Journal of Criminal Justice 46 (2016) 52–63

Contents lists available at ScienceDirect

Journal of Criminal Justice

Exploring the impact of ambient population measures on London crime hotspots

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

[a] School of Geography, University of Leeds, Leeds LS2 9JT, UK
[b] Institute for Canadian Urban Research Studies, School of Criminology, Simon Fraser University, 8888 University Drive, Vancouver V5A 1S6, Canada

ARTICLE INFO

Article history:
Received 22 December 2015
Received in revised form 7 March 2016
Accepted 8 March 2016
Available online xxxx

Keywords:
Ambient population
Crime pattern analysis
Clustering
Social media

ABSTRACT

Purpose: Crime analysts need accurate population-at-risk measures to quantify crime rates. This research evaluates five measures to find the most suitable *ambient* population-at-risk estimate for 'theft from the person' crimes.
Method:
1. Collect 'ambient' datasets: the 2011 Census, aggregate mobile telephone locations, and social media.
2. Correlate the population measures against crime volumes to identify the strongest predictor.
3. Use the G_i^* statistic to identify statistically significant clusters of crime under alternative denominators.
4. Explore the locations of clusters, comparing those that are significant under ambient and residential population estimates.

Results and Discussion: The research identifies the Census *workday* population as the most appropriate population-at-risk measure. It also highlights areas that exhibit statistically significant rates using both the ambient and residential denominators. This hints at an environmental backcloth that is indicative of both crime generators and attractors – i.e. places that attract large numbers of people for non-crime purposes (generators) as well as places that are used specifically for criminal activity (attractors). Regions that are largely residential and yet only exhibit hotspots under the ambient population might be places with a higher proportion of crime attractors to stimulate crime, but fewer generators to attract volumes of people.

Introduction

Crime 'hotspots' can be defined as "those areas where the local averages (e.g. concentrations of crime) are significantly different to the global averages" (Chainey & Ratcliffe, 2005 p. 164). Hotspots indicate the times and/or places that exhibit the largest relative volumes of crime, and should therefore be prioritised for intervention. However, raw counts of crime taken in isolation disregard the impacts of the population-at-risk of crime victimisation. Hence crime *rates* are often used to quantify hotspot relative to the population-at-risk. The residential population is the most commonly-used denominator. However, for some types of crime such as assaults (Boivin, 2013), robbery (Zhang et al., 2012) and violent crime (Andresen, 2011), this denominator is not suitable as it does not reliably quantify the size of the potential victim population. Daily flows of people will radically alter the characteristics of urban spaces and recent research has shown that these mobile populations can have a substantial impact on crime rates (Andresen & Jenion, 2010; Felson & Boivin, 2015; Stults & Hasbrouck, 2015).

* Corresponding author.
E-mail addresses: n.s.malleson@leeds.ac.uk (N. Malleson), andresen@sfu.ca (M.A. Andresen).

New data are becoming available that offer an alternative to the traditional residential crime rate. The aim of this research is to identify, from a suite of new population measures, the most reliable denominator for an inherently non-residential crime: *theft from the person*. This aim will be accomplished as follows:

1. Identify data sets that might reliably represent the population at risk of *theft from the person* and marry them to a shared geography;
2. Calculate correlations between the population measures and crime volumes to identify the strongest crime predictor;
3. Use a Local Indicator of Spatial Association (LISA) statistic to identify areas with statistically significant crime rates, contrasting the clusters that emerge under the residential and ambient population denominators.

The paper has been structured as follows: the following section (Background) outlines the relevant literature that this project will draw on; the Data and Methods section discusses the data and the methods used; the Results section presents the results; and the remaining two sections complete the paper with a discussion and conclusions respectively.

http://dx.doi.org/10.1016/j.jcrimjus.2016.03.002
0047-2352/

Background

It is well known that crime concentrates in certain places (Sherman et al., 1989; Andresen & Malleson, 2011; Weisburd et al., 2012; Weisburd & Amram, 2014) and at certain times (Felson & Newton, 2015). A body of criminological research has evolved around the impacts that characteristics of the residential population and the surrounding physical environment will have on those concentrations. However, it is increasingly evident that "daily nonresidential activities distribute crime unevenly over space, beyond residential effects" (Felson & Boivin, 2015, p 1). Boggs (1965) recognised, some 50 years ago, that rates of some types of crimes were dependent on the ambient population within an area, rather than the residential population. But criminological research into the use of alternative population denominators has been relatively limited until recently, primarily due to a lack of appropriate and available data. Nevertheless, there is some previous research in this area. A number of studies have made use of the LandScan[1] population database (Andresen, 2006; Andresen & Jenion, 2010; Andresen, 2011; Kautt et al., 2011; Andresen et al., 2012; Kurland et al., 2014). These studies are consistent in their support for the use of populations other than the residential. LandScan uses satellite imagery to estimate size of the ambient population at a resolution of 1 km^2. Although these data have been shown to offer some advantages over residential measures for certain types of crime, they are limited by their temporal and spatial resolution. They provide an average estimate of the ambient population, so populations cannot be disaggregated into day, evening, weekend, and so on. Furthermore, it is well known that crimes cluster at micro-places (Sherman et al., 1989; Andresen & Malleson, 2011; Weisburd et al., 2012; Weisburd & Amram, 2014), so a 1 km^2 resolution will smooth heterogeneity at smaller geographies.

More recently, Felson & Boivin (2015) use a large travel survey to explore the relationship between visitors to an area and crime occurrences. The authors find strong correlations between the number of visitors in an area and both property and violent crimes. Similarly, Stults & Hasbrouck (2015) compare rates of crime using two different denominators: the residential population and the population who commute to the city. The authors find evidence that when the size of the commuting population is controlled for, the crime rates in some cities vary dramatically.

Although relatively sparse, the available literature on the ambient population points to a growing recognition that hotspots for some crime types are, to an extent, a manifestation of the flows of the population more generally. This is not, however, a recent observation. Routine Activities Theory (Cohen & Felson, 1979) and the Geometry of Crime (Brantingham & Brantingham, 1981) associate structural differences in everyday, routine activities with variations in crime rates. Individuals structure their day by travelling between important *nodes* (i.e. places of work/school, friends' houses, etc.) along commonly-used *paths* (Brantingham & Brantingham, 1993). Inevitably, crime tends to follow these activities, and clusters around the nodes and paths that drive them through space. As Kinney et al. (2008) put it:

"The general patterns of movement towards and away from activity nodes such as work or school locations, major shopping areas, entertainment districts or bedroom suburbs provide a very general image of where crimes will concentrate".

A body of literature has also extended these ideas in order to classify buildings and land use types in terms of their impact on crime rates. The most important of these, in the context of the crime of *theft from the person*, are crime *generators* (Brantingham & Brantingham, 1995). These are places that attract large numbers of people, such as shopping centres, and transport hubs, and can facilitate opportunistic crimes due in part to the large volume of potential victims (Kinney et al., 2008). The size of the populations in these places, by definition, will not be accurately captured by counting the number of residents.

From a theoretical perspective therefore, it is well known that variations in the behaviour of potential victims will influence crime locations. It is surprising that there is not more research on the ambient population. A likely explanation is that, unlike residential population counts and land use data that are readily available, measures of population flows are more difficult to find (Felson & Boivin, 2015). Also, *opportunity* is a complex phenomenon that does not lend itself well to "simple counts of available targets" (Clarke, 1984). For example, whether an opportunity is acted upon depends on the perceived ease and attractiveness of the opportunity (Clarke & Cornish, 1985). To compound matters, these are subjective traits that can have widely different impacts on different individuals – e.g. one opportunity might both dissuade a prolific offender and at the same time encourage someone who is otherwise law abiding (Clarke, 1984). Whilst residential population data are usually rich in socio-demographic attributes, ambient population counts are hard to come by, let alone rich enough to begin to unpack some of the complexities highlighted by Ronald Clarke.

Although this research does not begin to unpack opportunity, it contributes to contemporary literature on the ambient population on in two ways. Firstly, it experiments with new data sources that have rarely been applied to spatio-temporal crime analysis, let alone to explore the impacts of the ambient population specifically. Secondly, it is novel in its treatment of geography. With the exception of Felson & Boivin (2015) who use the Census Tract geography, previous studies have generally focused on relatively large areas. Crime clusters at micro-places (Sherman et al., 1989; Andresen & Malleson, 2011; Weisburd et al., 2012; Weisburd & Amram, 2014) and, therefore, analysis at larger geographies can hide important lower-level patterns. Whilst broader analysis is valuable, it limits insight into the relationship between crime and human activities more generally. It will also have limited *direct* policy impact (that is not to say that the findings will not influence policy, only that they cannot help to better tailor interventions at specific neighbourhoods).

To summarise, although a body of work is *beginning* to emerge around the impact of different denominators on crime rate calculations, no prior work has directly compared the impacts that a number of ambient population measures will have on city-wide crime rates. Stults & Hasbrouck (2015) summarise the problem thus:

"Many cities, and particularly large commercial centers, attract large numbers of persons during the day who return to their homes outside of the city at night. The routine activity of daily commuting to work puts this group at higher risk of crime while within the city boundaries, and any crimes they commit or that are committed against them are included in reported levels of crime for the city. However, this population is not included in the denominator when calculating the city's crime rate since the traditional method of calculation includes only the residential population."

A failure to apply the most appropriate population denominator in a crime rate calculation might lead to the identification of spurious hotspots. This will have an inevitable impact; crime reduction policies aimed at the areas with the most significant crime problems might be poorly targeted and mis-specified.

Methods and data

Data

Table 1 outlines the data sources used in this study with descriptive statistics. Each source will be described separately in the following sections. It is worth noting that the input data have very different absolute numbers, which has implications for crime rate calculations and hotspot analysis. Although there would be some utility in combining different population-at-risk measures to improve the subsequent measure's overall accuracy, this needs to be done with great care to mitigate against potential biasses introduced by these different absolute numbers. We address this risk here by treating each population-at-risk measure separately and ignoring the absolute crime rates that they

Table 1
The data sources used here, their geography, and the corresponding variable names used in the remainder of the paper.

Data source	Geography	Variable name	N	Mean	St. dev.	Min	Max
Census residential population	LSOA	*Residential*	8,173,941	1691	264	985	4933
Census workday population	LSOA	*Workday*	4,500,481	931	4216	80	220
Geo-located twitter messages	Points	*Tweets*	204,159	42	86	0	2209
Mobile telephone activity counts	Regular grid	*Mobile*	10,037,250	2076	6628	0	347,303
Population 24/7 population estimates	Regular grid	*Pop247*	7,450,617	1541	3290	0	164,678
Theft from the person offences	Points	*Crime*	3,558	0.736	3.279	0	89

produce; focussing instead on the *difference* in crime rates between areas. This approach also has the advantage of closely reflecting the general definition of a crime 'hotspot' – where local area averages are significantly greater than global averages (Chainey & Ratcliffe, 2005).

To illustrate the differences in the spatial structure of populations measured by the different data sets, Fig. 1 maps all of the input data. Immediately obvious from Fig. 1 is that the residential population is far more evenly distributed than the census workday population, Twitter messages, aggregate mobile telephone activity, and population 24/7 daytime estimates. This should come as no surprise given that work-related land use is more concentrated than residential-related land use, but points to the importance of identifying an appropriate population at risk measure.

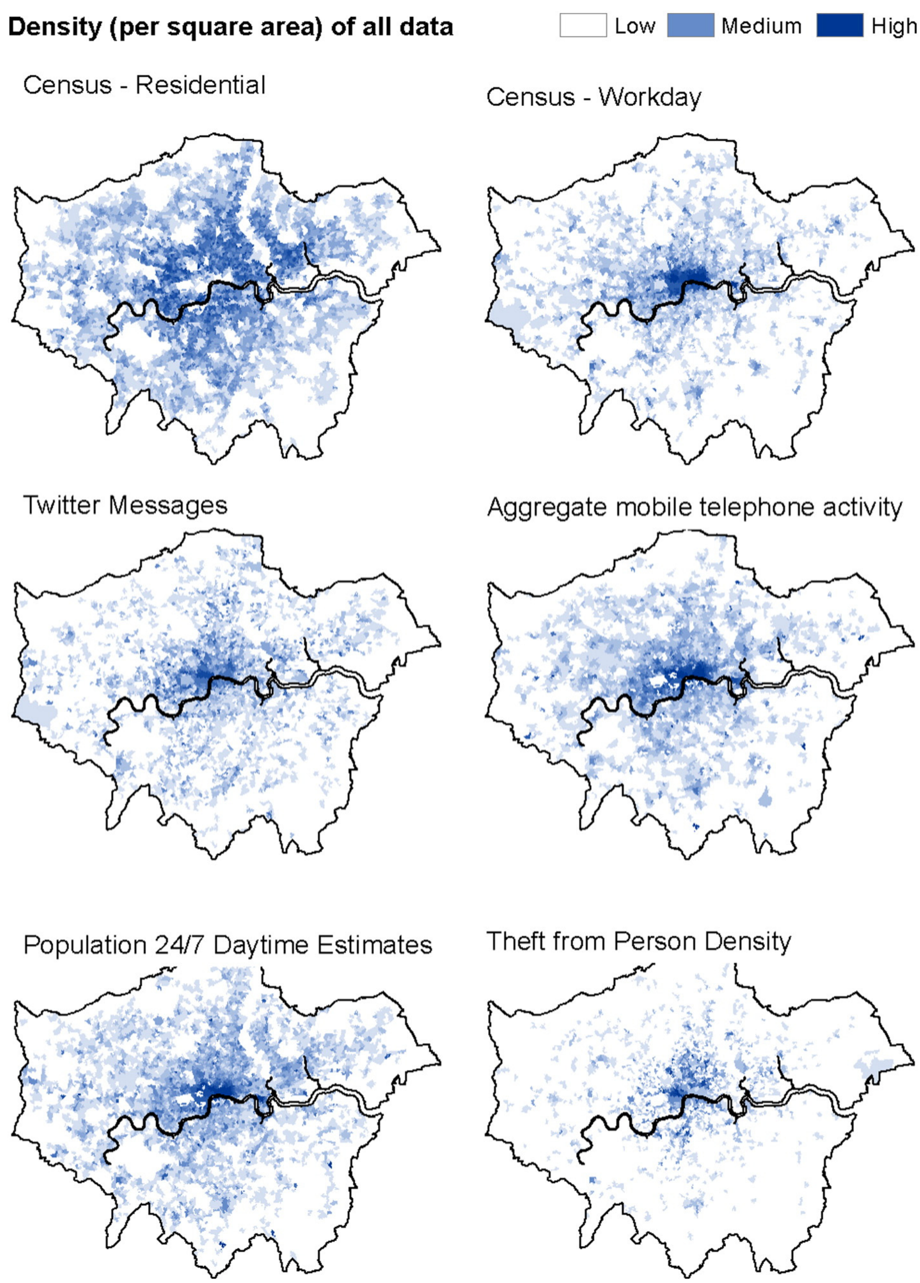

Fig. 1. Maps of the input data at the Lower Super Output Area (LSOA) geography. To reduce the visual impact of variations in area sizes, the data are displayed as densities (counts divided by the square area of the constituent LSOA).

Time period

The time-span of the study period must be defined. Ideally, the datasets would be disaggregated to reflect the temporal dynamism of crime patterns and the underlying ambient population, e.g. by daytime, evening, and weekend. Unfortunately, however, the available crime data have been aggregated to a single month which significantly limits the benefits of breaking down the population-at-risk. Such a fine temporal analysis will be left for future work when more temporally accurate crime data are available.

For this study, May 2013 was chosen. This is the most suitable time period because: it the same year and month that the aggregate mobile phone data are available for; it is the same year that the Twitter data are available for (September 2013); and it is relatively close to the 2011 Census day (31st March 2011). It also has the advantage that it avoids major holiday periods such as Christmas and the summer school holidays that will distort regular activity patterns.

Crime data

The crime data used here are all *theft from the person* offences. These are offences in which "there is theft of property, while the property is being carried by, or on the person of, the victim" (Office for National Statistics, 2014b). This crime type is particularly instructive for this study because *theft from the person* requires the presence of a victim (unlike residential burglary) that is best analysed through an understanding of *where* people actually are, not where they sleep. Data on recorded offences were obtained from the three organisations that hold police crime data in London – the Metropolitan Police, the City of London, and the British Transport Police – for May 2013. The crime data have been made available via the police.uk service and have undergone some spatial anonymisation. Although this reduces their accuracy, the impacts on this research are limited. Tompson et al. (2015) found that at the Lower Super Output Area (LSOA) geography 85% of all areas exhibited no statistically significant difference between the anonymised and raw crime data. The LSOA geography is used for the statistical analysis.

Residential census data

Crime rates calculated using measures of the ambient population are compared to those produced under the residential population. Hence the 'usual resident population' in each LSOA was obtained from the 2011 UK Census.

Workday census data

An innovation in the 2011 UK Census was to include a new measure of the population: the *workday* population. This is defined as the number of people who work in an area, plus those who are resident in the area and either do not work (including those under 16) or work from home. This includes all full- and part-time employment, including regular '9–5' work, night work, irregular shifts, etc. For full details see Office for National Statistics (2013). A new geography was created with which to publish workday population data called the 'workplace zone' (WZ), but for consistency with the other measures used here the workday data were obtained using the LSOA geography.

Population 24/7 data

Although workday census data potentially provide a closer approximation to the ambient population than residential data, they are also limited in that they do not include activities other than employment and cannot be temporally disaggregated to account for atypical work patterns. In an effort to improve the accuracy with which the ambient population can be measured, the Population 24/7 project (Smith et al., 2015; Martin et al., 2015) attempts to redistribute the census population across a regular spatial grid according to likely activities of residents. For example, students will be redistributed to schools during school hours, visitors can be distributed to attractions (museums etc.) during their opening times, and people who are likely to visit hospitals during the day can be redistributed there accordingly. For full details, see Martin (2011).

For this research, the Population 24/7 project kindly provided an estimate of the ambient population at 14:00 on a weekday, distributed over a 200 m^2 regular grid. The data take account of home, work, education, health care and some large visitor attractions.

Aggregate mobile phone counts

An estimate of aggregate mobile phone activity was obtained from Telefónica Digital; a large telecommunications provider. These data are similar to those used in other criminological research such as Bogomolov et al., (2014). The data set contains hourly counts of mobile telephone activity for the period 13–19 May 2013, geo-located to 3,891,216 cells that span the spatial extent of London. 'Activity' is loosely defined; rather than representing the number of mobile devices in an area it actually represents the number of times that a device communicates with the network. These events might be user-driven (i.e. by making a call) or generated automatically by the device as it communicates with the provider. Hence the measure is only a proxy for the size of the population in a cell, but correlates reasonably well with other measures of the ambient population (Fig. 3, discussed later, illustrates this). As the data are aggregate counts they are entirely anonymous. No information about the activity of an individual device is included, nor could it be elucidated from the aggregate data.

It is possible to extract the population for any hour in the data set and use this as an estimate for the ambient population. Here, the weekly daytime population, defined as 14, 15, 16 May (Monday, Tuesday, Wednesday) between 1–3 pm – is extracted. Another time period could have been chosen, such as 16:00–19:59 which is a 'hot time' for theft from the person (Smith et al., 2006; Newton et al., 2014), but 1–3 pm during the week has the closest correspondence with the *workday* and *pop247* data.

Social media data

The final data set used here is a collection of geo-located messages posted to the Twitter social media service. Approximately 200,000 messages with GPS coordinates originating in the study area were collected between 19th and 23rd September 2013 using the Twitter Streaming API. Ideally, the project would have a choice over the exact time period over which to collect the data, but at the time of writing no historical lookup service was available and the researchers were therefore limited to the data that they had collected historically. There are inevitable questions over the reliability of these data as a measure of the ambient population, but these are beyond the scope of this paper. The interested reader might refer to Malleson & Andresen (2015a,b).

Points of interest

One other aspect that will be interesting to explore, although it does not feature as part of the core analysis, is how the different types of amenities in an area influence its crime rate. In particular, this research will attempt to identify land uses that are "characterized by high activity functions" (Kinney et al., 2008 p72); i.e. crime generators. These land uses have been found to be places where assaults cluster (Kinney et al., 2008). A full analysis of this relationship is beyond the scope of this paper – see Brantingham & Brantingham (1995) for example – but it is interesting to explore whether there is a relationship between the residential population, the ambient population, crime rates, and amenity types.

To this end, Open Street Map (OSM) building data were obtained from the Geofabrik[2] service for the London area. The dataset contains 260,786 separate objects. In OSM, building types are classified using tags and these tags can be used to estimate whether a building might be crime generator. OSM does not have an official schema for choosing an appropriate tag for an object. Hence there is great flexibility in naming objects, but similar objects might be given different tags by different users. Fortunately the community is converging on a set of common tags, so of the 368 unique tags in the London data set, the most commonly-used 2% of tags are used to label 31% (80,848) of the objects. These tags can therefore be used to assess how many different types of amenity are present in the different crime clusters.

Geographical analysis

In order to statistically analyse the input data, each source must be married to an equivalent geography. Here the Lower Super Output Area (LSOA) geography is used because it strikes a balance between being small enough to limit the extent to which lower-level heterogeneity is obscured but is large enough to capture wider patterns that might not be in evidence at a lower resolution. LSOAs have been designed by the Office for National Statistics to contain between 1000–3000 people and 400–1200 households. There are 4835 LSOAs in the study area.

The data derived from the 2011 UK census – residential population and workday population – have been released at the LSOA geography so no spatial manipulation is necessary. The crime data and Twitter data are released as points, so can simply be aggregated to the LSOA areas. The remaining data – mobile phone activity and Population 24/7 outputs – are released using regular grids, so must be converted to the LSOA geography. This is achieved by intersecting the grid cells with the LSOA geography, splitting the proportions equally across divided cells, and then aggregating cells to their parent LSOAs. A similar conversion for the LandScan database was also necessary (Andresen & Jenion, 2010). This has been implemented by adapting a technical procedure for R outlined in Brunson & Comber (2015).

Hotspots and clustering

'Hot spots' are locations that have a disproportionately large number of occurrences compared to their surroundings over a particular time period. Although there was already a strong traditional of exploring spatial patterns of crime, interest in crime hotspots became particularly acute in the 1970s and '80s as detailed, electronic, high-resolution, spatial data sources such as censuses and police records began to allow more nuanced research into the geography of crime (e.g. Sherman et al., 1989).

There are a variety of methods that can be used to estimate hotspot locations. For a review, the interested reader is directed to Chainey & Ratcliffe (2005). Here, the Getis-Ord G_i^*(Getis & Ord, 1992; Ord & Getis, 1995) statistic has been chosen. G_i^* operates by examining each location i (LSOAs in this case) together with its neighbouring locations j. It calculates whether or not the total number (or rate) of occurrences in i and j is greater or lesser than would be expected by chance when compared to surrounding locations up to a distance d away from i. If a difference is found, then the areas i and j are assumed to be associated and different to their surroundings, i.e. a hotspot or coldspot. A benefit to the procedure over some other hotspot methods is its ability to provide a Z score and p-value for each location. These indicate whether the null hypothesis (that there is no association between the crime level in areas i and j) can be rejected. It is therefore possible to identify *statistically significant* hotspots.

The statistic was executed with the ArcMap GIS software using an inverse distance weighting scheme such that the impact of neighbouring areas on the area under study is reduced with distance. ArcMap estimated an appropriate threshold distance of 2.32 km and this was ultimately used (experiments with other thresholds produced little difference in the overall results).

Results

Correlation and spatial models

Frequency distributions for the explanatory variables (provided in Fig. 2) illustrate that none of the variables are normally distributed. Therefore Spearman's rank correlation coefficient (ρ) statistic is used over Pearson's product-moment coefficient (r) to calculate correlations.

Fig. 3 illustrates the strength of the relationships between the constituent variables at the LSOA level. Unsurprisingly, the residential population (*residential*) is poorly correlated with the ambient population measures. This is to be expected as the data inherently measure different populations. However, it is interesting to note that there are some particularly strong correlations between some of the measures of the ambient population. In particular, *workday* (the census estimate of the size of the population during the day), *mobile* (the aggregate mobile telephone data) and *pop247* (the Population 24/7 daytime estimates) all exhibit strong correlations. This suggests that each variable is capturing a similar population. Also interestingly, the social media data (*tweets*) are comparatively poorly correlated to the other ambient measures: $\rho = 0.35$ (*workday*); $\rho = 0.23$ (*mobile*); $\rho = 0.26$ (*pop247*). Again this is not entirely surprising as the social media data have not been temporally disaggregated, whereas the other data sets specifically represent daytime weekday populations.

These correlations between the explanatory variables are interesting in themselves, but exploring the correlations between different data in detail is beyond the scope of this paper. Instead, the remainder of the discussion will focus on the correlation between the population estimates and crime rates. Table 2 summarises the correlations between crime and the population measures at the LSOA and OA (Output Area) geographies. OAs are a much smaller geography than the LSOA. The relative strengths of the relationships do not change with geography, suggesting that the results are not an artefact of the modifiable areal unit problem (Openshaw, 1984).

There are clearly some outliers in the data; particularly for the *pop247*, *mobile*, and *tweets* data, although these are not surprising and will not be discarded. Census populations are generally set up for statistical inference so exhibit little variation in the resident populations, especially in an urban area. However, when these areas are used to quantify the ambient population, some very highly populated areas will naturally emerge, particularly in city centres. Although these are outliers from a statistical perspective, they actually represent reality.

It is also worth pointing out that some areas have zero values, either for the number of crimes or for the population data, and these will have some influence on the correlations. These zeros are, however, not surprising. In terms of crime counts, even if a longer time period were used, we would still expect a large number of areas with no crime. For example, Andresen & Malleson (2011) found that with robbery (a comparable crime type), 37% of all dissemination areas in their study showed no crime in a three year period. In terms of the population data, it is natural that there will be many places that have no employed people, tweets, etc.

The most important observation is that all of the variables are positively correlated and are significant at the 1% level. None of the correlations are overly strong, but this is to be expected as victim abundance is only one aspect of an otherwise extremely complex process that includes the offender and guardian locations (Cohen & Felson, 1979), a complex environmental mosaic (Brantingham & Brantingham, 1981), and offender decision making (Clarke & Cornish, 1985). However, by identifying the population measure that is most closely related to the crime patterns of interest, we are in a stronger position to begin to quantify crime hotspots that actually consider the risk of victimisation, not simply the volume of criminal events. In this case, the most strongly correlated variable is *workday* (the workday

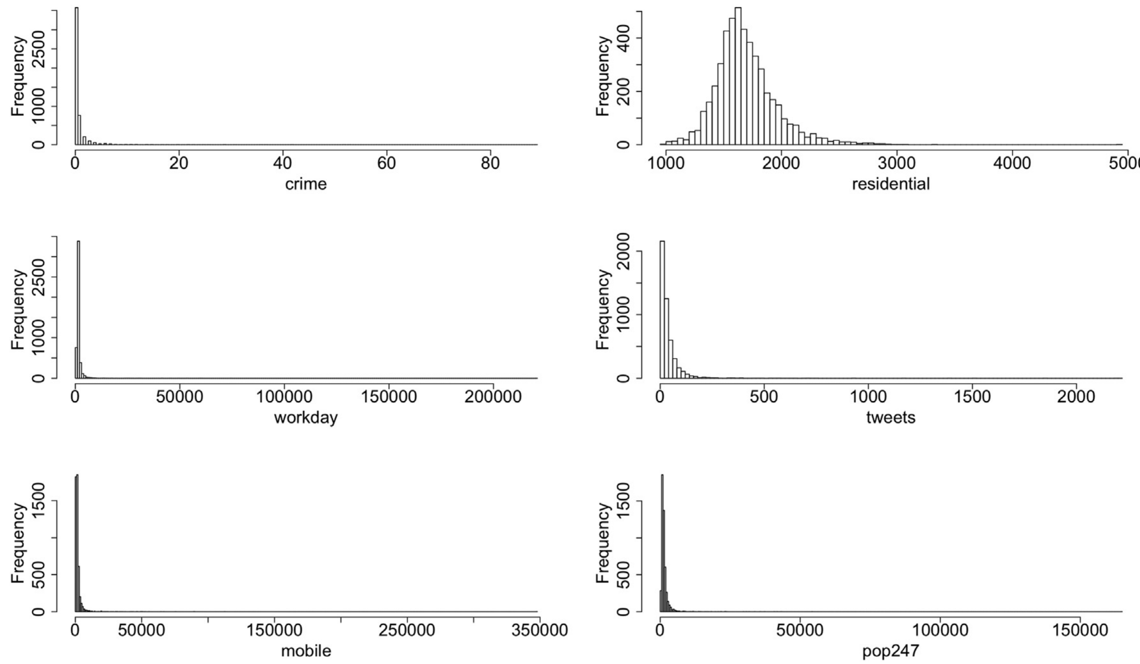

Fig. 2. Frequency distributions for all of the variables.

population estimate from the 2011 census). The following section will explore the impact that changing the denominator has on hotspot calculations.

Shifting hotspots

To quantify the impacts of different population denominators, the crime rate was calculated for each Lower Super Output Area using two different denominators: the *residential* population and the *workday* population. *Workday* has been chosen over the other possible population-at-risk measures because it was the most strongly correlated with the crime data. It would be interesting explore the crime rates that arise under the other measures as well, but this is well beyond the scope of a single paper. It is important to note that similar results were obtained at the smaller Output Area geography so the later results are unlikely to be wholly an artefact of the modifiable areal unit problem (Openshaw, 1984). The G_i^* statistic (Getis & Ord, 1992) was used to identify clusters that have significantly high or low crime rates given the rates in surrounding neighbourhoods.

Fig. 4 illustrates the significantly high crime rates at the 95% level ($p<0.05$) using both denominators. A large area in the centre of the city exhibits a significant crime rate using the *residential* population, but this disappears when the *workday* population is used. This is not surprising as the city centre has a low residential population and a high ambient population. This distribution of clusters has been found in similar research in different contexts (Andresen & Jenion, 2010; Andresen, 2011; Malleson & Andresen, 2015a,b). However, moving beyond the city centre the results become more interesting. This will be explored further in the Discussion.

Crime generators

A drawback with the analysis, that will be elaborated on in the Discussion, is that the crime data used here have been temporally aggregated to a single month. Hence it is not possible to separate daytime and night-time crimes. The measures of the ambient population used here predominantly estimate the *daytime* population, but without information about the times of offences it is equally likely that the *evening* population – neither residential, nor ambient daytime – is the most suitable denominator. In the absence of measures of the evening population, we turn instead to the physical environment as a means of estimating the presence of daytime versus evening populations. The flow of the ambient population will be driven in part by the location of *crime generators*. These are high-activity places that attract large numbers of visitors and can inadvertently help to facilitate crime opportunities.

Open Street Map (OSM) data can be used to estimate whether particular buildings are crime generators. The full set of individual OSM buildings has been analysed to derive a subset of non-residential, 'high activity' land uses that are likely to influence rates of assault. Kinney et al. (2008) perform an extensive analysis and identify a number of land uses that exhibit heavy concentrations of assaults. Some of these have been disregarded as it is not possible to reliably map them onto the tag system used by open street map contributors but most can be directly translated. The following lists the most important land uses as identified by Kinney et al. (2008) (in bold) and how these are mapped on to OSM tags (in italics):

- **Shopping Centre**: *retail, commercial, post_office, supermarket* and *shops*
- **Neighbourhood Pub**: *pub, bar* and *nightclub*
- **Fast Food Restaurant**: *fast_food* and *restaurant*
- **Motel & Auto Court**: *parking, garage* and *garages*
- **SkyTrain**: *station, transportation* and *taxi*
- **Schools, Universities and Colleges**: *school, university* and *college*
- **Recreational and Cultural Buildings**: *place_of_worship, church, hospital, community_centre, library, public_building, theatre, train_station, castle, museum, attraction, arts_centre, courthouse, cinema, social_facility, townhall, gallery, civic* and *swimming_pool*

To analyse the abundance of difference types of amenity in each crime cluster – workers, residential, and intersecting (i.e. where residential and workers clusters overlap) – a 500 m buffer is first drawn around each cluster. This acknowledges the fact that a nearby amenity might still have an impact on an area without immediately intersecting it. Fig. 5 and Table 3 then illustrate the number and

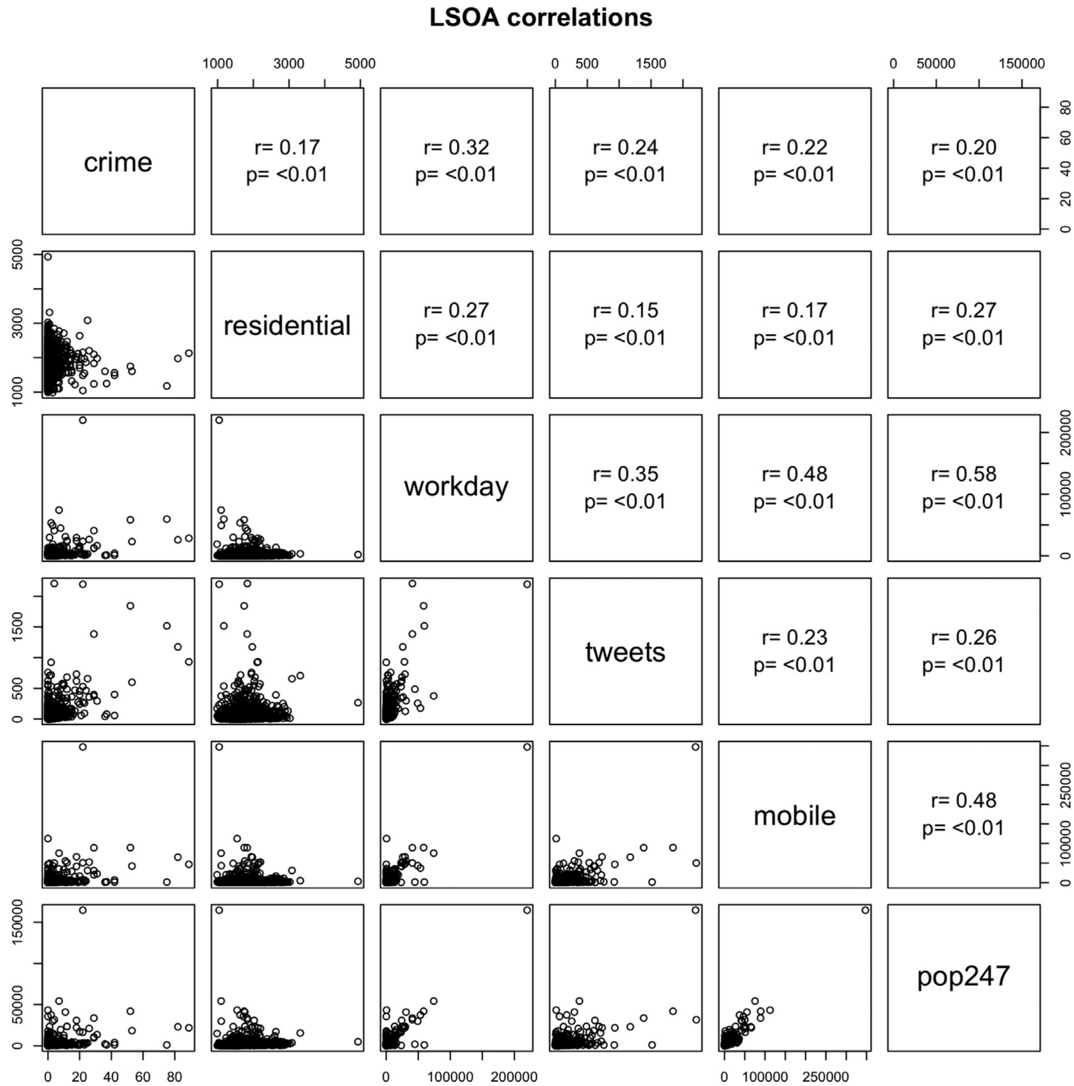

Fig. 3. Spearman's rank correlations and p values for all data at the LSOA geography.

percentage of different amenities that intersect the different crime cluster types. The results will be discussed in the following sections.

Discussion

Spatial patterns

Fig. 4 illustrates the locations of three different types of crime clusters that are significant at the 95% level: those calculated using the residential population as a denominator; those using the size of the working population (as a proxy for the ambient population); and those areas where the residential and ambient clusters intersect. There are two particularly notable features to the locations of the resulting clusters.

Table 2
Spearman's rank correlation coefficient (ρ) for the explanatory variables correlated against the crime data at the OA and LSOA geography. All are significant at the 1% level.

Explanatory variable	OA correlation	LSOA correlation
Workday	0.2	0.32
Tweets	0.17	0.24
Mobile	0.15	0.22
Pop247	0.15	0.2
Residential	0.025	0.17

Firstly, there are clusters surrounding the city centre under *both* the ambient and residential crime rates. These are particularly prevalent in an area to the north of the city centre as illustrated by the lower inset in Fig. 4. The environmental backcloth in this region is such that the area is predominantly residential, but also provides a range of amenities for locals and visitors. These will act as crime generators by attracting large numbers of people, and may also act as crime attractors by specifically encouraging illegal behaviour. In fact, areas that exhibit a similarity between ambient and residential crime hotspots might be indicative of a mix of both crime generators and attractors – i.e. places that attract large numbers of people for non-crime purposes (generators) as well as places that are used specifically for criminal activity (attractors).

On the other hand, those areas that are largely residential and yet only exhibit hotspots under the ambient population might be places with a higher proportion of crime attractors to stimulate crime, but fewer generators to attract the large volumes of people. This leads to the second interesting finding. Fig. 6 highlights some areas to the west of the city centre that exhibit a significantly high crime rate under the ambient population, but not under the residential population. Hence these areas have a larger volume of crime than would be expected given the size of the population-at-risk (the ambient population). Although they have a raised volume of crime compared to their surrounding neighbourhoods (see the lower inset in Fig. 6) these volumes are relatively low compared to broader volumes across the city. If the residential population were used as a

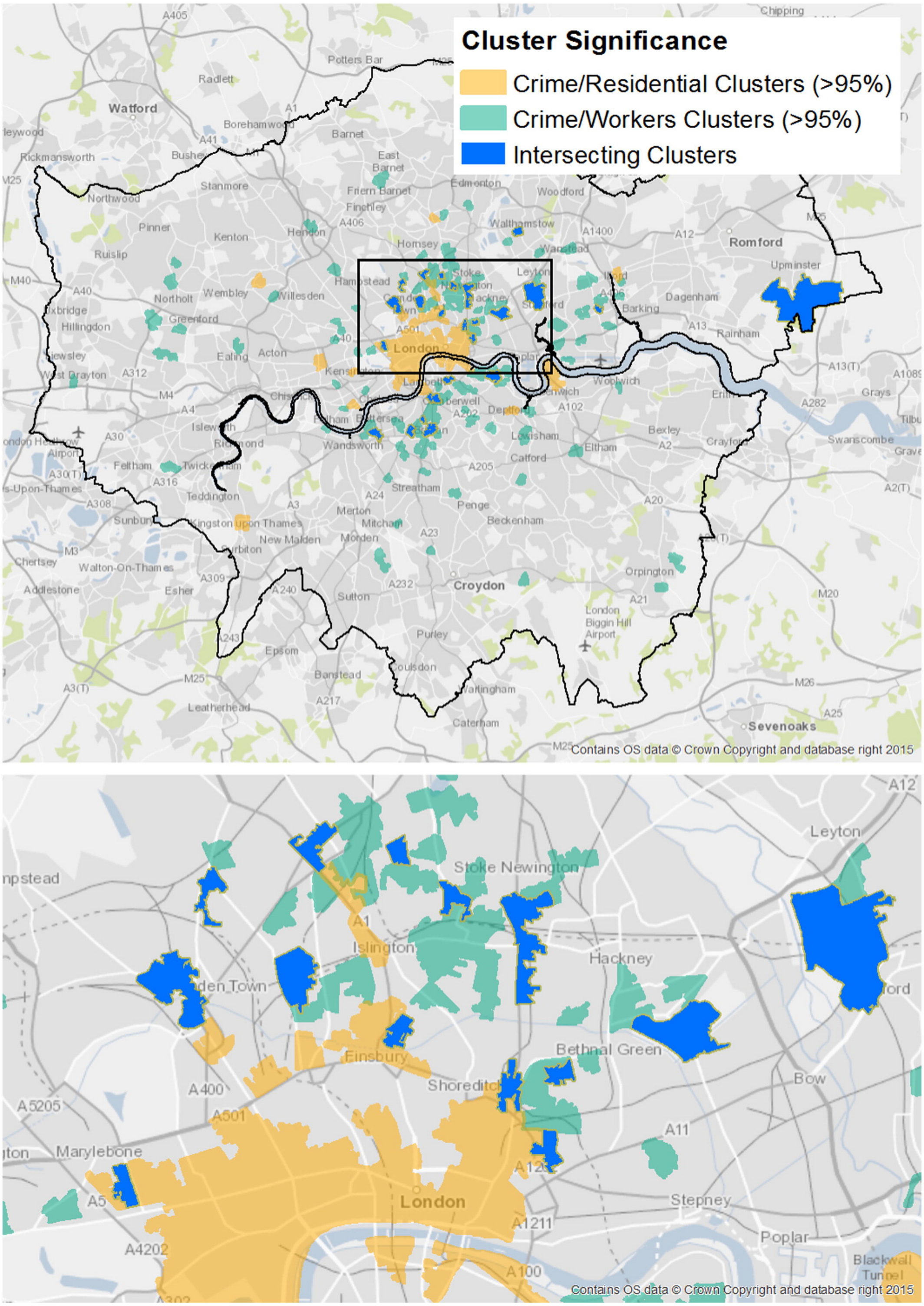

Fig. 4. Significant G_i^* clusters of high crime rate using residential population and workday population as denominators throughout the study area. Also illustrates intersecting clusters, focusing on part of North London (lower inset)

denominator, these areas would not have been identified as hotspots. It is worth nothing that there are other areas in the city that exhibit this pattern; these clusters are only used as an example. The following sections will discuss the implications of this result in more detail.

Daytime or evening ambient population?

As evidenced in Table 2, there is a much stronger association between reported thefts and the size of the ambient population than there is between thefts and the size of the residential population. However, in this work the ambient population has been measured by the size of the *daytime* population and it is conceivable that in these areas there is also a high *evening* ambient population. Therefore although the daytime population is a superior population-at-risk measure to the residential population, there is the possibility of misattributing this association to the daytime population when it is actually the evening population that are being victimised. This is an important distinction and therefore the discussion continues on two fronts: (1) with an

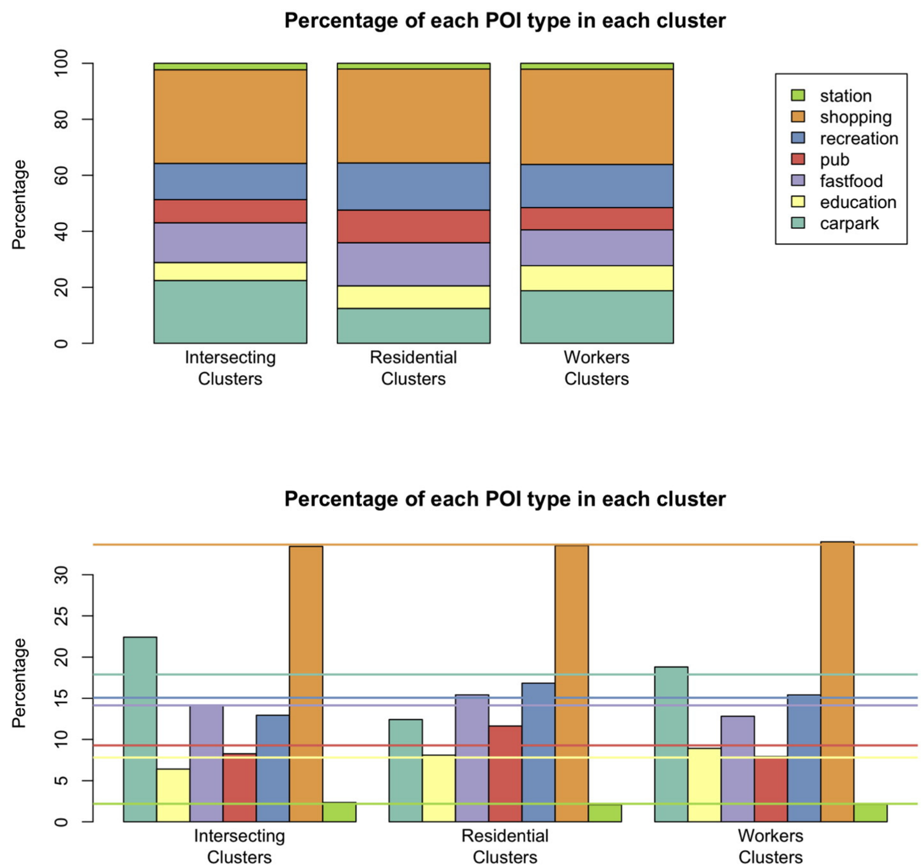

Fig. 5. The percentage of different building types in areas that are classified as clusters under the residential population, ambient population, and where those two clusters intersect. Horizontal lines show the mean percentage of each POI across the three cluster types.

analysis of the physical characteristics of different cluster areas to search for quantifiable differences in the types of amenities; and (2) an overview of the potential for disaggregating the victim population temporally and by socio-economic characteristics.

Crime generators

Although the ambient population appears to be a more suitable denominator than the residential population, it naturally does not fully explain the variation in crime rates across the city. Part of the explanation almost certainly lies in the environmental backcloth (Brantingham & Brantingham, 1993). We have already suggested that the crime hotspots calculated using the ambient population (e.g. those in the west of the city in Fig. 6) are unlikely to contain many crime generators (such as underground stations, music venues, pubs, and schools) as the presence of these facilities would have increased the size of the ambient population.

Table 3
The number (N) and percentage (%) of different point of interest types that intersect the different crime cluster types.

2*Amenity type	Intersecting clusters		Ambient clusters		Residential clusters	
	N	%	N	%	N	%
Carpark	556	22.42	856	18.79	623	12.42
Education	159	6.411	406	8.913	406	8.096
Fastfood	352	14.19	584	12.82	773	15.41
Pub	205	8.266	361	7.925	583	11.63
Recreation	321	12.94	702	15.41	844	16.83
Shopping	829	33.43	1549	34.01	1682	33.54
Station	58	2.339	97	2.13	104	2.074

To explore this assertion quantitatively, the number of different building types, as determined from their Open Street Map tag, in each cluster area was counted (see Table 3 and Fig. 5). Although the differences are not large, areas classified as a hotspot under the ambient population appear to have a smaller proportion of restaurants, recreation buildings, and pubs (the largest difference). They have slightly more educational establishments, shopping centres, public transport stops (although these differences are only marginal), and many more car parks. These results, therefore, go some way to supporting our assertion that these areas do indeed have fewer crime generators in them, particularly with respect to the substantially smaller proportion of pubs. If there are few generators in the area, then there might be crime *attractors* (i.e. facilities/activities that exist primarily to facilitate crime) that attract people who might also commit theft. Further research is required to explore this in more detail, but using crime rates in this way would certainly be an interesting way to hypothesise about the presence of crime attractors that would otherwise be extremely difficult to locate. By their nature, crime attractors will not be present in data such as Open Street Map.

The analysis of building types also begins to provide some evidence for the type of ambient population that might be present: the slightly larger proportion of schools and shopping centres in particular suggest a larger daytime population than an evening population. Whilst further research is needed to be more confident in this, this preliminary analysis does not cause us to rule out the use of data measuring daytime as opposed to nighttime populations. In the least, we can be more sure that the number of workers in an area, as a proxy for the ambient population, offers a substantial improvement to simply using the residential population, and the daytime vs. evening ambient population is not of critical importance at this stage.

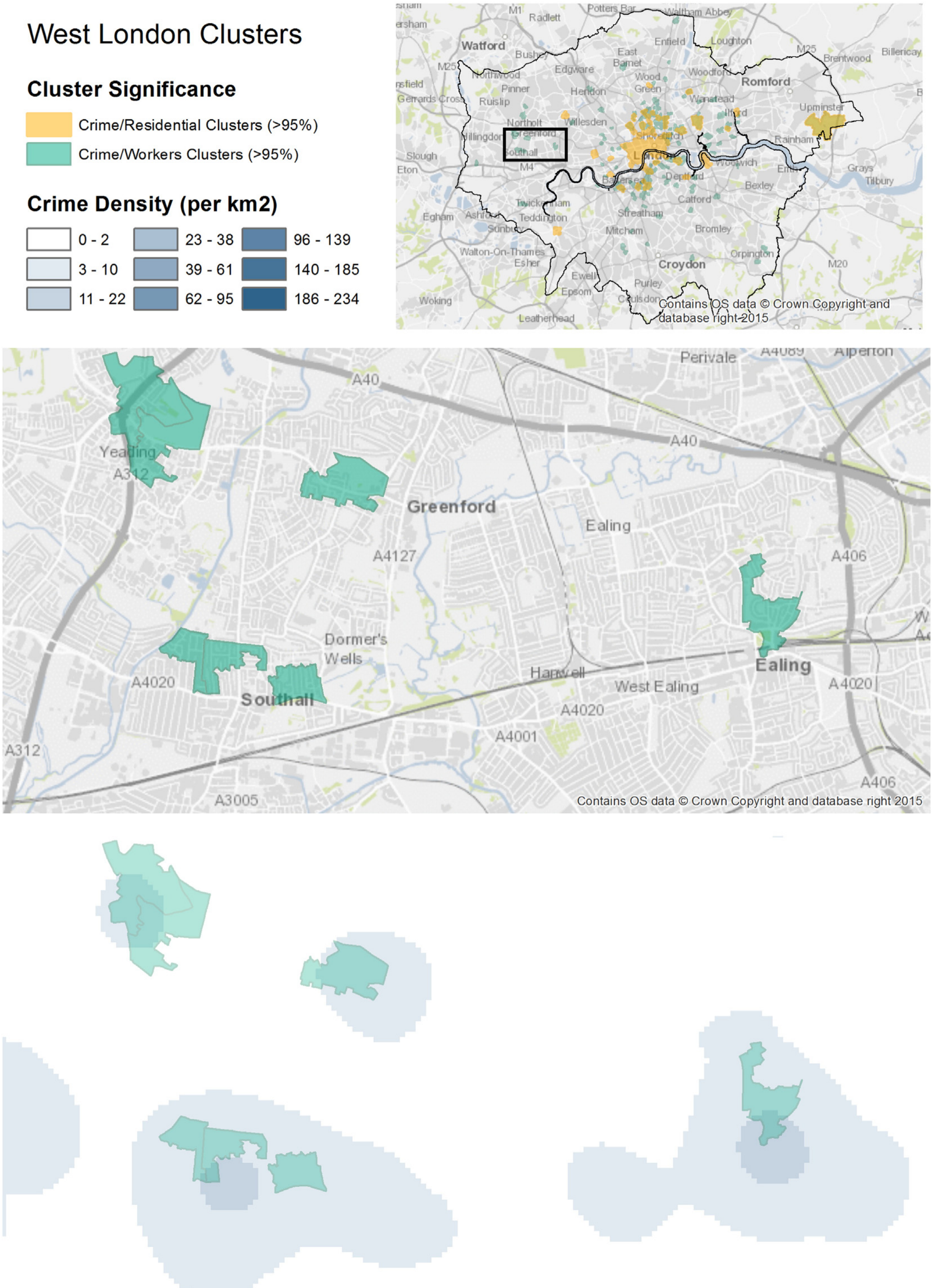

Fig. 6. Significant G_i^* clusters of high crime rate using residential population and workday population as denominators focussing on some specific clusters to the west of the city centre.

Victim disaggregation

The socio-economic characteristics of the areas are another important consideration as the clusters might highlight places with large numbers of people who are particularly vulnerable to theft from the person. For example, evidence suggests that women are twice as likely to be victims as men (Office for National Statistics, 2014a) and particularly those in the 18 and 24 age group (Home Office, 2013). One of the advantages with the approach put forward in this paper, is that some of the data sets used have the capacity to provide more accurate information about the characteristics of the underlying ambient population. This helps to move away from the "simple counts" that (Clarke, 1984) is wary of. For example:

- The 2011 Census statistical release includes a number of cross-tabulations for the *workday* population that allow sub-divisions by characteristics such as age, gender, level of education, etc.;
- Some characteristics of Twitter users can be imputed (Adnan et al., 2014; Sloan et al., 2015) although unlike Census statistics the estimation will inevitably introduce error;

• *pop247* datasets are derived from the Census, so can be disaggregated by the same demographic categories;
• The *mobile* data include age and gender totals.

Therefore a next step for this avenue of research will be doing begin to disaggregate the ambient population by important demographic characteristics relevant to the type of crime under study. This will result in an even more appropriate population-at-risk measure and hence an even more accurate estimate of hot spot significance.

Caveats and limitations

Other than the temporal weaknesses already discussed, the crime data also suffer from spatial weaknesses. Each offence is aggregated to one of a finite number of *anonymous map points*[3] before being released to the public. This has the effect of confounding the data spatially. Tompson et al. (2015) compared the anonymised data to raw Police data and found that with the geography used here (lower super output area) approximately 15% of areas exhibited statistically significant differences between the anonymised and raw data.

The ambient population measures used here are probably a more appropriate population-at-risk measure than the residential population. However, they bring their own limitations. The Twitter data have serious issues around accuracy and bias that limit the extent to which they reliably reflect the true underlying population, rather than just the group of people who use Twitter (see, for example, Malleson & Andresen, 2015a,b). The mobile telecommunications data share similar limitations, although bias will be much less substantial. Rather the main limitations of those data arise from inaccuracies in the location estimates of individual devices. The technical process with which devices communicate with base stations (masts) is complicated (Heine, 1998; Pauli et al., 2010) and there is no guarantee that the device will communicate with the closest station. Also, where base stations are widely distributed there will be a large error in the estimate of the device location. No attempt at triangulating the phone location is made in the data available here.

The data derived from the 2011 census (*workday*, *pop247*, and *residential*) are the most reliable in terms of bias, although accuracy is also limited as the data have been aggregated spatially to reasonably large areas. These data are also limited with respect to the different types of activities that they can represent. Mobility in census data is limited entirely to journeys to work or school, whereas the other datasets have the potential to represent much more varied activities during both the daytime (e.g. buying lunch, visiting friends) and evening (e.g. going to the cinema and bars). It is important to note, however, that the census data are by far the most transparent.

Wider implications

This work has created more questions than it has answered. However, it has clearly shown that the population-at-risk matters in crime analysis. Although this is not a new discovery – Boggs, (1965) wrote about the importance of disaggregating the population-at-risk more than 50 years ago – it is useful to quantify the difference between the ambient and residential population distributions for a crime type that, by definition, is not dependent on the number of residents in an area. In addition, the work found that the census *workday* population is the most reliable ambient measure of those examined. This is a very encouraging finding – the data are robust and publicly available, so could be easily be taken up by others in criminological research.

The lack of temporal information associated with the publicly available crime data has been cited as a limitation that confounded the timings of offences and the presence of potential victims. Opportunity is more complex than the "simple counts of available targets" (Clarke, 1984 p 75), but without a more refined representation of the victims of crime, research is not able to move beyond simple counts. Also, if temporal information were to be included with the offence data it would not only be possible to begin to segregate crimes by the time of day but would also open up a new avenue of exploration around the competing influences of the physical environment and victim activities. We could begin to explore the spatio-temporal dynamism of crime hotspots and compare this to the dynamism inherent in the ambient population. Where hotspots are relatively stable over space-time and the ambient population is not, this suggests that the physical characteristics are more important than the flows of the victim population. Conversely, where hotspots move with the ambient population, the reverse might also be true. This type of knowledge would help to unpack the competing effects of the physical environment and victim activities.

The main finding has shown that using ambient population-at-risk measures highlights clusters of high crime that appear significant given the number of potential *ambient* victims in the area. This could provide Police forces with a framework on which to improve their estimates of hotspot significance, allowing them to target areas that have an unusually high rate, not just volume.

Conclusions and future work

This paper has tried to find the most appropriate population-at-risk measure for the crime of theft from the person. Whilst the measures of the ambient population perform better than the residential population (as expected) there is room for improvement. In particular, as there is no time of offence available in the crime data it is not possible to disaggregate the ambient population by daytime, evening, weekend etc. Undoubtedly the size, socio-economic characteristics, and behavioural factors of the population will vary considerably by these times. Nevertheless, the ambient measures still offer an improvement over the residential population for estimating hotspot significance. This is an interesting finding that has obvious implications for crime reduction by improving our understanding of where truly significant hotspots occur. It is particularly encouraging that the most reliable ambient measure that we found was the Census *workday* population. This is a robust and publicly available dataset that could easily be taken up by others. It is also encouraging that the ambient data sources used here offer opportunities for further victim disaggregation that will make a more nuanced assessment of 'opportunity' possible. For the crime studied here, the residential population results in an over-estimation of hotspot significance in areas with few residents but many potential victims. Finally, the paper discussed the relationship between cluster significance and the presence or absence of crime generators (places with lots of people) and attractors (places that attract people specifically to commit crime). A closer examination of the physical and social characteristics of the revealed clusters will no doubt be interesting.

Acknowledgements

The authors would like to thank Professor Dave Martin and his colleagues working on the Population 24/7 project who kindly provided the *pop247* data used here.

This work has been funded by the Economic and Social Research Council (ESRC) grant number ES/L009900/1

Notes

[1] LandScan website: http://web.ornl.gov/sci/landscan/
[2] http://download.geofabrik.de/
[3] Full details about the anonymisation process of publicly available UK crime data are available from https://data.police.uk/about/ and reviewed in detail by Tompson et al. (2015).

References

Adnan, M., Longley, P. A., & Khan, S. M. (2014). Social dynamics of Twitter usage in London, Paris, and New York City. *First Monday*, *19*.

Andresen, M., & Malleson, N. (2011). Testing the stability of crime patterns: Implications for theory and policy. *Journal of Research in Crime and Delinquency*, *48*, 58–82.
Andresen, M. A. (2006). Crime measures and the spatial analysis of criminal activity. *British Journal of Criminology*, *46*, 258–285.
Andresen, M. A. (2011). The ambient population and crime analysis. *The Professional Geographer*, *63*, 193–212.
Andresen, M. A., & Jenion, G. W. (2010). Ambient populations and the calculation of crime rates and risk. *Security Journal*, *23*, 114–133.
Andresen, M. A., Jenion, G. W., & Reid, A.,. A. (2012). An evaluation of ambient population estimates for use in crime analysis. *Crime Mapping: A Journal of Research and Practice*, *4*, 8–31.
Boggs, S. L. (1965). Urban crime patterns. *American Sociological Review*, *30*, 899–908.
Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. *ICMI '14 Proceedings of the 16th International Conference on Multimodal Interaction* (pp. 427–434). New York, NY, USA: ACM.
Boivin, R. (2013). On the use of crime rates. *Canadian Journal of Criminology and Criminal Justice*, *55*, 263.
Brantingham, P., & Brantingham, P. (1995). Criminality of place. *European Journal on Criminal Policy and Research*, *3*, 5–26.
Brantingham, P. L., & Brantingham, P. J. (1981). Notes on the Geometry of crime. *Environmental Criminology* (pp. 27–54). Sage Publications.
Brantingham, P. L., & Brantingham, P. J. (1993). Nodes, paths and edges: Considerations on the complexity of crime and the physical environment. *Journal of Environmental Psychology*, *13*, 3–28 (Cited by 0334).
Brunson, C., & Comber, L. (2015). *An introduction to R for spatial analysis and mapping*. London: SAGE.
Chainey, S., & Ratcliffe, J. (2005). *GIS and crime mapping*. Chichester: John Wiley and Sons.
Clarke, R. V. (1984). Opportunity-based crime rates: The diffculties of further refinement. *British Journal of Criminology*, *24*, 74–83.
Clarke, R. V., & Cornish, D. B. (1985). Modeling offenders' decisions: A framework for research and policy. *Crime and Justice*, *6*, 147–185.
Cohen, L., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*, 588–608.
Felson, M., & Boivin, R. (2015). Daily crime flows within a city. *Crime Science*, *4*, 31.
Felson, M., & Newton, A. (2015). Crime patterns in time and space: The dynamics of crime opportunities in Urban areas. Special issue. *Crime Science*, *4*.
Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, *24*, 189–206.
Heine, G. (1998). *GSM Networks: Protocols, terminology, and implementation*.
Home Office (2013). *Theft from the person: Information pack for partners*. London: Home Office.
Kautt, P., Kurland, J., & Morina, A. D. (2011). The event effect: Demonstrating the impact of denominator selection on floor and ceiling crime rate estimates in the context of public events. *Crime rates, types and hot spots*. Nova Publishers.
Kinney, J. B., Brantingham, P. L., Wuschke, K., Kirk, M. G., & Brantingham, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment*, *34*, 62–74.
Kurland, J., Johnson, S. D., & Tilley, N. (2014). Offenses around stadiums: A natural experiment on crime attraction and generation. *Journal of Research in Crime and Delinquency*, *51*, 5–28.
Malleson, N., & Andresen, M. A. (2015a). The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, *42*, 112–121.
Malleson, N., & Andresen, M. A. (2015b). Spatio-temporal crime hotspots and the ambient population. *Crime Science*, *4*.
Martin, D. (2011). *SurfaceBuilder247 User Guide*. (ESRC. URL: http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1811/outputs/Read/ece508b5-6438-4e96-99eb-8fd68f1d3b99).
Martin, D., Cockings, S., & Leung, S. (2015). Developing a flexible framework for spatiotemporal population modeling. *Annals of the Association of American Geographers*, *105*, 754–772.
Newton, A. D., Partridge, H., & Gill, A. (2014). Above and below: Measuring crime risk in and around underground mass transit systems. *Crime Science*, *3*, 1.
Office for National Statistics (2013). *The workday population of England and Wales: An alternative 2011 census output base*. London: Office for National Statistics.
Office for National Statistics (2014a). *Chapter 1: Property crime — Overview. Crime statistics, focus on property crime, 2013/14 Release*. London: Office for National Statistics.
Office for National Statistics (2014b). *Crime in England and Wales, year ending March 2014*. London: ONS.
Openshaw, S. (1984). *The modifiable areal unit problem volume 38 of Concepts and Techniques in Modern Geography (CATMOG)*. Norwich: Geo Books.
Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, *27*, 286–306.
Pauli, V., Naranjo, J. D., & Seidel, E. (2010). *Heterogeneous LTE networks and inter-cell interference coordination*.
Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place*. *Criminology*, *27*, 27–56.
Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user metadata. *PLoS ONE*, *10*, e0115545.
Smith, A., Newing, A., Quinn, N., Martin, D., Cockings, S., & Neal, J. (2015). Assessing the impact of seasonal population fluctuation on regional flood risk management. *ISPRS International Journal of Geo-Information*, *4*, 1118–1141.
Smith, C., Bowers, K. J., & Johnson, S. D. (2006). Understanding bag theft within licensed premises in Westminster: Identifying initial steps towards prevention. *Security Journal*, *19*, 3–21.
Stults, B. J., & Hasbrouck, M. (2015). The effect of commuting on city-level crime rates. *Journal of Quantitative Criminology*, *31*, 331–350.
Tompson, L., Johnson, S., Ashby, M., Perkins, C., & Edwards, P. (2015). UK open source crime data: accuracy and possibilities for research. *Cartography and Geographic Information Science*, *42*, 97–111.
Weisburd, D., & Amram, S. (2014). The law of concentrations of crime at place: The case of Tel Aviv-Jaffa. *Police Practice and Research*, *15*, 101–114.
Weisburd, D. L., Groff, E. R., & Yang, S. -M. (2012). *The criminology of place: Street segments and our understanding of the crime problem*. Oxford; New York: OUP USA.
Zhang, H., Suresh, G., & Qiu, Y. (2012). Issues in the aggregation and spatial analysis of neighborhood crime. *Annals of GIS*, *18*, 173–183.