactions of the Institute of British Geographers*, **40**, 81-95.

Hettmansperger, T.P., and McKean, J.W., 1998, *Robust Nonparametric Statistical Methods*. (London: Arnold).

MacQueen, J., 1967, Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Maths and Statistics Problems*, vol. 1, 281-297.

Mantel, M., 1967, The detection of disease clustering and a generalised regression approach. *Cancer Research*, **27**, 209-220.

McLafferty, S., Williamson, D., and McGuire, P.G., 2000, Identifying crime hotspots using kernel smoothing. In Goldsmith *et al.* (eds.) *Analyzing Crime Patterns*, (Thousand Oaks CA: Sage), pp. 77-85.

Miller, H. J., and Han, J., 200, *Geographic Data Mining and Knowledge Discovery*. (London: Taylor & Francis).

Murray, A. T., 2000, Spatial characteristics and comparisons of interaction and median clustering models. *Geographical Analysis*, **32**, 1-18.

Murray, A. T., and Estivill-Castro, V., 1998, Cluster discovery techniques for exploratory spatial data analysis. *International Journal of Geographical Information Science*, **12**, 431-443.

Openshaw, S., 1998, Building automated geographical analysis and explanation machines. In Longley *et al.* (eds.) *Geocomputation: A Primer*, (Chichester: Wiley), pp. 95-115.

Openshaw, S., Charlton, M. E., Wymer, C., and Craft, A. W., 1987, A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, **1**, 359-377.

Phillips, J. D., 1999, Spatial analysis in physical geography and challenge of deterministic uncertainty. *Geographical Analysis*, **31**, 359-372.

Ratcliffe, J.H., and McCullagh, M.J., 1999, Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems*, **1**, 385-398.

Sabel, C.E., Gatrell, A.C., Löytönen, M., Maasilta, P., and Jokelainen, M., 2000, Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. *Social Sciences & Medicine*, **50**, 1121-1137.

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

Snow, J., 1855, *On the Mode of Communication of Cholera*. (London: Churchill Livingstone).

Sokal, R., and Sneath, P., 1963, *Principles of Numerical Taxonomy*. (San Francisco, Freeman).

Townsley, M., Homel, R. and Chaseling, J., 2003, Infectious burglaries. A test of the near repeat hypothesis. *British Journal of Criminology*, **43**, 615-633.

Trickett, A., Osborne, D.K., Seymour, J., Pease, K., 1992, What is different about high crime areas? *British Journal of Criminology*, **32**, 81-90.

Tsui, H. Y., and Brimicombe, A. J., 1997a, Adaptive recursive tessellations (ART) for Geographical Information Systems. *International Journal of Geographical Information Science*, **11**, 247-263.

Tsui, H. Y., and Brimicombe, A. J., 1997b, Hierarchical tessellations model and its use in spatial analysis. *Transactions in GIS*, **2**, 267-279.

Statement on trade names, trademarks: Where a name is used by a company to distinguish their product which they may claim as a trade name or trademark, then that name appears here with an initial capital or all in capital letters. Mention of such proprietary or commercial products does not constitute an endorsement, nor indeed a refutation of these products.