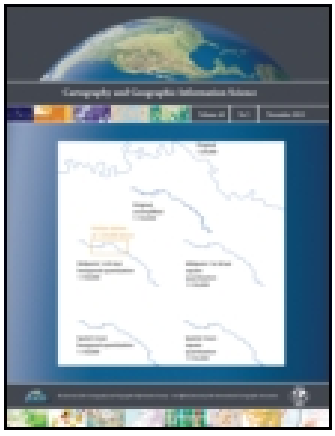

Cartography and Geographic Information Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tcag20

The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns

Nick Malleson[a] & Martin A. Andresen[b]

[a] School of Geography, University of Leeds, West Yorkshire, LS2 9JT, United Kingdom

[b] School of Criminology, Institute for Canadian Urban Research Studies, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6 Canada

Published online: 10 Apr 2014.

Cartography and Geographic Information Science, 2015
Vol. 42, No. 2, 112–121, http://dx.doi.org/10.1080/15230406.2014.905756

The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns

Nick Malleson[a] and Martin A. Andresen[b]*

[a]*School of Geography, University of Leeds, West Yorkshire, LS2 9JT, United Kingdom;* [b]*School of Criminology, Institute for Canadian Urban Research Studies, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6 Canada*

(*Received 28 November 2013; accepted 13 March 2014*)

Crime rate is a statistic used to summarize the risk of criminal events. However, research has shown that choosing the appropriate denominator is non-trivial. Different crime types exhibit different spatial opportunities and so does the population at risk. The residential population is the most commonly used population at risk, but is unlikely to be suitable for crimes that involve mobile populations. In this article, we use "crowd-sourced" data in Leeds, England, to measure the population at risk, considering violent crime. These new data sources have the potential to represent mobile populations at higher spatial and temporal resolutions than other available data. Through the use of two local spatial statistics (Getis-Ord GI* and the Geographical Analysis Machine) and visualization, we show that when the volume of social media messages, as opposed to the residential population, is used as a proxy for the population at risk, criminal event hot spots shift spatially. Specifically, the results indicate a significant shift in the city center, eliminating its hot spot. Consequently, if crime reduction/prevention efforts are based on resident population based crime rates, such efforts may not only be ineffective in reducing criminal event risk, but be a waste of public resources.

Keywords: violent crime; spatial crime analysis; twitter; population at risk

Introduction

The spatially referenced crime rate is a statistic often used to represent the risk of criminal events. Spatially referenced crime rates help to reveal clusters of crime in space and/or time based on an underlying population at risk. However, the choice of an appropriate population at risk is non-trivial. Different crime types have different spatial opportunity sets that necessitate the separate analyses. Similarly, the population at risk varies for different crime rates and should be given the same consideration. As stated by Boggs, "a valid rate … should form a probability statement, and therefore should be based on the risk or target group appropriate for each specific crime category" (Boggs 1965, 900). Despite this importance, most research uses the residential (census) population as the population at risk, primarily because of data availability and constraints in terms of time and money. Although it has been claimed that it matters little which poulation at risk is used in the analysis (Cohen, Kaufman, and Gottfredson 1985), recent research suggests that the residential population is unsuitable as a measure of population at risk for crimes that involve mobile victims such as assaults (Boivin 2013); robbery (Zhang, Suresh, and Qiu 2012); and automotive theft, burglary, and violent crime (Andresen 2006, 2011).

In an attempt to address these limitations, our article utilizes "crowd-sourced" data to measure the ambient population. Specifically, we use messages from mobile devices (such as smart phones) that are posted to Twitter. These data have the potential to represent the ambient population at much higher spatial and temporal resolutions than previous research in spatial crime analysis, although there are also considerble difficulties associated with the data that must be overcome before they can be used by crime analysts in earnest. The research questions are:

(1) Are crime hot spots stable under the application of different population-at-risk measures?
(2) Which areas have the highest crime rates when using both residential (census) and mobile (social media) population-at-risk data?

Related work

The population at risk in crime analysis

Although a number of studies have made attempts (Andresen 2006, 2011; Zhang, Suresh, and Qiu 2012; Boivin 2013), it is needless to say that there is no consensus on the appropriate way to measure the population at risk in the scientific community (Andresen and Jenion 2010). This is partially because there are so few available data sets at a spatial resolution that can be useful to researchers, particularly in the context of spatial crime analysis. Boggs (1965) is the earliest known example to

*Corresponding author. Email: andresen@sfu.ca

systematically show the impact of using different populations-at-risk measures in crime rate calculations. She considered the business/residential land use ratio for business crime, parking space availability for vehicle theft, and sidewalk area (as a proxy for pedestrians) for street robbery. In her subsequent analysis, Boggs (1965) found that her alternative populations at risk mattered a lot for some crime types and very little for other crime types. More recently, Andresen and colleagues have used the LandScan Global Population Database as the population at risk (Andresen 2006, 2011; Andresen and Jenion 2010; Andresen, Jenion, and Reid 2012). The LandScan data provide an estimate of the ambient population, on a global scale, at a spatial resolution of approximately 1 km^2 – this area varies with the distance from the equator. Though largely instructive, there are limitations with these data: (1) the spatial resolution is relatively poor for spatial crime analysis (approximately the size of a census tract) because recent research has shown that analyzing crime at scales greater than the street segment may hide important lower-level patterns (Andresen and Malleson 2011); and (2) the ambient population estimate is a yearly average, such that no account is taken for seasonal variations or the differences in population counts at different times of day. In an attempt to allieviate some of these problems, this article will use data contributed by individuals to social media services to estimate ambient population at risks.

Social media data for mobile populations

In recent years, the emergence of vast new administrative and commercial data sources, coupled with warnings about a "crisis" in an empirical sociology that continued to rely entirely on traditional small studies (Savage and Burrows 2007), has spurred some research to engage with new forms of "crowd-sourced" data to gain insight into social processes. These data, commonly contributed informally by citizens rather than being obtained from a formal survey, are becoming ubiquitous and will undoubtedly have a dramatic impact on future social science research. With respect to population dynamics in particular, traditional large-volume social science data lack information regarding where people are throughout the day, and instead represent the *night time* distribution of the population. A benefit of new forms of crowd-sourced data, and social media in particular, is that new technologies enable researchers to capture large volumes of information regarding peoples' *daily* behavior. This may prove to be instructive for understanding urban dynamics and developing more accurate population-at-risk estimates. And in the context of this article, such data may prove to be useful for spatial crime analysis.

The number of sources for such data is increasing, with the more widely used being Twitter, mobile device data from service providers, public transport usage, Foursquare, Flickr, and Facebook. Research in the United States has found that two-thirds of online adults (66%) use social media platforms (Smith 2011) and that 26% of American Internet users aged 18–29 have been found to use Twitter (Smith and Brenner 2012). Data from these sources are also voluminous. For example, there were supposedly over 100 million active Twitter accounts in 2011 (Twitter 2011) and 270,000 tweets per minute produced worldwide in 2012 (TechCrunch 2012).

Social media data have recently been used for a wide variety of different purposes – a full review of applications would be an extensive undertaking (and one that would be outdated before it is published). However, examples of the application of social media data to the study of social phenomena are more limitted. Examples include research into the fear of missing out (Przybylskia et al. 2013), well-being (Hong et al. 2012), and happiness over time (Bliss et al. 2012). Others make some limitted use of the data, but still resort to traditional sampling methods (see, for example, Fischer and Reuber 2011; Wohn et al. 2013). The most relevant research for this project are those that have started to make use of the *geographical locations* of social media messages, although given the novelty of utilizing these data sources, examples are still rare. Relevant research includes: the mathematical analysis of human mobility patterns (Cheng et al. 2011); the development of neighborhood boundaries based on the characteristics of those who commonly frequent them (Cranshaw et al. 2012); the identification of events such as earthquakes (Crooks et al. 2013) and other geographical patterns (Stefanidis, Crooks, and Radzikowski 2013) in social media data; and the use of Google search trends to estimate the locations of new outbreaks of influenza (Ginsberg et al. 2009). However, we are unaware of any research that uses social media data to better understand the risk of criminal victimization.

Despite their relatively widespread (and increasing) use, these data sources do have limitations. Such data are inherently "messy" in the sense that they are not gathered using a systematic and statistically guided methodology such as a census. As a result, data structures may be poorly defined, missing data are commonplace, and there are no systematic "corrections" for these issues because these data are still so new to research. Additionally, because of these issues, we must also be concerned with generalizability. For example, Li, Goodchild, and Xu (2013), as part of a special issue on mapping cyberspace and social media, found that higher socioeconomic status groups are overrepresented in Twitter and Flickr. This is not inherently problematic, particularly in the current context of measuring populations at risk, because these higher socioeconomic groups *may be* representative of the underlying population distribution, on average. The main difficulty arises in testing

such a hypothesis. However, even if only a portion of the actual population is being captured, the bias inherent in residential populations for measuring the population at risk may be reduced.

Study area and data overview

Leeds and the census data

Our study area is Leeds, United Kingdom (UK). The Leeds local authority district is the third largest in the UK (behind London and Birmingham) with a residential population estimated at 757,655 in 2012 (Office for National Statistics 2012). Leeds has a central business and retailing district with a high concentration of shops, businesses, and entertainment facilities. This district attracts large volumes of people from within Leeds, Bradford, Manchester, and a number of smaller towns/villages on the outskirts of the city. Such areas have long been known to have high levels of crime because they attract large volumes of people (Schmid 1960a, 1960b) and the center of Leeds is no exception; the district has high volumes of violent crime relative to surrounding areas. Related to the alternative population-at-risk literature in spatial crime analysis, relatively few people live in the city center, upwardly biasing any representations of criminal event risk using the resident population.

In order to measure the residential population, we have used the number of people residing in each *Output Area* (OA) at the time of the 2011 UK census. The OA geography is the smallest area for which census statistics are released. Each OA has a recommended size of 125 households, but can vary based on natural boundaries and the presence (or absence) or high-density housing.

Crime data

The criminal event data used in the analyses below include all individual occurrences of violent crime in 2011 within the Leeds Local Authority District (N = 10,625) that were reported to the police. These data were obtained from the police.uk service (http://www.police.uk); all police-recorded criminal events in England and Wales have been available to the public since December 2011, although only 44% of violent crimes were made known to the police (Flatley 2013a). "Violent crime" includes a variety of crime types ranging from minor assaults to serious incidents of wounding and murder (Flatley 2013b). A drawback with these data is that it is not possible to disaggregate the crime type further (for example, it might be advantageous to analyze robbery and assault separately as research has shown that the spatial patterns of spcific crime types can be rather different (Andresen and Linning 2012)). For privacy reasons, the police.uk service aggregates individual crime points to the nearest "anonymous map point" that can be the center of a street segment, a public place, or a commercial building. These points are defined with catchment areas that have at least eight unique postal addresses, approximately the size of a city block. Although such an aggregation process inevitably induces some spatial inaccuracy, the impact is unlikely to influence any results because the direction in which the criminal event points are moved is random in the aggregate. Also, because Leeds is a rather densely populated city, it is unlikely that any individual criminal event points will be displaced far from their actual location. Additionally, we could disaggregate the data temporally, which is an obvious application of social media data because of the availability of the time when messages are posted. We do not undertake such an analysis, and leave it for future research, because the first comparison in the spatial crime analysis context is with how crime data are mapped in the majority of research, an aggregated year.

Social media data

The data used in the current article are messages posted to the Twitter service from within the Leeds local authority district, 22 June 2011 to 14 April 2013. Although there are other social-media services that provide publicly available access to user contributed data (such as Flickr and Foursquare), data for this study originate solely from Twitter. Future work will explore the possibility of including a variety of sources (e.g., Stefanidis, Crooks, and Radzikowski 2013); currently Twitter is by far the most widely used service and it is not clear that the incorporation of additional services is necessary in this application. Because we are interested in the spatial dimension of criminal victimization risk, only messages with associated GPS coordinates have been included. Such data are commonly generated using mobile devices by users who have explicitly opted to publish their present location. A manual inspection of the data revealed that many high-volume accounts were not representing individuals (examples include weather forecasts, car advertisements, etc.). After deleting these data, the number of messages in our sample was almost 2 million, N = 1,955,655. In addition to the location, each individual message contains information regarding the user account, the text itself, and the time of the message. These additional fields allow for the creation of a temporally dynamic population at risk or an exploration of the characteristics of the individuals who make up the general population. Both of these factors could lead to even more accurate risk estimates, although this is not under investigation here and is a direction for future research.

The density of the messages overlaid with violent crime hot spots is shown in Figure 1.[1] As would be expected, message densities are greatest in urban areas and particularly in the city center. This is precisely what

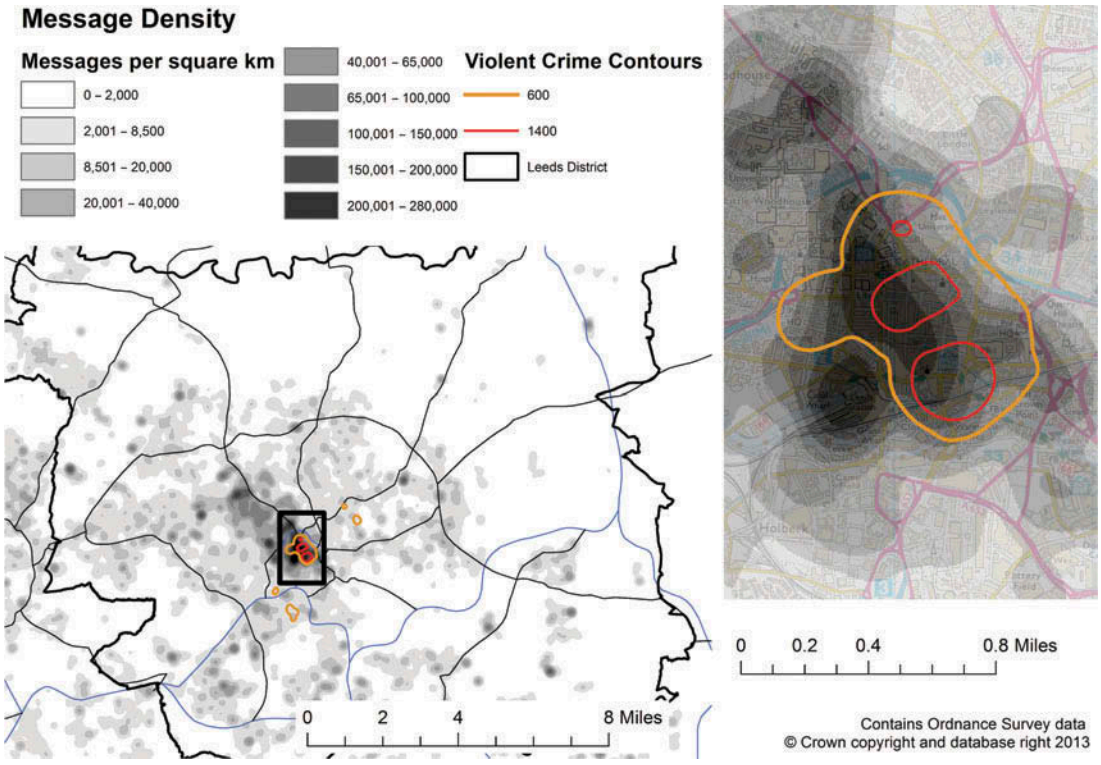

Figure 1. Kernel density of social media messages and violent crime contours. The contours depict the areas with the largest volume of violent crime (densities of 600 and 1400 crimes per km^2, respectively, obtained using Kernel Density Estimation).

would be expected, based on what we know regarding the ambient population. Consequently, as hypothesized above, despite not having a representative sample of individuals based on socioeconomic status, based on local knowledge of the study area, these data may appear to be representative of where people actually are. And, of great interest for the current article, the largest densities of messages appear to coincide with the greatest densities of violent crime – this is not the case with the resident population.

Methods and results

The aim of this article is to highlight the areas that suffer high rates of crime, using both residential and mobile population-at-risk estimates. To answer this question, the research will apply two complementary statistics that can be used to identify clusters in spatial data. Both search for clusters of crime by comparing volumes in individual areas to their surrounding neighbors and to global averages. They are known as Local Indicators of Spatial Association (LISA) and offer the advantage of testing for *statistical significance* of apparent clusters – see Anselin (1995) for a discussion of LISA statistics. Both statistics will be used to search for statistically significant crime hot spots using census data and social media data as the populations at risk.

*Statistic 1: Getis–Ord GI**

The first statistic to be applied is the Getis–Ord GI* (Getis and Ord 1992; Ord and Getis 1995). 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. We use first-order queen's contiguity in the analyses below.

Figure 2 maps the GI* indices for the two violent crime rates. Output areas with insignificant p values ($0.05 < p < 0.95$) are not shown, regardless of their Z value. The distribution of significant GI* scores proves to be instructive. When considering the residential violent crime rate, there is a statistically significant cluster in the city center as well as in some of the surrounding neighborhoods. The violent crime cluster in the city center is expected, particularly because of the low residential population and large volume of criminal events. The surrounding neighborhoods that exhibit clusters of violent crime

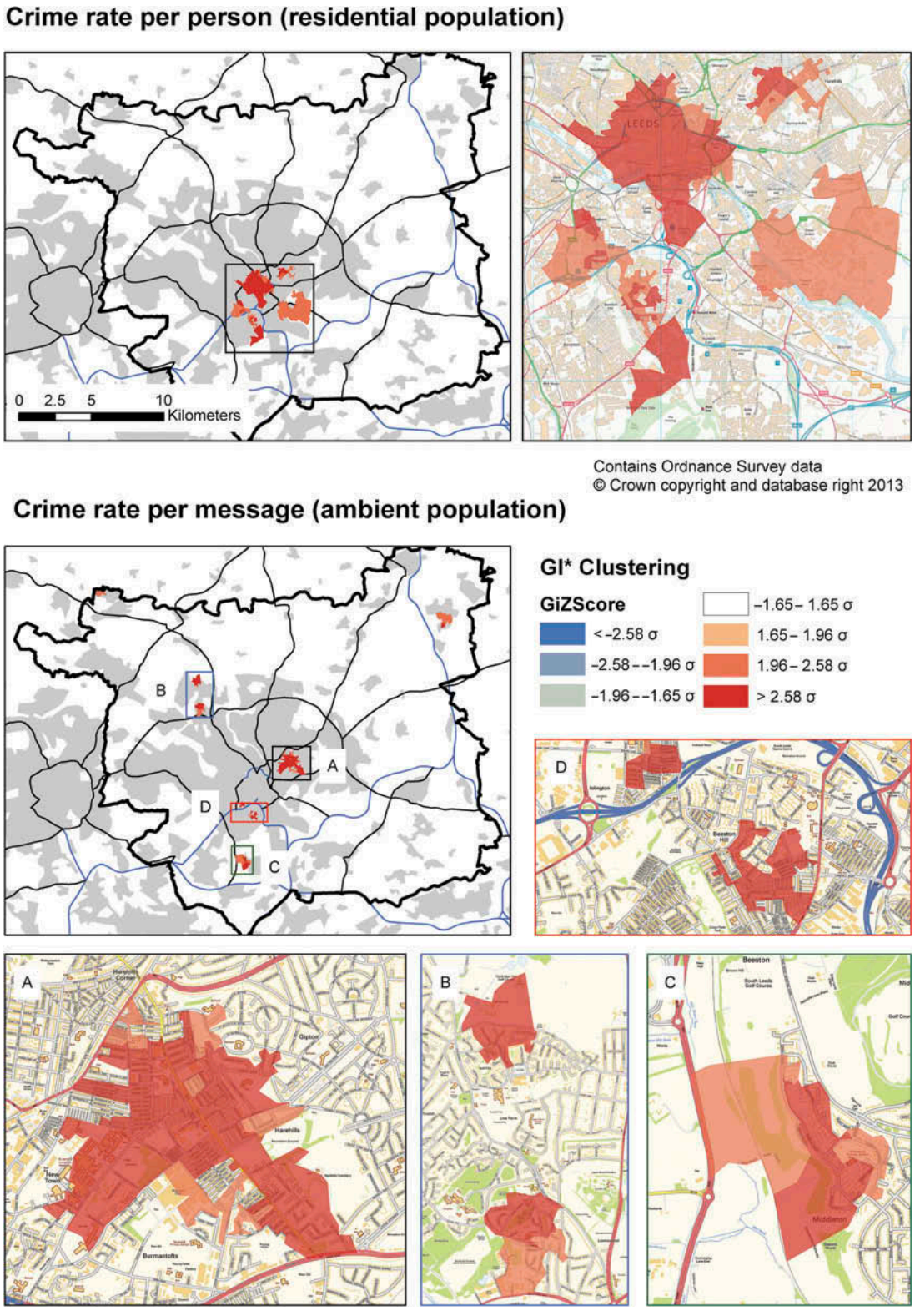

Figure 2. GI* *Z* values for crime rates (using ambient and residential population denominators) in Leeds.

largely consist of industrial estates that also have a low population density. The most notable exceptions are violent crime clusters surrounding a large hospital (St. James's University Hospital) to the north-east and two small areas in neighborhoods to the south-west. It should be noted, however, that the violent crime cluster

surrounding the university hospital may simply be a reporting issue: violent criminal events are coded to occur at this location because this is where they are reported.

A number of violent clusters emerge when using the ambient violent crime rate (see Figure 2 insets A–D). Curiously, none of the violent crime clusters includes the city center area, suggesting the violent crime rate there is not significant when using the ambient population to measure the population at risk. Rather, the violent crime clusters are in diverse neighborhoods with no obvious single explanation for their existence. Each of these neighborhoods may have high violent crime rates given the size of the population at risk. This is clearly a direction for further research.

A drawback with the GI* statistic is that it requires the spatial aggregation of point data into areas (output areas in this case). Therefore, it is susceptible to the modifiable areal unit problem (Openshaw 1984). Hence a second statistic is also used that avoids aggregation to the output area geography in order to further assess the differences in the two violent crime rate calculations.

Statistic 2: the Geographical Analysis Machine

The Geographical Analysis Machine (GAM) (Openshaw 1987) is an algorithm originally developed during research investigating child leukemia cases near a nuclear reactor (Openshaw, Charlton, and Craft 1988). However, GAM has also been applied to research areas such as food poverty (Farrow et al. 2005) and the analysis of crime clusters (Corcoran, Wilson, and Ware 2003). The clustering algorithm operates by iterating over a set of distinct search points that form a regular grid and then calculating the concentrations of events within a given radius of each point. For all search points, i, the algorithm calculates the number of expected events, e_i, standardizing against the underlying background population:

$$e_i = \left(\frac{\sum_0}{\sum_p}\right) p_i, \tag{1}$$

where $\sum_0$ is the total number of observations (crimes), $\sum_p$ is the size of the base population (number of residents or number of messages), and p_i is the size of the base population within search circle i. Then the difference, d_i, is calculated using the actual number of observations, a_i, and the expected number of observations, e_i, that occur within circle i:

$$d_i = a_i - e_i. \tag{2}$$

If a larger number of cases are found than would be expected ($d_i > 0$), a Poisson test for statistical significance is performed. The test calculates the probability that the number of observed events is the same as the number of expected events ($d = 0$). If this probability is lower than a set threshold – in this case the threshold is 0.0099 – then the null hypothesis is rejected and the difference is statistically significant at the specified threshold. In these cases the search circle is stored as a potential cluster. The GAM output is a list of search points and the difference between the expected and actual number events (d_i) when d_i is statistically significant.

This algorithm has been chosen to comlement the GI* analysis because, importantly, it minimizes the impact of the modifiable areal unit problem by defining arbitrary search locations on a regular grid and also by varying the search radius for each search point. In this manner, clusters that appear at one resolution can be discarded if they disappear at others. A further advantage of the GAM algorithm is that it will process raw point data directly – spatial aggregation is not a prerequisite.

In the following, multiple analyses were run with the search radii being increased in 100 m increments from 200 m to 1 km. All significant search points at all radii were used to generate a single density map. The difference between the expected and actual numbers of crimes at each search point (i.e., the output of the algorithm) was used to calculate the density. In this manner, the most dense areas will be those that have a large difference *at multiple resolutions*. Clusters that are only significant at a small number of search radii will add marginally to the density of their area. The results are mapped in Figure 3.

The first notable result is that the GAM outputs are largely in agreement with those of the GI* analysis. Both techniques reveal broadly similar cluster locations regardless of the population at risk used. Considering the number of social media messages, the large volume of violent crime in the city center is only marginally higher than would be expected given the ambient population. In other words, the risk of violent criminal victimization is not particularly high at the city center. However, the algorithms both identify violent crime clusters in neighborhoods to the north- and south-east regardless of the population at risk used. The consistency with which these areas have been identified as crime hot spots suggests that they are indicitative of an exceptionally high volume of crime, whereas the city center hot spot is more likely to be an artifact of the size of the ambient population.

Discussion and conclusions

In this analysis, we have shown that different spatial patterns of crime rates emerge when using two different population-at-risk measures: the residential population (measured by the 2011 UK census) and the ambient population (measured by counting the number of messages

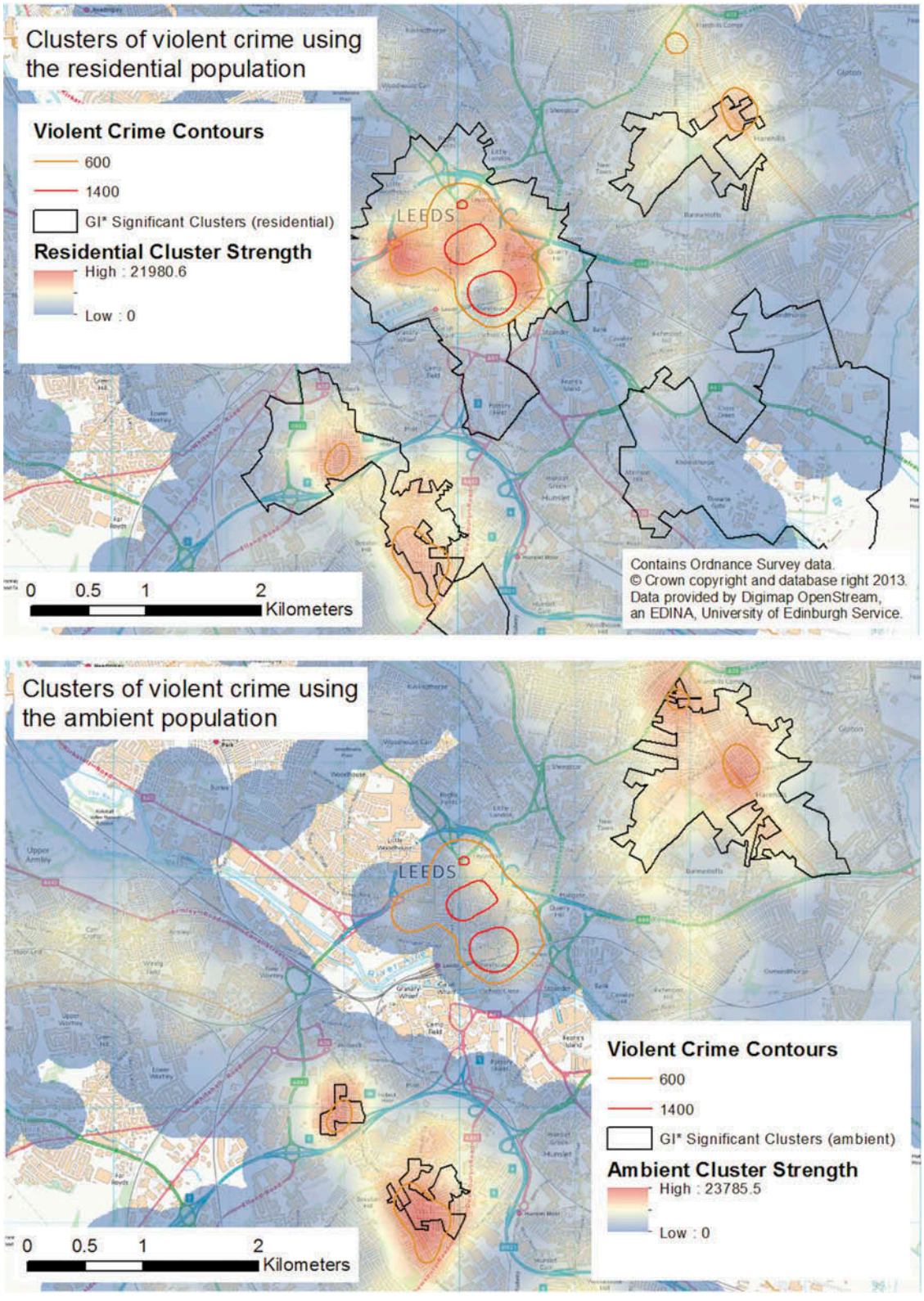

Figure 3. Clusters of violent crime calculated using the Geographical Analysis Machine with ambient and residential population at risk. "Cluster strength" is the sum of all significant search circles at all radii from 200 m to 1 km.

posted to the Twitter social media service). One may say that such a conclusion is an obvious one, but it is important to recognize that the use of an ambient population measure is justified by theory as well as previous empirical research despite the widespread use of the residential population in geography of crime literature. Perhaps most striking are the results from the Leeds city center. Though this area has a large volume of violent criminal events, it does not exhibit a statistically significant rate when the ambient population is used to measure the population at risk. Consequently, despite the high volume of violent criminal events, there is not a statistically significant elevation in risk of violent criminal victimizaton when considering a theoretically informed population at risk. No such conclusion would have been reached with the residential population.

Additionally, there are a small number of neighborhoods very close to the city center that exhibit significantly high violent crime rates when considering both populations at risk, regardless of the clustering method. There is no obvious reason for such high rates of violent crime. These neighborhoods score rather high on the deprivation scale, with two of the neiborhorhoods scoring 114 and 128, highest in England out of a total of 32,482 neighborhoods. Given that deprivation is a highly complex phenomenon, considering a multitude of social factors, it may be the case that this plays some role through (a lack of) oppportunity in terms of legitimate activities for residents social tension that leads to violence. This is clearly an area of future research interest as well.

Though we have had some interesting, and theoretically expected, results, our analysis is not without its limitations. Most specifically, we must be cautious with the use of Twitter data and making generalizations about general population movements. How well do the spatial locations of social media messages reflect the actual spatial locations of the ambient population, in general? We know that some socioeconomic groups are overrepresented in these data, but is this necessarily a problem? Also, to what extent does multiple-counting (users of Twitter who frequently tweet) bias the spatial distribution of the population at risk? These users may simply tweet in locations where there are more people anyway, not causing any spatial bias, or they may make it appear as though more people are present than actually are present. Additionally, despite the user rates of social media are increasing, the percentage of messages that include accurate geogrpahic information are as low as 1–2% (Leetaru et al. 2013; Gelernter and Mushegian 2011). Finally, there is the potential for participation inequality stemming from the differences in the prevalence of social media useage across different social groups. A body of work has explored the impacts of the "digital divide" (e.g., Yu 2006; Fuchs 2008) and it is possible that the higher crime rates identified in the north-east and south-west neighborhoods are an artifact of lower Twitter usage in these relatively deprived communities. However, it is not clear how well general trends in digital access are reflected in Twitter usage – further research is required to establish whether or not the ambient population in these neighborhoods is poorly represented by Twitter data. The persistance of the hot spots regardless of the population at risk used here does, however, add strength to the results.

In general, there are potential problems that must be investigated for the appropriate use of crowd-sourced data. However, if they can be resolved, there is great potential, particularly for spatial crime analysis. For example, Twitter data, or social media data more generally, could be used to estimate particular sub-populations at risk of particular crime types such as young people who visit bars during the evening. Therefore, the population at risk could be tailored according the the most likely victims of a particular crime category to answer the call made by Boggs (1965) almost 50 years ago: "the risk or target group appropriate for each specific crime category" (Boggs 1965, 900).

As discussed by Savage and Burrows (2007), the social sciences (spatial or not) must embrace these new forms of data that, although messy, biased and noisy, have the potential to describe social phenomena better than well-organized small surveys or even national censuses. Mayer-Schonberger and Cukier (2013) share this view:

> One of the areas that is being most dramatically shaken up by N = all is the social sciences. They have lost their monopoly of making sense of empirical social data, as big data analysis replaces the highly skilled survey specialists of the past… When data are collected passively while people do what they normally do anyway, the old biases associated with sampling and questionnaires disappear. (30)

We are confident that the messy, biased and noisy aspects of big data will soon be reduced for confident use in the social sciences. Though they may not disappear or be at the same low level as with more formal data gathering techniques, these limitations may simply become outweighed by the sheer volume of crowd-sourced data and the ways in which it can be utilized. We were able to obtain nearly 2 million individual datum with a minimal setup time and negligible financial cost. Also, with increased use and demand for such data, the providers of social media may very well enhance the quality of their data and metadata because they will realize the value of their commodity. We have argued above that its utility is significant for spatial crime analysis.

Future research in the area of spatial crime analysis has a number of obvious directions. The most obvious is to disentangle these data by day/night, weekday/weekend (or simply day of week), and so on. For this to be successful, a more nuanced definition of the crime type than that

provided by the police.uk data will be necessary – individual police forces do capture these data and might make it available for research purposes. This would allow for the identification of theoretically informed crime rates to be used for clusters in space and time. Additionally, with the possibility of linking social media users back to their home census geography unit, we could generate a profile from those census data of populations at risk in different locations. But of course, such research necessarily involves a new set of ethical implications that have yet to be properly addressed. However, if these ethical issues can be overcome and the public can see the social benefits that may emerge from this research, we may be able to significantly advance our knowledge of the spatial patterns of crime.

Note

1. The density per unit area is used in order to facilitate subsequent comparisons in the paper. Violent crime contours are present in order to show the overlap of violent crime with the messages.

References

Andresen, M. A. 2006. "Crime Measures and the Spatial Analysis of Criminal Activity." *British Journal of Criminology* 46 (2): 258–285. doi:10.1093/bjc/azi054.

Andresen, M. A. 2011. "The Ambient Population and Crime Analysis." *The Professional Geographer* 63 (2): 193–212. doi:10.1080/00330124.2010.547151.

Andresen, M. A., and G. W. Jenion. 2010. "Ambient Populations and the Calculation of Crime Rates and Risk." *Security Journal* 23 (2): 114–133. doi:10.1057/sj.2008.1.

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.

Andresen, M. A., and S. J. Linning. 2012. "The (In) Appropriateness of Aggregating across Crime Types." *Applied Geography* 35 (1–2): 275–282. doi:10.1016/j.apgeog.2012.07.007.

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. doi:10.1177/0022427810384136.

Anselin, L. 1995. "Local Indicators of Spatial Association – LISA." *Geographical Analysis* 27 (2): 93–115. doi:10.1111/j.1538-4632.1995.tb00338.x.

Bliss, C. A., I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds. 2012. "Twitter Reciprocal Reply Networks Exhibit Assortativity with Respect to Happiness." *Journal of Computational Science* 3 (5): 388–397. doi:10.1016/j.jocs.2012.05.001.

Boggs, S. L. 1965. "Urban Crime Patterns." *American Sociological Review* 30 (6): 899–908. doi:10.2307/2090968.

Boivin, R. 2013. "On the Use of Crime Rates." *Canadian Journal of Criminology and Criminal Justice/La Revue Canadienne De Criminologie Et De Justice Pénale* 55 (2): 263–277. doi:10.3138/cjccj.2012-E-06.

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

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)*, Barcelona, July, 81–88. Menlo Park, CA: AAAI press.

Cohen, L. E., R. L. Kaufman, and M. R. Gottfredson. 1985. "Risk-Based Crime Statistics: A Forecasting Comparison for Burglary and Auto Theft." *Journal of Criminal Justice* 13 (5): 445–457. doi:10.1016/0047-2352(85)90044-3.

Corcoran, J. J., I. D. Wilson, and J. Ware. 2003. "Predicting the Geo-Temporal Variations of Crime and Disorder." *International Journal of Forecasting* 19 (4): 623–634. doi:10.1016/S0169-2070(03)00095-5.

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)*, Dublin, May, 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. doi:10.1111/j.1467-9671.2012.01359.x.

Farrow, A., C. Larrea, G. Hyman, and G. Lema. 2005. "Exploring the Spatial Variation of Food Poverty in Ecuador." *Food Policy* 30 (5–6): 510–531. doi:10.1016/j.foodpol.2005.09.005.

Fischer, E., and A. R. Reuber. 2011. "Social Interaction Via New Social Media: (How) Can Interactions on Twitter Affect Effectual Thinking and Behavior?" *Journal of Business Venturing* 26 (1): 1–18. doi:10.1016/j.jbusvent.2010.09.002.

Flatley, J. 2013a. *Focus On: Violent Crime and Sexual Offences, 2011/12*. London: Office for National Statistics.

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. doi:10.1177/0894439308321628.

Gelernter, J., and N. Mushegian. 2011. "Geo-Parsing Messages from Microtext." *Transactions in GIS* 15 (6): 753–773.

Getis, A., and J. K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24 (3): 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski1, and L. Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457: 1012–1014. doi:10.1038/nature07634.

Hong, L., A. Ahmed, S. Gurumurthy, A. Smola, and T. Kostas. 2012. "Discovering Geographical Topics in the Twitter Stream." Proceedings of the 21st International Conference on World Wide Web, Lyon, 769–778.

Leetaru, K., S. Wang, A. Padmanabhan, and E. Shook. 2013. "Mapping the Global Twitter Heartbeat: the Geography of Twitter." *First Monday* 18 (5). doi:10.5210/fm.v18i5.4366.

Li, L., M. F. Goodchild, and B. Xu. 2013. "Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr." *Cartography and Geographic Information Science* 40 (2): 61–77. doi:10.1080/15230406.2013.777139.

Mayer-Schonberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Office for National Statistics. 2012. Mid-2012 Population Estimates. Accessed November 28, 2013. http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-england-and-wales/mid-2012/mid-2012-population-estimates-for-england-and-wales.html

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.

Openshaw, S., M. Charlton, and A. Craft. 1988. "Searching for Leukaemia Clusters Using A Geographical Analysis Machine." *Papers in Regional Science* 64 (1): 95–106. doi:10.1111/j.1435-5597.1988.tb01117.x.

Ord, J. K., and A. Getis. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and An Application." *Geographical Analysis* 27 (4): 286–306. doi:10.1111/j.1538-4632.1995.tb00912.x.

Przybylskia, A. K., K. Murayamab, C. R. DeHaanc, and V. Gladwelld. 2013. "Motivational, Emotional, and Behavioral Correlates of Fear of Missing Out." *Computers in Human Behavior* 29 (4): 1841–1848. doi:10.1016/j.chb.2013.02.014.

Savage, M., and R. Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41 (5): 885–899. doi:10.1177/0038038507080443.

Schmid, C. F. 1960a. "Urban Crime Areas: Part I." *American Sociological Review* 25 (4): 527–542. doi:10.2307/2092937.

Schmid, C. F. 1960b. "Urban Crime Areas: Part II." *American Sociological Review* 25 (5): 655–678. doi:10.2307/2090139.

Smith, A. 2011. *Why Americans use social media*. Technical report, Pew Research Centre. Accessed November 28, 2013. http://www.pewinternet.org/Reports/2011/Why-Americans-Use-Social-Media.aspx

Smith, A., and J. Brenner. 2012. *Twitter Use 2012*. Technical report, Pew Research Center. Accessed November 28, 2013. http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx

Stefanidis, A., A. Crooks, and J. Radzikowski. 2013. "Harvesting Ambient Geospatial Information from Social Media Feeds." *Geojournal* 78: 319–338. doi:10.1007/s10708-011-9438-2.

TechCrunch. 2012. "Analyst: Twitter Passed 500M Users In June 2012." Accessed January 19, 2013. http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

Twitter. 2011. "One Hundred Million Voices." Twitter Blog. Accessed January 2014. https://blog.twitter.com/2011/one-hundred-million-voices

Wohn, D. Y., N. Ellison, M. L. Khan, R. Fewins-Bliss, and R. Gray. 2013. "The Role of Social Media in Shaping First-Generation High School Students' College Aspirations: A Social Capital Lens." *Computers & Education* 63: 424–436. doi:10.1016/j.compedu.2013.01.004.

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. doi:10.1177/0961000606070600.

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. doi:10.1080/19475683.2012.691901.