Optimizing the Unit of Analysis of Environmental Criminology Studies through Spatial Cluster Detection in Survey Data

Christian Kreis
Netherlands Institute for the Study of Crime and Law
Enforcement
De Boelelaan 1077a
Amsterdam, Netherlands
ckreis@nscr.nl

Abstract

The current research proposes a non-parametric method to delineate spatial units of analysis for environmental criminology studies that comprise a minimum of data variability for a given area size. Kernel smoothing techniques serve to estimate the spatial variation of the intensity of a marked point pattern process in order to create area-level units of analysis that are as homogeneous as possible. Such “data-driven” spatial clusters are more likely to unearth neighborhood effects and enable researchers to delineate optimal areas for analysis when faced with a trade-off between data heterogeneity and neighborhood-level sample size. The geovisualization of the results for better communication with practitioners and intelligence-based decision making forms an important part of the study.

Keywords: Unit of analysis; Neighborhood effects; Kernel density estimation; Monte Carlo simulation.

1 Introduction

Choosing the neighborhood-level unit of analysis is an inevitable part of any multi-level analysis in environmental criminology. Crucial though this aspect of a study design is for its results, the question of the appropriate level of aggregation has only recently received more attention [7]. The upshot of this debate has been that smaller units of analysis are better because they are more homogeneous and thus a more accurate measure of neighborhood conditions [4]. Still, the default choice is to use administrative boundaries of convenience such as census tracts or postal ZIP code districts as neighborhood-level units of analysis.

The current study contends that while smaller is undoubtedly better, in studies using survey data more often than not a trade-off exists between neighborhood-level sample size and heterogeneity [4]. Moreover, if the areas delineated by administrative boundaries are heterogeneous, any gain in statistical power from an increased sample size risks being lost by the increase in variability.

The current study thus proposes to use methods developed in spatial epidemiology for the analyses of point pattern data [2, 3, 6] to delineate spatial units of analysis based on the local intensity of the phenomenon under study rather than administrative boundaries. Such “data-driven” neighborhoods it is contended are less likely to mask significant neighborhood effects and enable researchers to make a more informed choice when faced with a trade-off between sample size and data variability.

2 Data and Methodology

The data to test this proposition come from the 2011 Swiss Crime Survey (SCS), a large-scale longitudinal victimization survey. In 2011, the SCS data has been geocoded based on each survey respondent’s home address, which first makes it possible to aggregate individual respondents by variable geometry independently of any administrative boundaries. For the current study, the two Swiss cities of Bern and Zurich serve as testing ground, where 508 and 504 local residents have been interviewed, respectively.

The methodology to identify areas of high or low intensity of a given phenomenon is derived from a case-control study approach in spatial epidemiology, which tries to estimate the spatial variation in the relative risk of disease. In close analogy, individual survey respondents who exhibit a certain trait are considered “cases”, with the remainder serving as controls. Cases and controls are treated as realizations of two separate inhomogeneous poisson processes (IPP), and Kernel smoothing techniques are used to estimate the spatial variation in their intensity [1, 2].

In order to estimate the intensity of the IPP of the cases $\lambda_1(x)$ and the controls $\lambda_0(x)$, the study uses a standard Gaussian kernel with the same bandwidth, which is determined through cross-validation. The estimate of the local prevalence of a given phenomenon then is the relative risk or the log-ratio of the local intensities of the two IPP. In order to control for the influence of demographic characteristics of survey respondents, the study uses stratified random sampling to match the distribution of demographic covariates among the cases and controls [3].

The statistical significance of the observed spatial pattern is assessed by means of a Monte Carlo simulation, in which the assignment of cases and controls between the point locations is shuffled through random permutation for each iteration. The algorithm then checks for each cell of the grid used in the Kernel smoothing procedure the number of iterations for which the log-ratio of the intensity of the simulated IPP is above the observed value. Contour lines then serve to delineate the areas for which the associated p-value lies above or below a given threshold value (e.g. $p<0.05$, $p>0.95$) [1, 3].

In order to evaluate the validity of the spatial clusters thus established, the log-ratio of the intensity of the two IPP $\lambda_1(x)$ and $\lambda_0(x)$ is retrieved for each survey respondent and included as an explanatory variable in a multi-level model. The validity of the spatial clusters as the level II units of analysis can then be assessed by comparing the output of two multi-level binary regression models that include either the mean value per administrative unit or the discretized log-ratio of the intensity of the phenomenon under study as explanatory variable at the neighborhood level.

All computations were made using the R Statistical Computing Software [5].

3 Results

Figure 1 illustrates the results of the analysis of the relative intensity of fear of crime for the City of Zurich, measured as the number of survey respondents who are afraid of crime when walking alone at night through their neighborhood. The risk surface of fear of crime reveals some very distinct spatial patterns, with the areas with relatively lower levels of fear being located primarily in the Southern and Western parts of the city. By contrast, the areas with the highest levels of fear of crime are located in the Western and the North Eastern parts, with the areas with moderate levels of fear sandwiched in between.

The visual impression from the risk surface that the spatial extension of the areas of high or low fear is not congruent with the boundaries of the census tracts is confirmed by the quantitative analysis. In the binary regression models, not only is the discretized log-ratio of the intensity of the point process a better predictor than the mean value by administrative unit, the former model fit the data also much better overall.

4 Discussion

The current study assessed the validity of two procedures of aggregating individual-level survey data for multi-level analyses, one based on existing administrative boundaries versus another that identifies spatial clusters of flexible shape based on the local intensity of a marked point process. The study concludes that the debate about the appropriate scale of aggregation for neighborhood-level data misses the main point, if the spatial pattern of a given phenomenon has only a loose resemblance with administrative boundaries. This suggests that constructing area-level units of analysis based on the spatial distribution inherent in the data may help unearth neighborhood effects that would otherwise go unnoticed and enables researchers to better design area-based studies of criminological phenomena.

4.1 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

[1] Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R.* Springer, New York, 2008.

[2] Anthony C. Gatrell, Trevor C. Bailey, Peter J. Diggle, and Barry S. Rowlingson. Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. *Transactions of the Institute of British Geographers, New Series*, 21(1): 256-274, 1996.

[3] J. E. Kelsall and P. J. Diggle. (1998) Spatial variation in risk of disease: a nonparametric binary regression approach. *Journal of the Royal Statistical Society: Series C*, 47(4): 559-573, 1988.

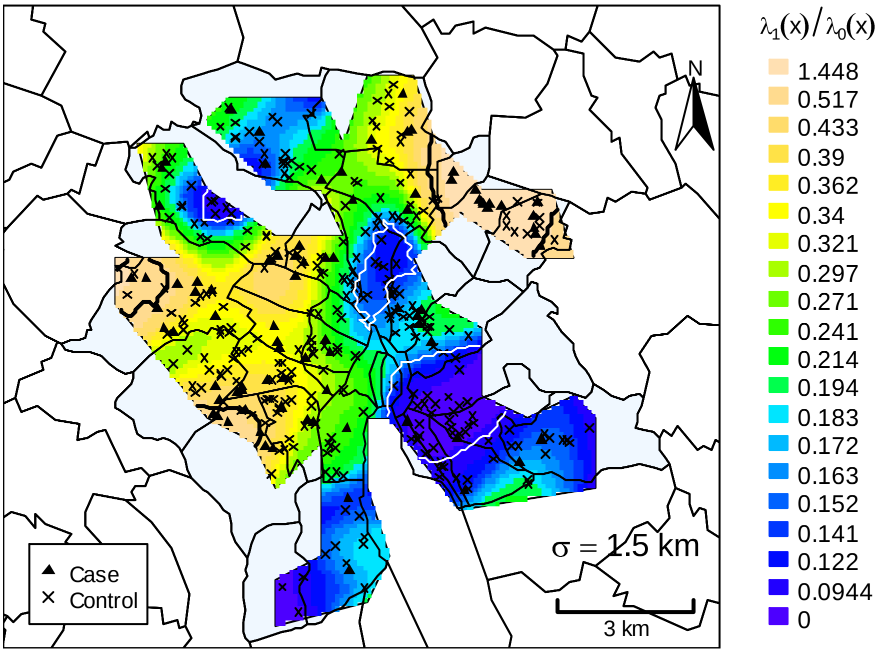
[4] Dietrich Oberwittler and Per-Olof Wikström. Why small is better: Advancing the study of the role of behavioral contexts in crime causation. In D. Weisburd, W. Bernasco, G.J.N. Bruinsma, editors, *Putting Crime in its Place: Units of Analysis in Geographic Criminology*, pages 35-59, Springer, New York, 2009.

[5] R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2012.

[6] Lance A. Waller and Carol A. Gotway. *Applied Spatial Statistics for Public Health Data.* Wiley, Hoboken, 2004.

[7] David Weisburd, Wim Bernasco, Gerben J.N. Bruinsma, editors. *Putting Crime in its Place: Units of Analysis in Geographic Criminology.* Springer, New York, 2009.

Figure 1: Hot/Cold Spots of Fear of Crime in Zurich. Kernel ratio of the intensity of cases and controls using a bandwidth of 1.5 km. The white and thick black lines delineate significant cold or hot spots of fear, respectively, whereas thin black lines indicate the census tract boundaries.



Source: Swiss Crime Survey 2011.