Using Social Media Data to Assess Spatial Crime Hotspots

Nick Malleson (n.malleson06@leeds.ac.uk), School of Geography, University of Leeds
Martin Andresen, School of Criminology, Simon Fraser University, Canada

1 Introduction

The crime rate is often used as a means of understanding the abundance of criminal events over space and time. The *population at risk* is a fundamental component of a crime rate calculation, although accurately estimating this population is fraught with difficulty. It has been recognised for some time that population at risk estimates vary by crime category (Boggs, 1965), but most research resorts to using the residential population due to data availability. However, recent research has identified crime categories for which the residential population does not provide a suitable estimate of the distribution of potential victims (Andresen 2006, 2011; Zhang et al. 2012; Boivin 2013). In many cases, the *ambient population* (i.e. the volume of people visiting an area during the day) represents a more accurate population at risk than the residential (or *night time*) population.

This paper explores the utility of using new 'crowd-sourced' data as a means of estimating the ambient population. Specifically, the data consists of messages from mobile devices (such as smart phones) that are posted to the Twitter social media service. These data are used as population at risk estimates in street crime rate calculations, using two local indicators of spatial association: GI* and GAM. The paper will show that when using the ambient population, an apparent street crime hotspot in the centre of the city disappears.

2 Literature Review

Despite early demonstrations that the population at risk had a considerable impact on some crime types (Boggs, 1965), as well as more recent demonstrations (Andresen 2006, 2011; Zhang et al. 2012; Boivin 2013), the crime science community has yet to converge at a consensus on the appropriate way to measure the population at risk in

crime analysis (Andresen and Jenion, 2010). The problem faced by researchers is that on the one hand the residential population is not a suitable measure of the probability of victimisation but on the other hand very few non-residential population measures exist. Recent attempts to more accurately measure the ambient population make use of the LandScan Global Population Database (Andresen 2006, 2011; Andresen and Jenion 2010; Andresen et al. 2012), although these data have a spatial resolution of approximately 1km^2 which is likely to hid important lower-level patterns (Andresen and Malleson 2011).

Fortunately, the emergence of vast new administrative, social and commercial data sources is inciting interest in new forms of 'crowd sourced' data that have the potential to address some of the fundamental drawbacks associated with the population at risk in crime analysis. In particular, this new information about peoples' daily behaviour may prove to be instructive for understanding urban dynamics and hence developing more accurate estimates of the ambient population.

Even though these data are being used much more regularly, examples the use the geographical locations of social media messages are still rare. The most relevant examples include: the analysis of human mobility patterns (Cheng et al. 2011); the development of neighbourhood boundaries (Cranshaw et al. 2012); the identification of events such as earthquakes (Crooks et al. 2013) and other geographical patterns (Stefanidis et al., 2013) in social media data. However, we are unaware of any research that uses social media data to better understand the risk of criminal victimization.

3 Study area and data overview

The study area is Leeds, United Kingdom (UK). It is of particular relevance that Leeds has a central business and retailing area that attracts large volumes of people and has (as would be expected) high volumes of violent crime relative to surrounding. Data from the 2011 UK census are used to estimate the residential population at the Output Area geography. These data show that relatively few people live in the city centre, upwardly biasing any representations of criminal event risk using the resident population.

The crime data used here were extracted from the `police.uk` service (http://www.police.uk) and represent all occurrences of 'violent crime' – which includes a variety of crime types ranging from minor assaults to serious incidents of wounding and

murder (Flatley 2013b) – recorded by the police in 2011within the Leeds Local Authority District (*N*=10,625).

The crowd-soured data used here consist of messages posted to the Twitter service from within the Leeds between 22nd June 2011 and 14th April 2013 with associated GPS coordinates, *N*=1,955,655. Figure 1 illustrates the density of messages calculated using the Kernel Density Estimation algorithm.

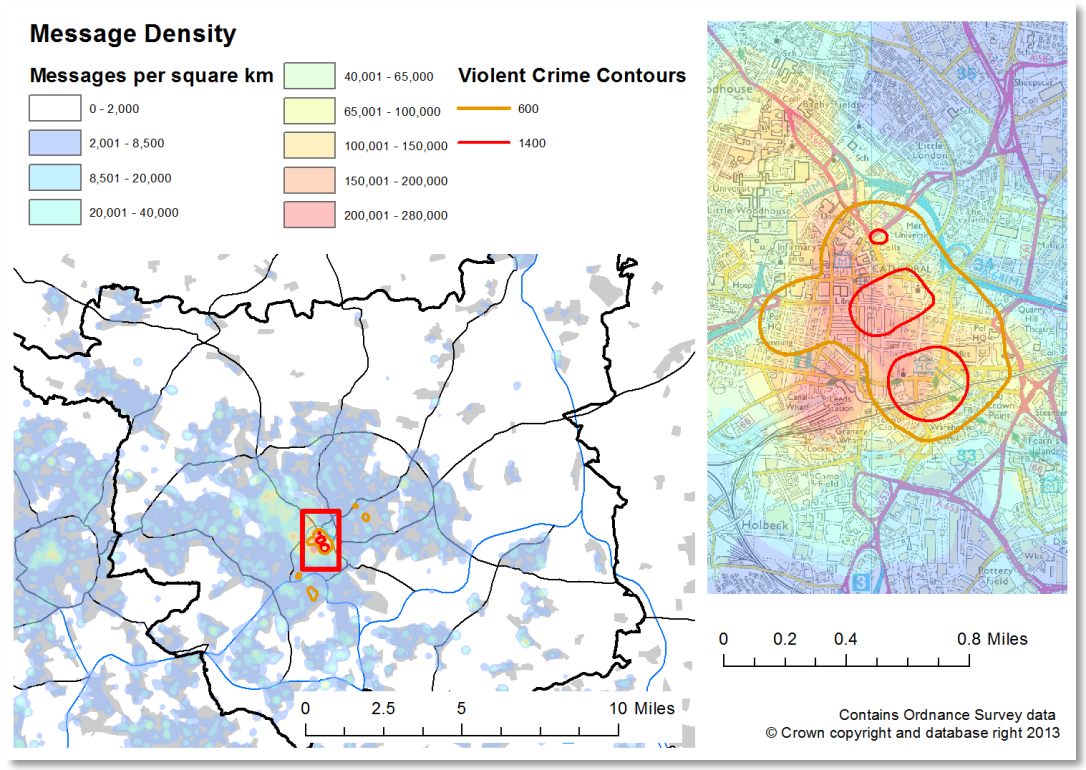

Figure 1 The density of social media messages and crime volume contours (generated from police.uk data) in Leeds.

4 Methods and Results

The aim of this work is to highlight the areas that suffer high rates of crime using both residential (census) and ambient (crowd-sourced) population at risk estimates. Two complementary statistics will be applied to search for statistically significant crime hot spots using the two different population at risk estimates.

The first statistic to be applied is the Getis-Ord GI* (Getis and Ord 1992; Ord and Getis 1995) and the results are illustrated in Figure 2. This is used here because its definition closely matches that of a 'hot spot' – local area averages that are significantly greater

than global averages (Chainey and Ratcliffe 2005) – and has hence become popular within spatial criminological research. Figure 2 maps the GI* indices for the two violent crime rates. The most striking result is that when the ambient population is used to measure the population at risk the statistically significant cluster in the city centre disappears (bottom maps).

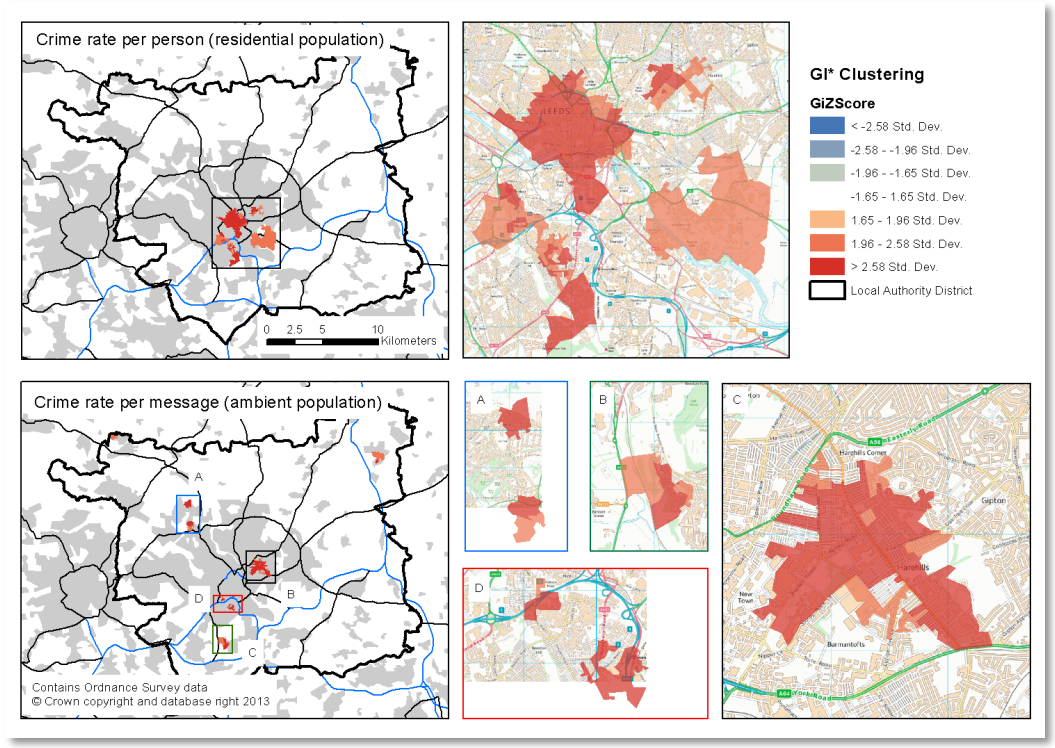

Figure 2. Crime clusters (Z values) calculated using GI* and two different measures of the population at risk. Output areas with insignificant p values ($0.05 < p < 0.95$) are not shown, regardless of their Z value.

A drawback with the GI* statistic is that, as it requires spatial aggregation, it is susceptible to the modifiable areal unit problem (Openshaw 1984). To mediate this problem, the Geographical Analysis Machine (GAM: Openshaw, 1987) algorithm will also be applied – for details about the algorithm see Charlton (2008). In the following analysis, the algorithm was executed multiple times with the search radii increasing in 100m increments from 200m to 1km. A single combined density map was produced from all significant search points at all radii. Hence the most dense areas will be those that have a significant crime volume given the underlying population at risk *at multiple resolutions*. The results are presented in Figure 3.

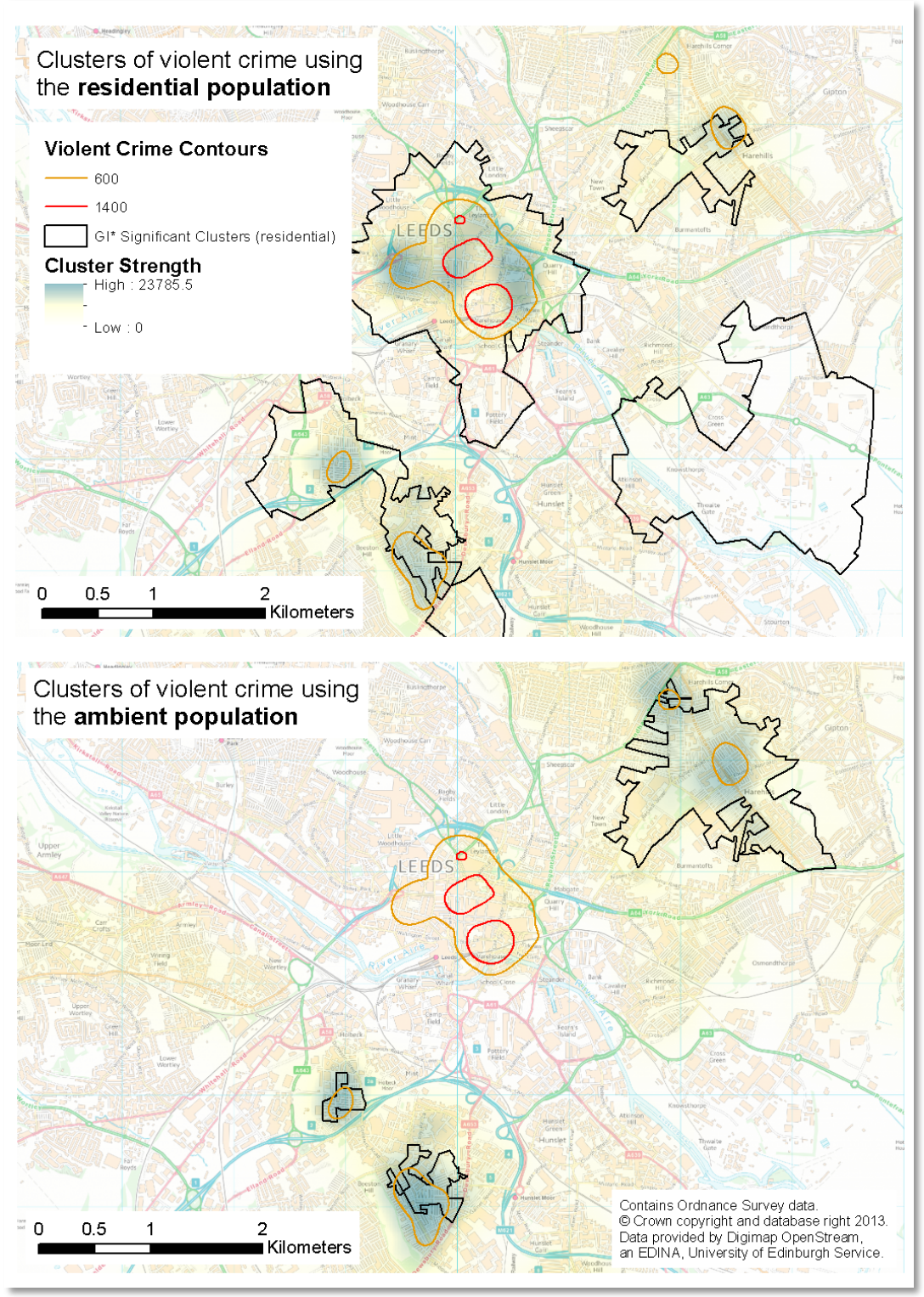

Figure 3 Clusters of violent crime calculated using GAM with ambient and residential population at risk. 'Cluster strength' is the sum of all significant search circles at all radii from 200m to 1km.

The extent to which the GAM outputs are in agreement with those of the GI* analysis us striking. Both techniques reveal broadly similar cluster locations regardless of the population at risk used. Importantly, they continue to suggest that the volume of violent crime in the city centre is only marginally higher than would be expected given the size ambient population.

5 Discussion and conclusions

This research has shown that very different crime hotspots emerge under the application of two different crime rate denominators: the residential population (measured by the 2011 UK census) and the ambient population (measured by the number of spatially referenced messages posted to Twitter). Importantly, the results suggest that the large volume of violent crime in the city centre does not lead to a statistically significant crime rate when the ambient population is used to measure the population at risk. Furthermore there are a small number of neighbourhoods that exhibit large volumes of crime *and* high rates regardless of the population at risk or statistical analysis method used. Explaining these high rates is a clear area for future research.

It is important to be cautious with the assumption that Twitter support generalisations about the general ambient population. In particular, it is not clear which groups of people are not well represented by twitter data – the 'digital divide' might distort these results (e.g. Yu, 2006; Fuchs, 2009) – or what impact a small number of highly prolific Twitter users will have on general patterns.

Assuming these drawbacks can be addressed, or at least better understood, one of the most exciting possible future developments would be to estimate particular sub-populations at risk of particular crime types – such as young people at risk of robbery when they visit bars during the evening. Both the population at risk and the crime events could be temporally disaggregated by day/night, weekday/weekend, and so on. The subsequent application of spatio-temporal cluster hunting algorithms could be applied to create an even more nuanced assessment of genuine hotspot locations. Of course, the use of crowd-sourced data necessarily involves a new set of ethical implications that have yet to be properly addressed. However, if these issues can be overcome there are considerable social benefits that may emerge from a much clearer understanding of the true distribution of crime.

6 References

Andresen, M.A., 2006. “Crime Measures and the Spatial Analysis of Criminal Activity.” *British Journal of Criminology* 46 (2): 258–285.

Andresen, M.A., and G.W. Jenion. 2010. “Ambient Populations and the Calculation of Crime Rates and Risk.” *Security Journal* 23 (2): 114–133.

Andresen, M.A., 2011. “The Ambient Population and Crime Analysis.” *Professional Geographer* 63 (2): 193–212.

Andresen, M.A., and N. Malleson. 2011. “Testing the Stability of Crime Patterns: Implications for Theory and Policy.” *Journal of Research in Crime and Delinquency* 48 (1): 58–82.

Andresen, M.A., G.W. Jenion, and A.A. Reid. 2012. “An Evaluation of Ambient Population Estimates for Use in Crime Analysis.” *Crime Mapping: A Journal of Research and Practice* 4(1): 7–30.

Boggs, S.L. 1965. “Urban Crime Patterns.” *American Sociological Review* 30 (6): 899–908.

Boivin, R. 2013. “On the Use of Crime Rates.” *Canadian Journal of Criminology and Criminal Justice* 55 (2): 263—277.

Cheng, Z., J. Caverlee, K. Lee, and D. Z. Sui. 2011. “Exploring Millions of Footprints in Location Sharing Services.” In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, July 2011, Barcelona, 81–88. Menlo Park, CA: AAAI press.

Chainey, S., and J.H. Ratcliffe. 2005. *GIS and Crime Mapping*. Chichester: John Wiley and Sons.

Charlton, M. 2008, “Geographical analysis machine (GAM)”, in KK Kemp (ed.), *Encyclopedia of geographic information science*, SAGE Publications, Inc., Thousand Oaks, CA, pp. 179-80.

Cranshaw, J., R. Schwartz, J. Hong, and N. Sadeh. 2012. “The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City.” In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (*ICWSM),

May 2012, Dublin, 58 – 65. Menlo Park, CA: AAAI press.

Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski. 2013. "#Earthquake: Twitter as a Distributed Sensor System." *Transactions in GIS* 17 (1): 124–147.

Flatley, J. 2013b. *Crime in England and Wales, year ending September 2012*. London: Office for National Statistics.

Fuchs, C. 2008. The Role of Income Inequality in a Multivariate Cross-National Analysis of the Digital Divide. *Social Science Computer Review* 27: 41–58.

Getis, A., and J.K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24 (3): 189–206.

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

Openshaw, S. 1987. "An Automated Geographical Analysis System." *Environment and Planning A* 19 (4): 431–436.

Ord, J.K., and A. Getis. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." *Geographical Analysis* 27 (4): 286-306.

Stefanidis, A., A. Crooks and J. Radzikowski. 2013. "Harvesting ambient geospatial information from social media feeds." GeoJournal 78: 1–20.

Yu, L. 2006. Understanding information inequality: Making sense of the literature of the information and digital divides. *Journal of Librarianship and Information Science* 38: 229–252.

Zhang, H., G. Suresh, and Y. Qiu. 2012. "Issues in the Aggregation and Spatial Analysis of Neighborhood Crime. *Annals of GIS* 18 (3): 173–183.