Evaluation of Spatial Cluster Detection Algorithms for Crime Locations

Marco Helbich and Michael Leitner

Abstract This comparative analysis examines the suitability of commonly applied local cluster detection algorithms. The spatial distribution of an observed spatial crime pattern for Houston, TX, for August 2005 is examined by three different cluster detection methods, including the Geographical Analysis Machine, the Besag and Newell statistic, and Kulldorff's spatial scan statistic. The results suggest that the size and locations of the detected clusters are sensitive to the chosen parameters of each method. Results also vary among the methods. We thus recommend to apply multiple different cluster detection methods to the same data and to look for commonalities between the results. Most confidence will then be given to those spatial clusters that are common to as many methods as possible.

1 Introduction

Geographic Information Systems (GIS) and spatial analysis have become valuable and indispensable tools used in day-to-day operations of governmental agencies. This is also true of police departments, which increasingly supplement and enhance their traditional criminological modus operandi with geographical information technologies for tactical and strategic decision-making. To improve the ability to gain knowledge from geospatial data and to understand the spatial processes contributing to the presence or absence of criminal offenses, spatial data mining

M. Helbich (✉)
GIScience, Department of Geography, University of Heidelberg, Berliner Strasse 48, 69120 Heidelberg, Germany
e-mail: marco.helbich@geog.uni-heidelberg.de

M. Leitner
Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, USA
e-mail: mleitne@lsu.edu

W. Gaul et al. (eds.), *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-24466-7_20,

tools are essential. The detection of local spatial crime clusters helps to improve the efficiency of strategies for prevention and serves as strategic planning tool for decision-making.

Because it is an important topic in spatial statistics, many different global and local spatial algorithms for different data structures exist to evaluate patterns (e.g. Anselin 1999; Kulldorff 1997; Openshaw et al. 1987). For this reason researchers are called upon to develop easy to understand and easy to use guidelines that practitioners can rely on for choosing appropriate cluster detection methods for different crime distributions. The research presented in this paper focuses on the local – and thus mappable – modeling of spatial clusters and on global trends of a spatial pattern. Following Knox (1989) a spatial cluster is "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance." To detect spatial clusters, methods typically screen a study area for evidence of hot spots of points (e.g., offense or disease events) without preconception about their likely locations (Besag and Newell 1991).

Fotheringham and Zhang's research (Fotheringham and Zhan 1996) is one of the few dedicated to the evaluation of the exploratory performance of the Geographical Analysis Machine (GAM, Openshaw et al. (1987) and the Besag and Newell statistic (BNS, Besag and Newell (1991) in conjunction with GIS. Their data set consists of single-family detached dwellings in the U.S. city of Amherst, NY. They concluded that both are suitable within a GIS framework and that BNS is less prone to producing false positive results. Nevertheless, the results of both cluster detection methods are criticized because of perceptual issues of their exploratory maps. Other algorithm comparisons can be found in Kulldorff et al. (2003) and Song and Kulldorff (2003) which emphasize the efficiency of Kulldorff's spatial scan statistic (SSS).

The primary aim of this paper is the evaluation of different local cluster detection algorithms that are frequently used in practice, such as the GAM (Openshaw et al. 1987), the BNS (Besag and Newell 1991), and the SSS (Kulldorff 1997). The main research questions are: Do crime incidences tend to cluster in geographical space? Do the algorithms detect the same spatial clusters? This paper will also discuss the pros and cons of each statistic and it compares each statistic using a real world example of crime locations.

The paper is structured as follows: Sect. 2 presents the study area, the data set, and the necessary data preprocessing steps. Section 3 briefly introduces the methodology. In its subsections the three selected algorithms of local cluster detection are briefly presented. This is followed by a discussion of the results of the empirical analysis. The paper concludes with a summary and some useful recommendation for crime analysts (Sect. 4).

2 Study Area and Data

The study area comprises of the U.S. metropolitan area of Houston, TX, which is located inside Harris County. Crime data (e.g., burglaries, robberies, burglaries of motor vehicle, auto thefts) for this study area were received for the entire

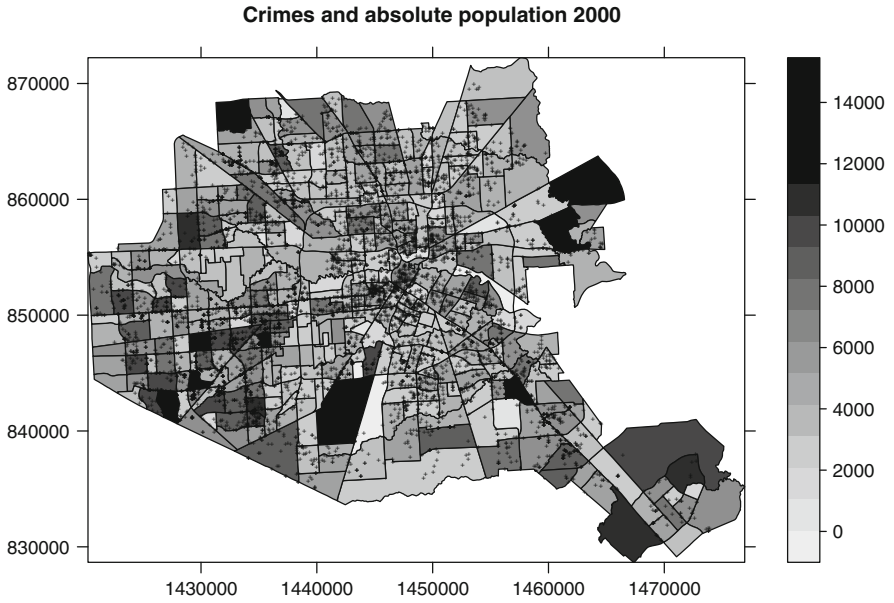

Fig. 1 Distribution of crime locations (back cross signatures) for August 2005 and the at risk population for the census tracts (shaded polygons). Each point signature represents one crime event

month of August 2005 from the Houston Police Department. The data included the offense date and time, offense type, police beat, and the address of the offense at the street block level. This allowed the geo-coding of crime locations using the TIGER (Topologically Integrated Geographic Encoding and Referencing system) street network data freely available from the U.S. Bureau of Census. Of the total number of crimes (8,528) that occurred during the month of August 2005, 8,057 (94.5%) were successfully geo-coded. As expected, the geo-coding rate varies by police district. Figure 1 shows the spatial distribution of the crime locations as well as the spatial distribution of the population in every census tract. In Figs. 1–5 positions are referenced to the Texas State Mapping System (Lambert Conformal Conic, EPSG Code: 3081), hence the x- and y-axes denote the distances (in meters) from the Longitude and Latitude of Origin, respectively.

3 Methodology and Results

This study compares three different and frequently used local cluster detection algorithms. As shown in Fig. 1 both crime and population patterns are distributed heterogeneously in space and generally there are more offenses in densely populated census tracts. Spearman's rank correlation coefficient confirms this hypothesis and shows some significant positive association ($\rho = 0.449$; $p < 0.001$) between the

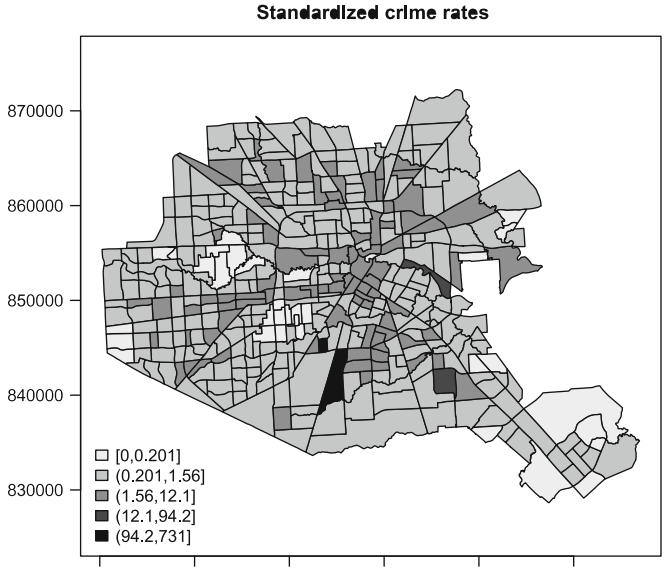

Fig. 2 Standardized crime rates for August 2005

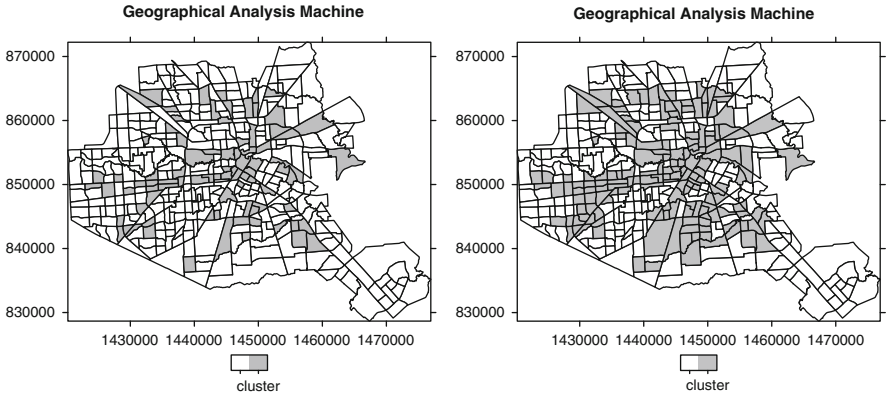

Fig. 3 Significant clusters (shaded in gray) detected by GAM with different parameter settings. The left map uses a cell width of 500 m and $r = 500$ m. The right map uses a cell width of 500 m and $r = 1{,}000$ m

number of crimes per census tract and its corresponding heterogeneously distributed population at risk. The latter is based on the 2000 census and assumed to be equal to the residential population in each census tract. We are aware that the estimate of the population at risk is biased in areas such as the central business district, industrial areas, or around big malls. Similarly, in the case of certain types of

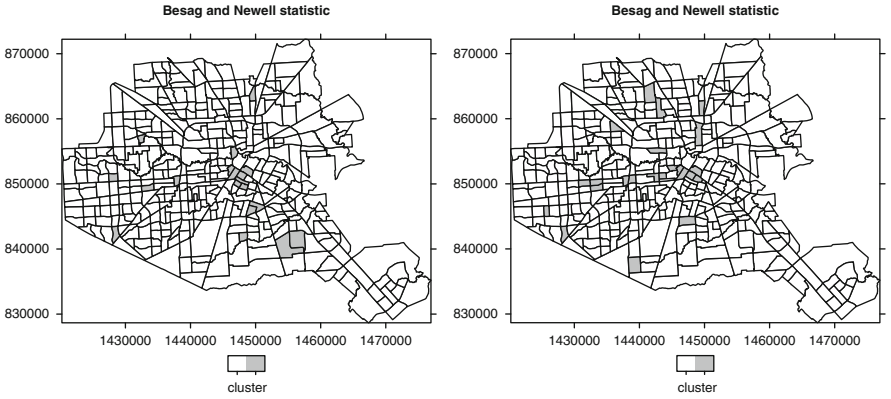

Fig. 4 Significant ($\alpha = 0.05$) clusters detected by BNS (shaded in gray). The left map uses $k = 20$. The right map uses $k = 30$

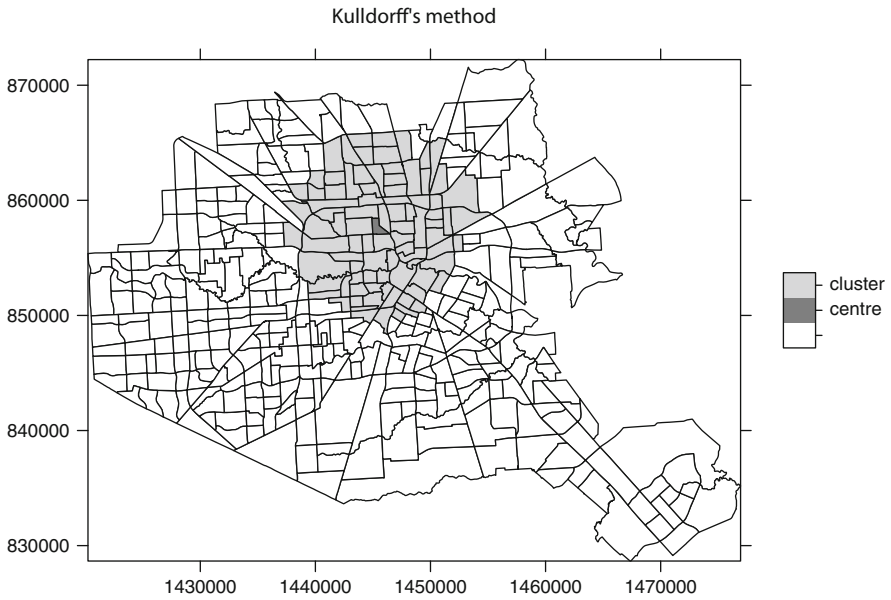

Fig. 5 Clusters (shaded in gray) detected by SSS with a significance value $\alpha = 0.05$ and a maximum fraction of the total population is 20% of the overall population

property crime (e.g., car theft), human population is only a proxy for the actual (car) population at risk (assuming constant number of cars per capita). Because population data were only available on the administrative level of census tracts, the point locations of crimes were aggregated to the same enumeration units. Thus, our reformulated research question is: Are crime locations more clustered relative

to the background population? Nevertheless, this aggregation process is not free of criticism. On the one hand it induces some aggregation bias, because it does not reflect the real-world situation exactly, but on the other hand the consideration of the at risk population leads to more distinguished clusters than purely spatial ones, as stated in Fotheringham and Zhan (1996). However, in the on-going analysis census tracts are represented by their centroids and attributed with total crime incidence and total background population. Because density plots of the number of crime locations and the population size per census tract follow approximately a Poisson distribution, we assumed that this distribution type is appropriate for further analysis. For computation we used the R statistical programming environment, in particular the DCluster package (Gomes-Rubio et al. 2005).

3.1 Standardized Crime Rates

A first crude approximation of the spatial crime distribution is by calculating a simple risk estimator like the "raw standardized crime rate" (SCR). SCR is defined as O_i/E_i, whereas O_i is the observed number of crimes in census tract i and E_i the expected number. E_i can be estimated as $E_i = POP2000_i * (\sum_{i=1}^{n} O_i / \sum_{i=1}^{n} POP2000_i)$, assuming independence between the crime locations and following a Poisson distribution (Gomes-Rubio et al. 2005). Hence, areas with a SCR above 1 are of interest, because these areas can be considered as high risk areas. Figure 2 visualizes the spatial distribution of the SCR for August 2005 for the study area. In this figure the pattern of high risk areas clearly follows the main traffic axes. Furthermore, the two census tracts with an extremely high SCR value in the south have nearly no residential land use and, therefore, only a marginal number of residential population, leading to more crime locations than inhabitants. It should be mentioned that these findings are also dependent on the observed time-frame. The following Sects. 3.2, 3.3, and 3.4 discuss the results for each algorithm.

3.2 Geographical Analysis Machine

The GAM (Openshaw et al. 1987) is the first search method attempting to explore local clusters in point and lattice data sets. The study region is superimposed by a regular grid, and circles with a user specified radius r are drawn around each grid cell. The observed number of crime locations within each circle are compared with its expected number under a Poisson distribution. Cells having a significantly higher than expected number of cases within a circle are saved. Crime clusters are thus shown cell-based on the basis of overlapping circles. This approach has suffered much criticism (Besag and Newell 1991). The major criticism is that GAM does not control for multiple testing, when GAM was first proposed. To counteract this main critique, Openshaw et al. (1987) advocate a significance value of 0.002 to

detect apparent clusters, which was used in this study as well when running the GAM. This means that circles with a significance value below 0.002 are declared as significant and therefore saved. Different combinations of parameter settings concerning the grid cell's width (100–1,000 m) and the length of the radii (100–1,000 m) are calculated. Due to a more suitable comparison with the results of the other algorithms (Sect. 4), a polygon-based visualization is applied, whereas each census tract containing at least a significant cell is identified as cluster. As a consequence the resulting clusters can be overestimated in their spatial dimension. Figure 3 shows two examples with different radii.

3.3 Besag and Newell Statistic

Another method to explore local clusters is the BNS method (Besag and Newell 1991). This statistic overcomes the drawback of GAM, examining clusters only based on a predefined and constant distance criterion, which is problematic if the population at risk is heterogeneously distributed in space and spatial objects are represented by centroid locations. BNS takes the second issue into account by requiring an a-priori user-defined cluster size k. Thus, the scanning circles, centered over each census tract, are expanded until a size of k objects is reached within a cluster. Under the null hypotheses of uniform risk (no clustering), each potential cluster is evaluated by a test statistic L_i concerning its significance. As above, the overall level of simultaneous testing is not controlled for. Researchers (e.g. Fotheringham and Zhan 1996) criticize this approach because of its ad hoc choice of a value for the parameter k. In this application the significance level has been set to 0.05. Figure 4 visualizes the detected spatial clusters.

3.4 Spatial Scan Statistic

The spatial scan statistic (SSS, Kulldorff (1997)) is a relatively novel methodology. The method can be described as follows: A two dimensional circular scan window moves from census tract i to census tract j throughout the study site and compares the probability of being a case inside the circle given the residential population at risk inside this circle as well as the probability of being a case given the residential population at risk outside (Gomes-Rubio et al. 2005). To detect clusters of different sizes, the scan window continuously increases its size up to a maximum of 20% of the total population falling inside each circle used in SSS. Due to the lack of prior knowledge about the resulting cluster sizes, all potential cluster sizes are being considered, when using this statistic (Kulldorff 1997). For each window size a Poisson-based likelihood function is calculated. The most likely cluster is the one that maximizes this likelihood function. The level of significance is usually set to $\alpha = 0.05$ and a parametric bootstrap of 999 simulation runs based on a

Poisson distribution is used to determine local significance (Gomes-Rubio et al. 2005). Compared to the above mentioned methods, the SSS accounts for multiple hypotheses testing. The result is presented in Fig. 5.

4 Discussion of the Results and Conclusions

This paper compares different methods used for exploratory analysis to detect spatial clusters of lattice data. Of the three methods tested, GAM and BNS are very sensitive with regard to the chosen parameters. Table 1 underpins these impressions and compares different aspects of the detected clusters, like the number of census tracts marked as clusters and their size in square miles. For instance, increasing the circle size of GAM leads to larger cluster surfaces. In detail, an increase of the GAM's search radius from 500 m to 1,000 m nearly doubles the detected clusters. Similarly, the setting of different k values of BNS results in different cluster morphologies. Compared to the GAM, BNS finds a lower number of significant clusters, whereas an increase in the parameter k results in more significant clusters. The most likely cluster of the SSS overlaps partially with the results of the GAM but lies spatially in-between the clusters found by the BNS. One advantage of the SSS algorithm is the distinction between a cluster center, having the highest probability of being a cluster, and its assigned parts with a lower but still significant probability. Overall, the number of significant census tracts marked as clusters by the SSS varies considerably, which is of course reflected in the amount of crime affected population, shown in Table 1.

Finally, the correlation between the detected clusters is estimated by Cramer's V. This coefficient of contingency measures the strength of the relationship between two nominal scaled variables, whereas 0 means no correlation and 1 a perfect correlation. The results are presented in Table 2. BNS and GAM result in

Table 1 Number of census tracts including clusters

	Number of clusters	Area (sqm.)	Population
GAM ($r = 500$ m)	84	95.9	348,654
GAM ($r = 1{,}000$ m)	165	183.1	758,772
BNS ($k = 20$)	19	15.3	41,449
BNS ($k = 30$)	23	17.9	70,219
SSS	87	77.5	345,934

Table 2 Correlations between the detected clusters

	SSS	BNS ($k = 20$)	BNS ($k = 30$)	GAM ($r = 500$ m)
BNS ($k = 20$)	0.110			
BNS ($k = 30$)	0.208	0.299		
GAM ($r = 500$ m)	0.190	0.348	0.427	
GAM ($r = 1{,}000$ m)	0.260	0.245	0.275	0.606

roughly similar results, having a higher correlation coefficient. Furthermore, the SSS correlates only slightly with the results of GAM and BNS. An explanation is that the SSS method resulted in one large compact cluster located in the central northern part and literally no clusters found in other parts of the study area.

Overall, the SSS seems most useful, because it has a comprehensive theoretical statistical background, shows the highest flexibility, in terms of availability of data models (e.g., Poisson-based, exponential model) and expandability (e.g., comprise further covariates, space-time analysis), without leaving it up to the user to make subjective decisions about important parameter settings. It is clear that this on-going research is not free of limitations. The main limitation is that a comparison between the cluster results from the different methods was purely based on a visual level and by using a simple correlation measure. Nevertheless, our pragmatic suggestion is to apply different techniques when trying to identify significant clusters and to have the most confidence in those clusters that have been detected by the majority of the cluster algorithms being used. Additional research is required to test and compare alternative cluster methods with each other. Only then can a more complete understanding of an adequate usage of such methods being made.

However, the comparison of these exploratory techniques using a real data set should sharpen the crime mappers awareness concerning the importance of choosing an appropriate method for strategic planning and decision-making because different methods provide different insights into the data.

References

Anselin L (1999) Local indicators of spatial association. LISA Geogr Anal 27:93–115

Besag J, Newell J (1991) The detection of clusters in rare diseases. J Royal Stat Soc A 154:143–155

Fotheringham A, Zhan F (1996) A comparison of three exploratory methods for cluster detection in spatial point patterns. Geogr Anal 28:200–218

Gomes-Rubio V, Ferrandiz J, Lopez-Quilez A (2005) Detecting clusters of diseases with R. J Geogr Syst 7:189–206

Knox E (1989) Detection of clusters. In: Elliott P (ed) Methodology of enquiries into disease clustering. Small Area Health Statistics Unit, London

Kulldorff M (1997) A spatial scan statistic. Commun Stat Theory Methods 26:1481–1496

Kulldorff M, Tango T, Park P (2003) Power comparisons for disease clustering tests. Comput Stat Data Anal 42:665–684

Openshaw S, Charlton M, Wymer C, Craft A (1987) A mark 1 geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inform Sci 1:335–358

Song C, Kulldorff M (2003) Power evaluation of disease clustering tests. Int J Health Geogr 2(9)