UK open source crime data: accuracy and possibilities for research

Lisa Tompson, Shane Johnson, Matthew Ashby, Chloe Perkins & Phillip Edwards

Taylor & Francis
Taylor & Francis Group

UK open source crime data: accuracy and possibilities for research

Lisa Tompson[a]*, Shane Johnson[a], Matthew Ashby[a], Chloe Perkins[b] and Phillip Edwards[b]

[a]*UCL Department of Security and Crime Science, University College London, 35 Tavistock Square, London, WC1H 9EZ, UK;*
[b]*Department of Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, UK*

In the United Kingdom, since 2011 data regarding individual police recorded crimes have been made openly available to the public via the police.uk website. To protect the location privacy of victims these data are obfuscated using geomasking techniques to reduce their spatial accuracy. This paper examines the spatial accuracy of the police.uk data to determine at what level(s) of spatial resolution – if any – it is suitable for analysis in the context of theory testing and falsification, evaluation research, or crime analysis. Police.uk data are compared to police recorded data for one large metropolitan Police Force and spatial accuracy is quantified for four different levels of geography across five crime types. Hypotheses regarding systematic errors are tested using appropriate statistical approaches, including methods of maximum likelihood. Finally, a "*best-fit*" statistical model is presented to explain the error as well as to develop a model that can correct it. The implications of the findings for researchers using the police.uk data for spatial analysis are discussed.

Keywords: open data; indices of similarity; spatial accuracy; spatial crime analysis; geomasking

Introduction

The political impetus to publish crime statistics online has grown in the last decade, and has accompanied a more general move towards the provision of open geographical data across private and public sector organizations. In the United Kingdom (UK), since 2011 data regarding individual police recorded crimes have been made openly available to the public via the police.uk website. For the first time, this presents researchers with the potential opportunity to test theories of crime causation at the micro- and meso-level using data for an entire country, and for evaluators of crime reduction initiatives to estimate the impact of interventions for any location within the UK. However, while open data offers many such benefits, it suffers from some notable limitations. For instance, to desensitize it so that victims are not identifiable (a critical criterion of the UK Data Protection Act, 1998), the data are obfuscated using geomasking techniques to reduce their spatial accuracy. At present it is unknown to what extent (if any) this process renders the data unsuitable for spatial analysis and theory testing. Nor is it known at what spatial resolution the open source crime data *might* be suitable for such analysis.

This paper has three primary aims:

- To quantify the spatial accuracy of the open crime data available on police.uk by comparing it with police recorded crime data at different levels of geographic resolution.

- To identify potential sources of systematic error in the police.uk data and test hypotheses using appropriate statistical approaches including methods of maximum likelihood estimation.

- To develop a "*best-fit*" statistical model of the true spatial distribution of (police recorded) crime using the police.uk and other relevant sources of open data. The intention is to explain the error as well as to develop a model that can correct for it.

Ultimately, the paper will assess the extent to which the open source data could be used to test theories of crime – or to answer other empirical questions – and what issues researchers should be aware of when using it.

The paper is organized as follows, in the next section we provide a comprehensive overview of the police.uk data. This is followed by a discussion of units of geography in spatial analysis. Next, we articulate our research questions and describe our analytical approach. We then present findings and discuss their implications for criminological enquiry and the conduct of empirical evaluations.

Police.uk data

Although the argument had been made before (see, for example, Ecclestone 1998), in the UK the Guardian newspaper is frequently credited with starting the open data movement in 2006 with the "free our data" campaign. In 2010 the UK Open Government License was created to overcome

copyright issues regarding the release of public datasets, and the data.gov.uk site was launched to provide access to myriad sources of data. The Coalition Government of the day conspicuously adopted a "transparency agenda" with regards to public service delivery that was imbued with the principles of the open data movement.

The Police – as a public service – and the data they collected fed into this new "democratic transparency" program (Hohl, Bradford, and Stanko 2010; Home Office 2010). The aims of publishing such crime data via crime mapping stemmed from three policy objectives (Chainey and Tompson 2012):

- To improve the credibility of crime statistics in the mind of the public;
- To provide a more community-focused police service;
- To inform, engage and empower the public to participate in crime prevention efforts.

Previously, the publication of crime data online via crime maps had been piecemeal, with some forces being very active in this area,[1] and others not. With political impetus (Home Office 2008), in December 2008 a standardized approach was adopted across the 43 police forces in England and Wales. This presented crime data mapped thematically by administrative areas (mostly but not always at the Census Middle Super Output Area – which have a mean population of around 2940 residents). However, as each Force was responsible for the processing and uploading of their data to the website, this resulted in an inconsistent range of geographies being used. The provision of spatial crime data remained of varying coverage and quality across the country, which was not conducive to analysis, academic or otherwise.

A step change occurred in December 2010, when crime data were published on the police.uk website. Data were presented at the street level alongside additional information intended to provide sufficient context for the public to be able to interpret observed patterns. To visualize the data the police.uk website uses graduated point symbology to depict clusters of crime on particular streets, which change in size according to the level of geography at which the maps are viewed.

The police.uk anonymization method

To comply with the UK Data Protection Act 1998, and in accordance with the Information Commissioner's Office (Graham 2012; ICO 2012) report on the implications of sharing sensitive crime data, the data visualized – and available for download – on the police.uk site are processed according to strict criteria. A key aspect of this process involves the obfuscation, or geomasking (e.g., Armstrong, Rushton, and Zimmerman 1999) of crime locations.

Geomasking adds "stochastic or deterministic noise to the original data matrix by modifying the geographic coordinates of the data points" (Kwan, Casas, and Schmitz 2004, 16). Such techniques aim to minimize the extent to which the anonymized data can be "reverse geocoded" – which would threaten an individual's privacy – whilst preserving the general spatial pattern in the data. Areal aggregation is one technique but this can lead to a loss of information (e.g., Wieland et al. 2008) and so spatial epidemiologists have devised a number of alternative methods using affine and randomizing transformations to name a few (e.g., Armstrong, Rushton, and Zimmerman 1999; Cassa et al. 2006; Wieland et al. 2008). In terms of the impact of geomasking techniques, Leitner and Curtis (2004) examined how different methods affected student's perceptions of the spatial pattern of a sample of Homicide victim's residences in Baton Rouge. Other scholars have examined how geomasking would affect the spatial patterns observed in geomasked public health data (e.g., Armstrong, Rushton, and Zimmerman 1999). However, as far as we are aware, no studies have examined how geomasking techniques have actually affected the spatial accuracy of crime data (such as the police.uk data) to which a geomask has been applied and that has subsequently been made publicly available. This is the aim of the current paper.

Considering the police.uk data in particular, it is likely that geoprivacy concerns (e.g., Wieland et. al. 2008) have also motivated the decision by the providers to aggregate data for some crime types together (e.g., burglaries to dwellings with burglaries to non-dwellings) and to only release the data on a monthly basis[2] (the last working day of the subsequent month). The less frequently the data are released, the less likely it is that residents will be able to correlate local news of crimes in their area with locations of crime events. Data release timing is thus used as a means of privacy protection (Kounadi, Bowers, and Leitner 2014).

In terms of the process employed in the UK, each of the 43 UK police forces upload crime incident data in a standardized format to a secure server on a monthly basis (in most cases). A spatial "jitter" is then applied so that no actual crime event location is identifiable (see, Bridwell 2007) as follows (Police.uk 2014):

- The location of each crime is compared against a master list of "snap points" to find the nearest (see below).
- The coordinates of the crime are then replaced with the coordinates of the snap point.
- If the nearest snap point is over 20 km away, coordinates of zero are assigned so that the crime is not shown on the subsequent map.

The master list of "snap points" – which is not publicly available – contains over 750,000 points, generated so that they appear over the *center point of a street*, or above a public place such as a commercial premise or public land.[3]

Points are never placed over specific dwellings. Each point has a catchment area that contains at least eight postal addresses[4] (in 2011 this threshold was 12), or no addresses at all. The list is continually updated by representatives in the police forces so that key locations (such as railway stations or premises of interest) are represented in the database.

This process has an obvious impact on the positional accuracy of the data. They are necessarily inexact when there is a chance of a victim being identified. However, for research and other purposes, the question is whether the spatial patterning of crime is preserved in the data or if it is obscured in some way and, if so, how? The accuracy of the point patterns observed in the data is, in part, determined by the database of snap points employed and these issues are in addition to the fact that there may be inherent recording inaccuracies in the original crime data. For example, some crimes may occur at places that are not represented in a gazetteer system – commonly referred to as non-addressables (for other examples see Ratcliffe 2001; Chainey and Ratcliffe 2005).

Developments and critiques of police.uk data

When initially launched, the police.uk website drew large-scale criticism due to anecdotal errors that were identified in the data (Daily Mail 2011a). For example, crimes were erroneously attributed to quiet cul-de-sacs by virtue of their close proximity to police stations – to which some offenses, for which the locations were unknown, had been geocoded by the police. Moreover, the data seemed not to represent the riots in the summer of 2011 (Daily Mail 2011b). In truth, the public – including reporters – were not aware of the nuances of geographic crime data, or the geomasking methods used by the police.uk site to protect the privacy of victims. However, the question of how accurate the spatial data are remains.

As well as the loss of spatial information caused by the use of snap points, some other privacy-driven limitations of police.uk are noteworthy. The grouping and categorization of crimes can further obfuscate the data. For example, all violent and sexual offences are aggregated into a single category. This category includes all assaults, regardless of the degree of injury caused, such that murders are counted equally with assaults that caused no injury at all. Finally, fraud offences are excluded from police.uk data entirely.

Over time, the police.uk site has undergone a number of developments. Some of the crime types have been disaggregated so that the crime categories are less coarse.[5] A customizable area tool was introduced so that users of the site can view local crime patterns by police geography (e.g., local neighborhood police area), or within a one-mile radius from a postcode, or for a self-drawn area. Temporal trends were also provided so that users could see how crime rates have changed over time. Crime

outcome information was also added so that the public could see the criminal justice outcome for an event.

Units of geographic analysis

Geographic data can be visualized and analyzed at multiple scales of interest. The choice of which unit of analysis to employ (e.g., micro, meso or macro) has been a perennial topic of debate amongst geographers for several decades (Anselin and Getis 1992). Research concerned with crime is no exception.

Typical units used in the spatial analysis of crime include preexisting census geography (e.g., Output Areas and multiples thereof in the UK, or Blocks, Groups and Tracts in the US) or politically designed administrative areas (e.g., Parishes, Wards, Local Authorities or Counties). These are often employed because other data are collected at these levels of geography that permit statistical comparison. In terms of theory testing and falsification, it is important that the geographical unit selected should match the spatial scale over which the theoretical mechanisms of interest are hypothesized to operate as crime can form very different patterns at different scales of analyses (Brantingham, Dyreson, and Brantingham 1976; see also, Hipp 2007; Ouimet 2000; Wooldredge 2002).

This illustrates the Modifiable Areal Unit Problem or MAUP (Openshaw 1984; see also Bailey and Gattrell 1995) which can occur when point data (including crime) are aggregated into areal units. In the above example, the problem concerns spatial scale and the fact that patterns displayed at large geographic units can mask important heterogeneity that may be observed for smaller units of analysis (Robinson 1950).

The extent to which this a problem will, as noted, depend upon the theory tested, or the purpose of the analysis undertaken. The point of central importance is that perceived patterns in aggregated data can be influenced by the precise choice of unit and the associated boundaries, and changing these can lead to potential errors of inference. This problem may (or may not) be exacerbated for research based on the analysis of police.uk data as the original point data are "allocated" to one set of boundaries (defined by the snap points) as part of the anonymization process.

Considering the spatial analysis of crime more specifically, this can be undertaken for at least three purposes: the targeting of crime reduction resources, applied evaluation research, and academic enquiry. In policing circles, the aim of analysis is typically to better understand the clustering, or concentration of crime so that resources can be allocated appropriately (e.g., Bowers, Johnson, and Pease 2004; Chainey and Ratcliffe 2005). For these purposes, it makes sense to employ a unit of analysis that is relevant to the decision-maker. In the case of tactical policing, small spatial units will frequently be preferred

as police commanders will have limited resources available and hence will likely want to deploy them to a limited number of precise locations.

For more strategic (longer-term) approaches to crime reduction that involve multi-agency groups (which bring together a consortium of public agencies responsible for dealing with the effects of crime) a meso-level administrative geography may be used, partly because this will be familiar to all partners but also because this will reflect the spatial scale over which they will jointly be responsible. Evaluation research may be conducted at a range of spatial scales, and should align with the area over which interventions are implemented and are consequently anticipated to have an impact.

In academic research, spatial analysis is usually conducted with the express purpose of generating insight into the causal processes that influence crime occurrence. As Weisburd, Bruinsma, and Bernasco (2009, 24) stress, "The unit of analysis for geographic crime studies cannot be divorced from the social contexts of crime and criminals." The immediate context, or environment, of crime is central to several criminological theories (Brantingham and Brantingham 1991; Cornish and Clarke 1986; Wikstrom 2006), yet the "environment" is a somewhat nebulous concept and no universal criterion exists to define it in a geographic sense. Most studies undertaken under the rubric of these theories use aggregated crime data as the outcome or dependent variable in spatial analysis, with the unit size varying tremendously across studies. For example, the unit of analysis used in research concerned with "micro places" has ranged from US census blocks (Bernasco and Block 2011); clusters of a hundred addresses on a street (Groff, Weisburd, and Morris 2009; Weisburd et al. 2004), and street segments (Johnson and Bowers 2010), also known as block faces (Groff, Weisburd, and Yang 2010; Smith, Frazee, and Davison 2000) to individual buildings and addresses (Polvi et al. 1991; Sherman, Gartin, and Buerger 1989).

Research completed in the tradition of social disorganization theory (Sampson and Groves 1989; Sampson, Raudenbush, and Earls 1997; Kubrin and Weitzer 2003) in particular focuses on the relationship between crime and neighborhood characteristics as, according to the theory, important social processes that operate at the neighborhood or community level influence the occurrence of crime. Such studies have examined patterns for "neighborhoods" varying in size, although the US census tract is the most common (Kubrin and Weitzer 2003). Clearly then, academic research and analyses conducted by practitioners alike frequently employ data aggregated to a variety of spatial scales, and might take advantage of data such as that provided on the police.uk website. Whether this is appropriate, of course, depends upon the spatial accuracy of the data.

The current study

To examine the spatial accuracy of the police.uk data, we compared it with the police recorded crime data used to generate it, using data for the West Midlands Police Force area. This area is predominantly metropolitan, covering three cities and 22 large- and medium- sized towns, and encompasses a wide range of urban geography. Both datasets were first aggregated to four different areal units of analysis using a geographical information system (GIS). The areal units considered were UK postcodes (PC), census output areas (OA), lower super output areas (LSOA), and middle layer super output areas (MSOA). The latter three types of area are commonly used to test ecological theories of crime (see above), while the first provides insight as to the spatial accuracy of the police.uk data at a very fine level of resolution. Figure 1 provides an illustration of the boundaries for the different Census units of analysis, while Table 1 provides descriptive statistics regarding the mean area, population and number of households for the same units.[6]

Three analytic strategies were adopted. For the first, we sought to quantify the extent to which the area level counts generated using the sample of police.uk data agree with those for the police recorded crime data at each spatial scale considered. For the second, we tested hypotheses (see below) regarding those area-level factors that
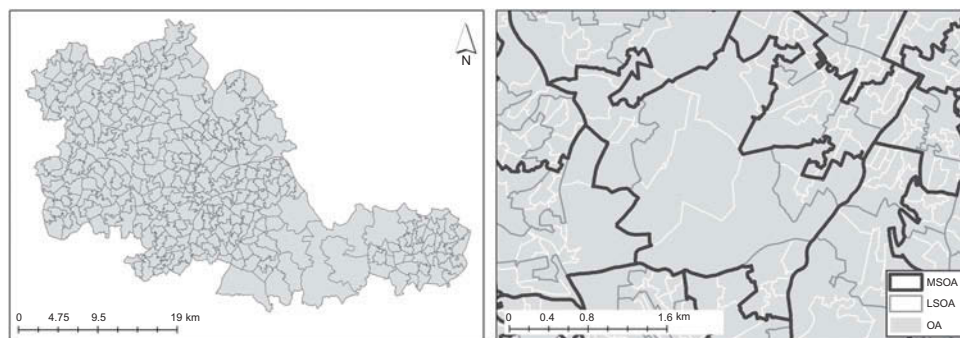

Figure 1. Illustration of the nested Census geographic units. Contains National Statistics data © Crown copyright and database right 2012. Contains Ordnance Survey data © Crown copyright and database right 2012.

Table 1. Summary information for population and households for the UK and the West Midlands study area.

Geography	UK				West Midlands			
	n	Median size (km^2)	Mean population	Mean households	n	Median size (km^2)	Mean population	Mean households
Output Area	181,408	0.07	309	129	8468	0.06	323	128
Lower Super Output Area	34,753	20.56	609	470	1680	20.78	620	455
Middle Super Output Area	7201	91.92	2940	2271	356	93.59	2928	2148

might significantly influence the spatial accuracy of the police.uk data. Finally, we sought to establish whether a spatial model that uses the police.uk data and that incorporates these factors can be used to improve the estimates and, if so, by how much.

Different types of crime typically exhibit different geographical patterns and varying levels of spatial clustering. Consequently, we examine the spatial accuracy of the data for different categories of crime, in this case:

- antisocial behavior;
- burglary (which includes burglaries to dwellings and non-dwellings);
- criminal damage and arson;
- robbery (commercial robbery and robbery to the person); and
- vehicle crime (including theft of, and from, a vehicle).

Before describing the methods employed and our findings, we consider some of the factors that might lead to systematic inconsistencies in the spatial patterns observed for analyses conducted using police recorded crime and police.uk data. First, we hypothesize that there will be differences across crime types. One reason for this is that some types of crime are harder to accurately geocode than others. For example, in the UK when a burglary to a dwelling is reported, the police can precisely determine the location of the dwelling using a gazetteer of postal addresses. However, for crimes that happen at "non-addressable" locations such as a park, subway or outbuilding, it is more difficult to assign exact coordinates to the crime location, even if the victim can provide an exact description of the location (which they might not be able to do). For such locations, it is likely that there will be fewer "snap points" in the police.uk database, which would be expected to lead to inaccuracy in the data.

Burglary is the only type of crime that, by definition, occurs in a defined building (although sometimes those buildings will be located in gardens or yards), which means that it is less subject to data recording inaccuracies. However, other types of crime are affected to differing degrees. For instance, most vehicle crime occurs on the street or in the vicinity of people's residences, but a

Table 2. Per cent of crimes recorded as happening at an addressable location.

Crime type	Offences at addressable location-types
Anti-social behavior	91.4%
Burglary	95.7%
Criminal damage and arson	95.7%
Robbery	91.3%
Vehicle crime	95.3%

small proportion happen in parking lots or other areas, which will be less well represented in police gazetteer systems.

A preliminary analysis of the police.uk data illustrates this point. The results shown in Table 2 – derived by inspecting the "location type" field included in the police.uk data – indicate the percentages of crimes that were indexed to addressable locations for each type of crime. Those for which the location was reported as occurring in a "park/open space," "parking area," "pedestrian subway" and "sports/recreation area" were classified as non-addressable. It is apparent that most crimes *are* indexed to addressable locations, but that geo-coding accuracy is likely to be highest for the crimes burglary, criminal damage and arson, and vehicle crime.

On the basis of the above logic, we expect the spatial accuracy of the police.uk dataset to vary between crime types, conditional on the degree to which they occur at locations that are more easily addressable. Specifically:

H1 – we anticipate the spatial accuracy of the police.uk data to be highest for the crimes of burglary, criminal damage and arson, and vehicle crime.

Given the way in which the police.uk "snap point" database was constructed (see above):

H2 – we anticipate the police.uk data to be the most accurate for larger geographic units of analysis, in this case the MSOA and LSOA census geographies.

On a similar note, because the residential population in an area will directly influence the number of snap points to be found within it:

H3 – we predict that – all else equal – more spatial error will be observed for areas with smaller populations,

However, the number of snap points in an area will not be determined solely by the number of residents within it. In particular, given the way the police.uk snap point database was constructed, the number of street segments within an area should – all else equal – influence the number of snap points within it. Thus:

H4 – we predict that at the area level, the spatial accuracy of the police.uk data will be highest for areas with a larger number of roads within them.

H5 – A final hypothesis that concerns the influence of the snap point database is that as it has been improved over time through the addition of new snap points (supplied by Police Forces to more accurately reflect locations of relevance to crime occurrence), the spatial accuracy of the police.uk data should also have improved over time, being more accurate for the most recent years.

The next three hypotheses consider the interaction between the morphology of the census geographies considered and the police.uk "snap point" database. First, census units are (approximately) standardized by population, and consequently the geographic size of a census area provides an indication of "urban density" or how much open space there is likely to be within it. As larger areas will have lower urban density (for a discussion of geomasking techniques based on population density, see Cassa et al. 2006), they are likely to have fewer snap points within them, and hence:

H6 – we predict that in larger areas – which will likely have lower urban density as census areas have similar population sizes – there will be a greater risk of crimes that occurred within them being relocated to a snap point outside their boundary.

Second, areas with more nondomestic land-use are also likely to have fewer snap points within them per unit area, and so:

H7 – we expect the spatial accuracy of the police.uk data to be lower for areas with more nondomestic land uses within them.

Third, census boundaries are often delineated by major roads, with one side of a major road being in one area, the one opposite in another. However, as the police.uk snap point data-set was derived in a different way using Voronoi polygons (see below):

H8 – we anticipate there to be larger differences in area level counts of crime estimated using the police recorded crime and police.uk data for areas with major roads in them.

Finally, urban form tends to differ in a variety of ways as the distance from the nearest urban center increases. For example, locations near to urban centers may have constraints associated with the amount and location of land available for housing stock. This may result in small pockets of high-density housing, but more generally residential housing may be more dispersed the nearer an area is to an urban center. An area may also have a greater mix of, and variation in, land uses the closer it is to an urban center. Such variation may influence the number of snap points in a census area, as well as their spatial distribution. In theory, this variation should affect the spatial accuracy of the police.uk data:

H9 – we predict there to be decreasing accuracy in the police.uk data the closer an area is to the city center.

Method

Data preparation

Police recorded crime data were provided by West Midlands Police (UK) for the years 2011 to 2013, and it is for this Force area that the analyses are conducted. Police.uk data from the same period (and same area) were downloaded from the website via the application programming interface (API). The data preparation was undertaken in the following sequence.

First, the crime categories provided in the police.uk were inspected to find categories, or combinations of them in the police recorded crime data that could serve as comparators. The crime type categories used by police.uk did not align exactly with the crime type categories commonly used by the police so required some matching. For example, for the police recorded crime data, arson and criminal damage were aggregated into one category for consistency with the police.uk data. In the case of antisocial behavior (ASB) it was necessary to exclude incidents from the police recorded crime data that had no geographic coordinates (many of which were classified as hoax calls). This resulted in a loss of approximately one per cent of the ASB data.

Next, the crime data from both data sources were "joined" to the nearest areal unit for each level of analysis (postcode, OA, LSOA and MSOA).[7] Spatial lag values representing the mean count of crimes in adjacent areas were created for each area, year and crime type.[8] The

distance to the nearest urban center was calculated as the distance of each unit boundary to the nearest city center. As the influence of urban centers is likely to be more pronounced the closer an area is to one, the logarithm of the distance was used in all analyses.

The "Area of Non Domestic Buildings" was calculated using data from the Generalized Land Use Database.[9] These data were collected at the 2001 census OA level, which led to instances where 2011 OAs were not directly comparable. Where a 2011 OAs comprised multiple 2001 OAs, the values were aggregated accordingly. Where 2011 OA had no corresponding 2001 OA (and therefore had missing nondomestic building data), we assigned the mean value observed across all OAs.

The length of different types of roads (any or major roads) in an area were computed using data from the integrated transport network (ITN). Twelve output areas had no roads at all, and so were given a value of 0 for each type of road.

Analytic strategy and results

How similar are area level counts of police.uk and police recorded crime data?

A number of analytic approaches could be used to examine the similarity of the spatial pattern in the two sources of data.[10] Here, we use Andresen's (2009) spatial point pattern test. Shown as Equation (1), this is a global index designed to examine the similarity in spatial patterns for two samples of spatial data aggregated to some areal unit.

$$S = \frac{1}{N} \cdot \sum_{i=1}^{N} s_i \qquad (1)$$

where, N is the number of areal units considered,

s_i is equal to one if the counts of events from two different sources in area i are similar, and zero otherwise.

To test for similarity in spatial patterns, one source of data is used as a reference data-set (sometimes referred to as the "gold standard") against which the other is compared. Confidence intervals are estimated for the point estimate of the crime count for the reference data-set and similarity is established if the values for the second data-set fall within this interval. In other words, the counts from two sources of data are considered to be "similar" if any difference between the counts could be explained by sampling error. The confidence intervals are estimated using a simple non-parametric bootstrap methodology. Implemented in the R programming language (http://www.R-project.org), a Monte Carlo simulation is used to generate the estimates as follows: (1) M events are sampled (with replacement) from the full set of (M) observed events using a uniform random number generator; (2) area level counts are computed for the sampled data; (3) this process is repeated

many times (in this case 100 times) to generate a distribution of counts for each area; (4) the mean count (and 95% confidence intervals) for each area are computed using the data from the 100 iterations. As two sources of data may have different overall counts of crime (due to the aggregation of offense types to categories – see above), and because the test is concerned with the extent to which the spatial patterns are similar, rather than testing the absolute differences in counts, it is the percentage of crime (i.e., the area count divided by the total) associated with each area that is compared.

To illustrate how the confidence intervals are constructed, Figure 2 shows an example of the mean values and 95% confidence intervals generated using the police recorded crime data for the crime of burglary at the census OA level (N = 8468). For each OA, the mean values and the associated confidence intervals generated using the MC simulation are plotted against the values observed in the police recorded crime data. As would be expected, the mean values obtained from the MC simulation are largely identical to the observed values. For this example, the values for the police.uk data (not shown in Figure 2) were within the confidence intervals computed for the police recorded crime data for 6123 of the 8468 comparisons. Consequently, the global value of S (similarity) was 6123/8468 = 0.72. As well as providing a global index of similarity, the approach allows the specific areas that *are* similar to be identified, mapped or analyzed. Figure 3 shows a map of those areas that had similar (and nonsimilar) values for the police.uk and police recorded crime data for burglary in 2012.

The above procedure was repeated for each spatial unit of analysis (MSOAs, LSOAs, OAs, and postcodes), for each type of crime (burglary, vehicle crime, ASB, criminal damage, and robbery), for each year data were available (2011, 2012, and 2013) and for the aggregate period 2011–2013. To simplify presentation, the (set of 20) results are displayed for each spatial unit of analysis separately.

Figure 4 shows the results for the MSOA geography. In addition to showing the index of similarity, the columns near the y-axis enumerate the overall counts of crimes for each source of data. It is apparent that the counts are similar, but not identical. This is due to the way in which recorded crimes are aggregated to more general categories of offenses in the police.uk data.

At this level of resolution, the indices of similarity are generally high, typically between 0.9 and 0.95. That is, around 90% of MSOAs have similar counts of crime for estimates generated using the two sources of data. This is particularly the case for the crimes of burglary, vehicle crime and robbery. With the exception of criminal damage, it also appears to be the case that levels of similarity have increased over time, as predicted (Hypothesis 5). Moreover, levels of similarity appear to be greater for comparisons made for each year than for the aggregate
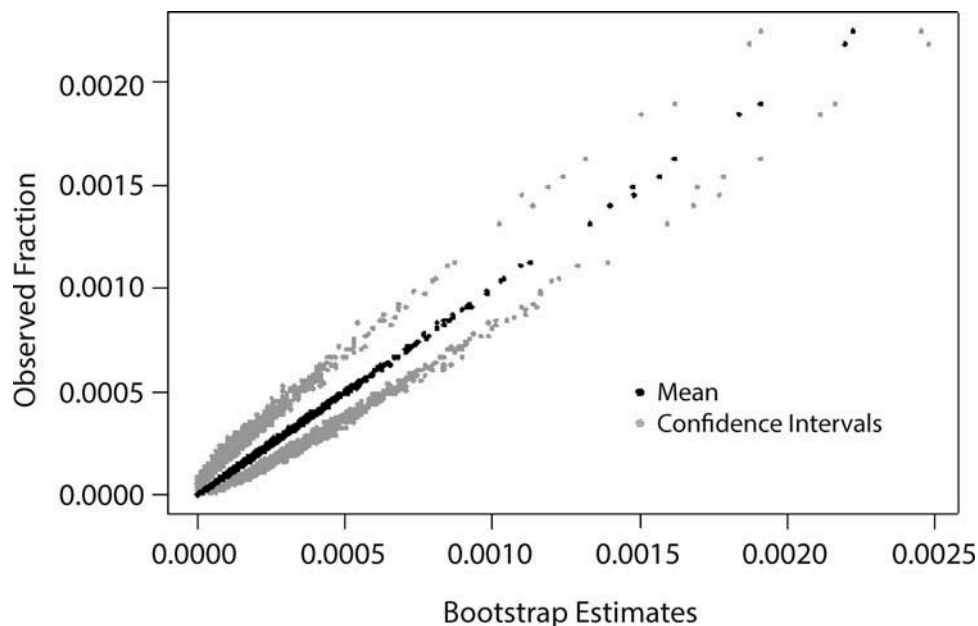
L. Tompson et al.

Figure 2.   Mean values and 95% confidence intervals for one sample of data plotted against the observed values.
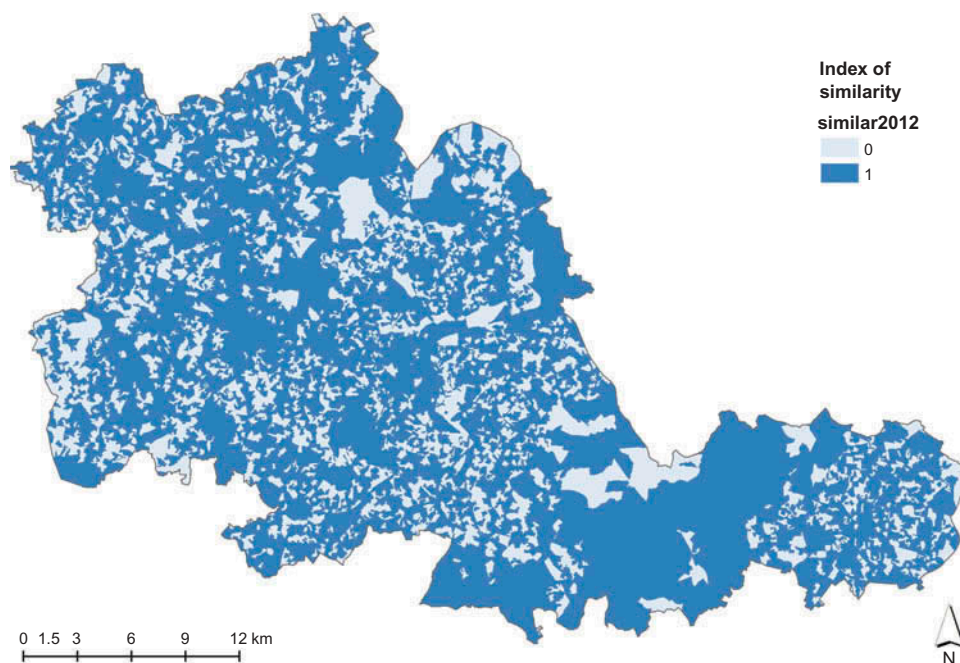
Figure 3.   Indices of similarity values mapped for 2012 burglary data. Contains National Statistics data © Crown copyright and database right 2012. Contains Ordnance Survey data © Crown copyright and database right 2012.

period 2011–2013. The pattern was least consistent for incidents of ASB, but the data appear to have improved substantially over time.

At the LSOA level, patterns are again quite similar across the two data sources (see Figure 5), although the overall level consistency is lower at around 0.85. The exception to this is for robbery for which the yearly

estimates (around 0.7) are the lowest for all crimes considered. Otherwise the patterns are in line with those observed at the MSOA level. This is not surprising given that LSOAs and MSOAs are both quite large areas (see Table 1).

Figure 6 presents results at the OA level. This shows, with the exception of ASB, patterns are more consistent
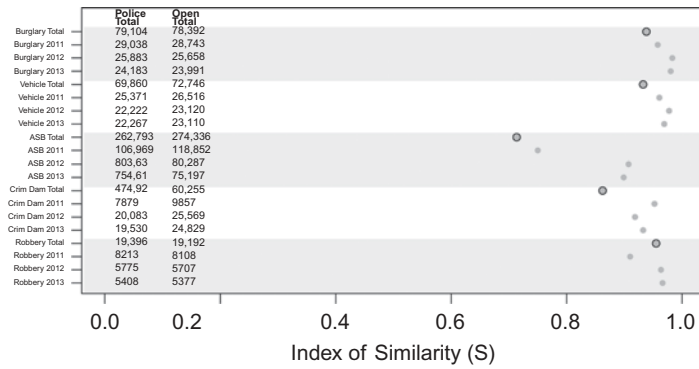
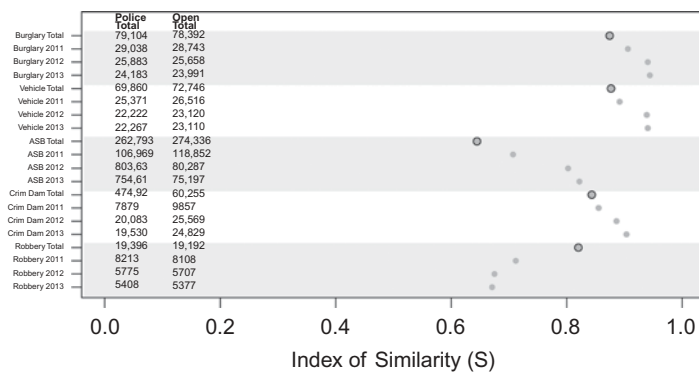Figure 4.   Analysis of similarity at the MSOA level.

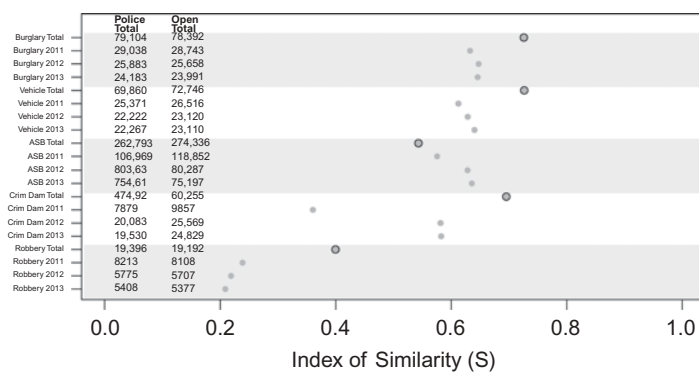Figure 5.   Analysis of similarity at the LSOA level.

Figure 6.   Analysis of similarity at the OA level.

for the aggregate period 2011–2013 than they are for any one-year interval. Consistent with Hypotheses 1 and 2, levels of similarity are generally lower for this (smaller) unit of analysis, and, are best for the crimes of burglary and vehicle crime (over 0.6 for each one-year interval and around 0.75 for the 2011–2013 interval). They were particularly poor for the crime of robbery, being only slightly above 0.2 for the three one-year periods considered. At

this level of resolution, with the exception of robbery, it also appears to be the case that the spatial accuracy of the police.uk data has improved over time (substantially for incidents of criminal damage), as predicted.

As shown in Figure 7 and consistent with Hypothesis 2, the index of similarity was the lowest observed for analyses conducted at the postcode level (0.2 or less). In line with Hypothesis 1, levels of consistency were
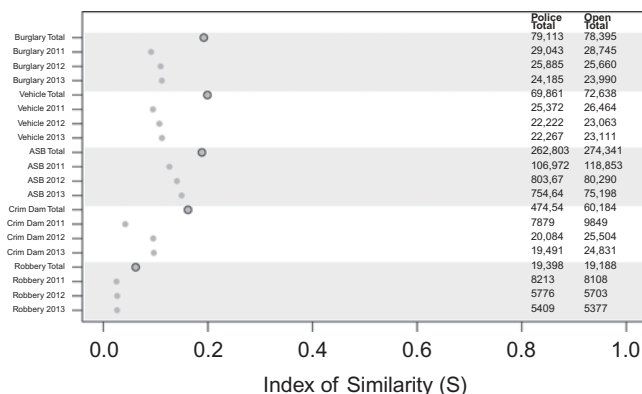
Figure 7. Analysis of similarity at the postcode level.

particularly poor for the crime of robbery. As with analyses conducted at the OA level, the level of consistency appears to have increased over time, but was highest for the aggregated interval 2011–2013.

What area level factors influence the spatial accuracy of police.uk data?

To test hypotheses regarding factors that are associated with the spatial accuracy of the police.uk data, we use logistic regression models to see if there are systematic differences between the types of areas for which the police.uk and police recorded crime are similar and those for which they are not. Analyses were conducted for each of the different census geographies considered (census data were not available at the postcode level), but those shown are for census OAs. The reason for this is that for the MSOA and LSOA geographies, the Police.uk data were on the whole rather accurate.[11] Table 3 shows the results from five logistic regression analyses, one for each type of crime. All analyses were conducted using data for the aggregate period 2011–2013.

The results provide fairly consistent support for Hypotheses 3, 4, 6 and 8, with the coefficients being in the direction expected in all cases, expect for the amount of major road in an area for the crime of robbery. That is,

areas with a greater amount of road in them (all coefficients were in the right direction, and four out of five were statistically significant), with a larger population (four out of five tests were statistically significant), a smaller geographic area (four out of five coefficients were in the right direction, and two out of five were statistically significant), or (with the exception of robbery for which the reverse was true) less major roads (with four out of five tests being statistically significant) in them tend to display similarity across the two sources of data.

Some support is also found for Hypothesis 7 that levels of similarity would be lower for areas with more nondomestic land use in them (four out of five coefficients were in the right direction, and two tests were statistically significant), although the findings are nonsignificant for vehicle crime and criminal damage. Moreover, for robbery we find that OAs with more area of non-domestic land use are actually more (rather than less) likely to have similar estimates, although the effect is small. Apropos Hypothesis 9, with the exception of robbery, we find that areas located further from the city center are more likely to have similar values across the two data-sets. In the case of robbery, the reverse appears to be the case. In sum, whilst at least partial support is provided for all of our hypotheses, there are some differences across crime types.

Table 3. Logistic regression analysis of sources of inconsistency in the police.uk data.

	Burglary		Vehicle Crime		ASB		Criminal Damage		Robbery	
	e^β	z	e^β	z	e^β	z	e^β	z	e^β	z
Road Length/km	1.59***	7.05	1.23**	3.64	1.24***	4.12	1.08	1.41	1.14**	2.61
Major Road Length/km	0.71*	−2.95	0.67***	−3.88	0.69***	−3.87	0.89	−1.15	1.54*	4.35
Area of Nondomestic	0.99***	−4.16	0.99	−1.35	0.99**	−3.51	0.99	−0.21	1.01***	5.18
Population/100	1.07*	2.016	1.09**	2.69	0.97	−1.08	1.20***	5.32	1.33***	8.96
Area	0.63*	−2.13	0.82	−1.34	1.06	0.36	0.84	−1.29	0.47***	−4.83
Log (distance to nearest city)	1.29***	6.16	1.37***	8.33	1.25***	6.22	1.03	0.89	0.76***	−4.83

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Can the police.uk estimates be improved?

Given that the analyses presented above suggest that there are systematic biases in the police.uk data, the possibility exists that simple spatial models could be used to improve the area level estimates generated using the data. To examine this, we use regression models to try to correct for the error in the data. In the first instance, we fit parameters using a training or in-sample data-set (2012 data). In the second, we apply these parameters to an out of sample data-set (2013 data) to examine the utility of the approach. Rather than reporting the results for all possible analyses, for the reason discussed in the previous section, we report those conducted using data for census OA areas.

Before presenting findings, we rehearse the logic of the approach. In the event that the police.uk data fit the official recorded crime data well, the generic model shown in Equation (2) should describe the data.

$$x_{i,t,c} = \beta_0 + \beta_1 y_{i,t,c} \tag{2}$$

where, $x_{i,t,c}$ is the count of official police recorded crime in location i at time t for crime type c

$y_{i,t,c}$ is the count of police.uk crime at location i at time t for crime type c

and the β are parameters to be estimated empirically

However, as suggested above, at lower levels of spatial resolution systematic errors are introduced into the data due to the way in which the police.uk data are generated. To improve the model (of official police recorded crime data), a spatial lag term may be added to model the spatial error associated with the geomasking procedure applied during the generation of the police.uk data. That is, we assume that some of the crimes that actually occurred in (say) area A will have been erroneously allocated to neighboring areas. Thus, by modeling the counts for adjacent areas we seek to correct for this error. The model can be written as follows:

$$x_{i,t,c} = \beta_0 + \beta_1 y_{i,t,c} + \beta_1 y^{|}_{i,t,c} \tag{3}$$

where, $y^{|}_{i,t,c}$ is the count of police.uk crime in the immediately neighboring areas of location i at time t for crime type c (the neighbors for each area were identified using the Queen's criterion).

Additional independent variables were included to model the biases discussed in the previous section, which results in a model of the form:

$$x_{i,t,c} = \beta_0 + \beta_1 y_{i,t,c} + \beta_2 y^{|}_{i,t,c} + \sum_{j=3}^{N+3} \beta_j z_{i,j} \tag{4}$$

where, N is the number of independent variables included in the model, and

$z_{i,j}$ is a vector of values for variable j for area i

A number of approaches were used to examine the utility of the models produced, but for comparison with the analyses presented above, we report the index of similarity, calculated using the values predicted by the spatial model. In the event that the model is of value, these should be greater to those reported in Figure 6.

Table 4 shows the coefficients for each type of crime, estimated using a linear model and the police.uk 2012 data.[12] It also shows the index of similarity for the 2013 data computed using the police.uk data alone, and using the estimates computed using the spatial model. In the case of burglary, all of the coefficients were statistically significant and the estimates based on the spatial model are clearly an improvement on the police.uk data alone.

In each case, estimates based on the spatial model offer an improvement over the police.uk data. As the parameters estimated for one year (2012) were used to compute the predictions for another (2013) this suggests that the influences identified are stable to some degree, and that there is benefit in taking this kind of approach. Considering general trends in the coefficients, three (police.uk, police.uk lag and nondomestic area) of the

Table 4. Linear regression analysis of police recorded crime data, and index of similarity for original police.uk and (police.uk) modeled data.

	Burglary	Vehicle Crime	ASB	Criminal Damage	Robbery
Intercept	1.26**	0.29	4.09**	0.29	0.77**
Police.uk$_{2012}$	0.66**	0.44**	0.87**	0.43**	0.79**
Police.uk$_{2012}$ Lag	0.19**	0.14**	0.06**	0.14**	0.08**
Total Rd	0.01**	−0.01*	0.00	−0.01*	0.01**
Major Roads	−0.15	0.12	−0.18**	0.12	−0.01
Area	−0.57**	0.05	1.11*	0.01	−0.22**
Nondomestic	0.01**	0.01*	0.01*	0.01**	0.01**
Dist City	−0.13**	0.01	−0.39**	0.01	−0.09**
S$_{original}$ 2013	0.65	0.64	0.64	0.58	0.21
S$_{model}$ 2013	0.79	0.71	0.71	0.71	0.29

*$p < 0.05$, **$p < 0.001$.

independent variables differed in magnitude but were in the same direction for all crime types. For the total length of road variable, the coefficients were small but varied in direction. For the major roads and distance to the city center variables, those coefficients that were statistically significant were in the same direction.

Thus, there is some consistency in the findings across crime types, but the coefficient estimates are by no means identical. This is not surprising for at least two reasons. First, differences in magnitude are to be expected as the number of offenses differs across crime types. Second, different types of crime exhibit different spatial patterns, and so will be susceptible to different types of spatial error in the police.uk data. The findings thereby suggest that different models will be necessary for different types of crime.

Discussion

The provision of publicly accessible data regarding the location of crime events for the whole of England and Wales presents a substantial opportunity for academics to test criminological theories, or to examine how crime patterns vary over space and time. It also presents substantial opportunities for practitioners engaged in crime reduction more directly, as data sharing protocols often impede access to crime data – even for crime reduction agencies. This is particularly the case for the analysis of crime problems that cross geographical/jurisdictional borders. Consequently, the data may offer solutions to practical issues that agencies face on a day-to-day basis, as well as facilitating research that informs our understanding of crime. Of course, the data also provide the public with the opportunity to examine online maps that show the locations of crime events. However, to protect the anonymity of victims, the police.uk data have necessarily been obfuscated with a geomasking technique, which influences their spatial accuracy. The aim of this paper was to examine the extent to which this is the case and to determine at what level(s) of spatial resolution – if any – the police.uk data are suitable for analysis (formal or otherwise).

Comparing the police.uk data to the police recorded crime data used to generate it, as expected (Hypothesis 1), we find that for large areal units, such as MSOAs the spatial accuracy of the data is very good. The same is true for the slightly smaller LSOA unit of geography, particularly for the crimes of burglary, vehicle crime and criminal damage. For smaller geographical units, particularly at the postcode level, however, it becomes clear that there is considerable spatial error in the data.

This has important implications for the types of theories that researchers might plausibly test using police.uk data. The social processes deemed relevant to some ecological theories are hypothesized to function at a micro-level of place (see Weisburd, Bernasco and Bruinsma

2009), and so, for these theories, the police.uk data may have limited application. However for those theories which assume social forces exert their influence at the neighborhood level (e.g., social disorganization theory), the use of the police.uk data may be entirely appropriate meaning that, for the first time, researchers can test such theories using data for an entire country.

The inaccuracy prevalent at smaller levels of geography, such as postcodes, also affects the interpretation of crime patterns by users of the police.uk website. The negative media coverage in the weeks following the launch of the site indicates that many users do not take the time to acquaint themselves with the way in which the data are obfuscated – despite this information being available and accessible. There is thus considerable potential for crime patterns at the local level to be misinterpreted by users of the police.uk website.

However, even at the small area level, all may not be lost. Our analyses suggest that there are systematic biases associated with the way the police.uk data are generated. The identification of such biases is useful insofar as it presents the opportunity to model and correct for these in any analysis of the data. The analyses presented above illustrate one way in which such modeling might be achieved. Alternative methods of modeling (spatial) spill-over effects or to correct for spatial autocorrelation produced by unobserved heterogeneity (e.g., see Anselin 1988) can similarly assess the effect of the "jitter" procedure applied to the police.uk data. Unfortunately, producing modeled estimates of crime counts in the way illustrated in this paper will only be possible for those with access to the original police recorded crime data. And, those with access to such data would presumably have no use for the police.uk data. However, even those without access to such data can include independent variables in their statistical models to attempt to correct for the biases in the police.uk data identified here.

Moreover, if the estimated parameters reported here prove to be applicable to other geographical areas, it may be possible to produce a "general" model that could be used more universally. Whether this is feasible is an empirical question, and it is important to point out that the analyses reported here are for a sample of data. While they are generally consistent with a-priori expectations, it is possible that different patterns would be observed for other police force areas. Consequently, replication studies will be important to establish the generalizability of our findings.

At this point, it is important to revisit the issue of statistical aggregation to areal units discussed in the introduction. Aside from the points already considered, one issue associated with the analysis of data aggregated to some geography is that the boundaries employed are not created with crime analysis in mind; they are generally artificially constructed for political or administrative

purposes. For example, census areas may be constructed so as to produce neighborhoods with homogeneous socio-demographic characteristics, but they often fail to reflect the sociospatial distribution of land use, people and crime events that will be of interest to crime analysts (Rengert and Lockwood 2009). This issue is above and beyond those associated with the spatial accuracy of the data analyzed.

Regardless of such issues, in the current paper we examined the spatial accuracy of the data at the areal unit level. The primary reason for this was, as discussed in the introduction, that crime data are frequently analyzed in this way for a variety of purposes. However, future studies could examine the spatial accuracy of the data at the street segment level. Street segments represent a fine spatial scale – similar in size to the postcode geography examined here – but given that the "snap point" database was essentially generated at the street segment level, it is possible that the police.uk data are more accurate at this unit of analysis than at the postcode level. Future studies might also disaggregate the data by time as well space. The police.uk data do not indicate the day or time on which offenses take place but they do indicate the month of occurrence and so it would be useful to know if the estimates are accurate on finer timescales than one-year.

In conclusion, the provision of open access data is increasing, with attendant benefits to the academic and other communities. While such data may be imperfect, based on the analyses presented here, the police.uk data appear to hold considerable promise as long as they are analyzed at a suitable geographical resolution.

Disclosure statement

The authors are not aware of any conflicts of interest that are related to this work.

Notes

1. For example West Midlands, West Yorkshire, Surrey, Sussex, Devon and Cornwall and the Metropolitan Police.
2. However this is also because crime records are subject to coding changes after they have been initially recorded. For example, some incidents originally recorded may not be classified as crimes once the Police have investigated the circumstances. In addition, the data processing requirements on Forces would be very consuming if they had to provide data in "real time" or at lower resolution than a month.
3. The "snap point" masterlist was created in the first instance by taking the center point of every road in England and Wales from the Ordnance Survey Locator dataset. Then points of local relevance from the PointX dataset (such as transportation hubs and large retail premises) were added. These data were then subjected to analysis which used Voronoi polygons to determine how many postal addresses were contained in the catchment area for each snap point. Any snap point which had fewer than eight addresses associated with it were discarded to protect the privacy of victims. Once the snap point masterlist had been created by the police.uk website developers it was passed to Police Forces for a human assessment. Over time a number of snap points were added to the masterlist based on this feedback.
4. It is unclear from the information on the police.uk website whether these eight postal addresses have to be residential. For the purposes of this study we are assuming that non-residential postal addresses are not included in the snap point list, as these are not as sensitive to the identification of victims as residential postal addresses.
5. For example from December 2010 until August 2011 there were six categories (ASB, burglary, robbery, vehicle crime, violent/sexual crime and other crime). From September 2011 until April 2013 there were 11 categories (the existing ones plus criminal damage/arson, drugs, other theft, public disorder/weapons and shoplifting). Since May 2013 there have been 14 categories (the existing categories, minus public/disorder and weapons (which was split into two), plus bicycle theft and theft from the person).
6. The median was computed for the area size due to the presence of extreme outliers in the data for that variable.
7. This was achieved by joining crime events to their nearest postcode centroid (which fell within the West Midlands Police area), and then using a lookup table – provided by the Office for National Statistics – to generate the Census geography information. This resulted in a handful of postcodes on the boundary of the study area being associated with Census geography units that fell outside the study area, and hence the loss of a small number of crime data points. This way of deriving the areal unit information was considered preferable to spatial joins available in ArcGIS and the R statistical programming software which double-counted crime events which fell on boundary lines of areal units.
8. Areas were classed as being adjacent to one another if their boundaries touched at any point (i.e., "queen" contiguity). This was done in ArcMap 10.1 using the spatial join feature to identify adjacent areas (by joining the data to *itself*). Further information is available from the authors on request.
9. Available from http://www.neighbourhood.statistics.gov.uk/
10. For example, we could compute the root mean squared error (or simply the absolute difference) for the difference between the police.uk and police recorded crime counts.

However, the same pattern of results emerge using this method, and so we discuss this method no further (findings are available upon request).

11. For the larger units of analysis, the trends were consistent with those reported but the coefficients associated with the census and land use data were largely non-significant.

12. Due to the data being count data, we also used a negative binomial regression model to estimate parameters. The results, which were largely the same, are available upon request.

References

Andresen, M. A. 2009. "Testing for Similarity in Area-Based Spatial Patterns: A Nonparametric Monte Carlo Approach." *Applied Geography* 29 (3): 333–345.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Anselin, L., and A. Getis. 1992. "Spatial Statistical Analysis and Geographic Information Systems." *The Annals of Regional Science* 26 (1): 19–33.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman. 1999. "Geographically Masking Health Data to Preserve Confidentiality." *Statistics in Medicine* 18: 497–525.

Bailey, T. C., and A. C. Gattrell. 1995. *Interactive Spatial Data Analysis*. London: Longman.

Bernasco, W., and R. Block. 2011. "Robberies in Chicago: A Block-Level Analysis of the Influence of Crime Generators, Crime Attractors, and Offender Anchor Points." *Journal of Research in Crime and Delinquency* 48: 33–57.

Bowers, K. J., S. D. Johnson, and K. Pease. 2004. "Prospective Hot-Spotting the Future of Crime Mapping?" *British Journal of Criminology* 44 (5): 641–658.

Brantingham, P. J., and P. L. Brantingham, eds. 1991. *Environmental Criminology*. 2nd. Prospect Heights, IL: Waveland Press.

Brantingham, P. J., D. A. Dyreson, and P. L. Brantingham. 1976. "Crime Seen through a Cone of Resolution." *American Behavioral Scientist* 20 (2): 261–273.

Bridwell, S. A. 2007. "The Dimensions of Locational Privacy." *Societies and Cities in the Age of Instant Access* 88: 209–225.

Cassa, C. A., S. J. Grannis, M. Overhage, and K. D. Mandl. 2006. "A Context-sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection." *Journal of the American Medical Informatics Association* 13 (2): 160–165.

Chainey, S., and J. Ratcliffe. 2005. *GIS and Crime Mapping*. London: Wiley.

Chainey, S., and L. Tompson. 2012. "Engagement, Empowerment and Transparency: Publishing Crime Statistics Using Online Crime Mapping." *Policing: A Journal of Policy and Practice* 6 (3): 228–239.

Cornish, D., and R. V. Clarke, eds. 1986. *The Reasoning Criminal*. New York: Springer-Verlag.

Daily Mail. 2011a. "It's Crimebotch UK: Outcry as Leafy Streets Labelled Crime Hotspots on New Home Office Website (Which Crashed as soon as It Went Live)." Accessed June 23. http://www.dailymail.co.uk/news/article-1352354/Police-crime-map-Home-Office-website-crashed-2-centres-hotspots.html.

Daily Mail. 2011b. "Police Deny Crime Figures Were 'Airbrushed' following Manchester Riots in the Summer." Accessed June 23. http://www.dailymail.co.uk/news/article-2045547/UK-riots-Police-accused-fiddling-crime-figures-counting-Manchester-riots-single-offence.html#ixzz1cMejjn9S.

Ecclestone, A. 1998. "Freedom of Information & the Internet: An Electronic Window on to Government. The Campaign for Freedom of Information." Accessed June 23. http://www.cfoi.org.uk/pdf/cyberwindow.pdf.

Graham, C. 2012. *Anonymisation: Managing Data Protection Risk Code of Practice*. Wilmslow: Information Commissioner's Office.

Groff, E. R., D. Weisburd, and N. Morris. 2009. "Where the Action Is at Places: Examining Spatio-Temporal Patterns of Juvenile Crime at Places Using Trajectory Analysis and GIS." In *Putting Crime in Its Place: Units of Analysis in Spatial Crime Research*, edited by D. Weisburd, W. Bernasco, and G. Bruinsma, 61–86. New York: Springer.

Groff, E. R., D. Weisburd, and S.-M. Yang. 2010. "Is it Important to Examine Crime Trends at a Local "Micro" Level?: A Longitudinal Analysis of Block to Block Variability in Crime Trajectories." *Journal of Quantitative Criminology* 26: 7–32.

Hampton, K. H., M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. A. Serre, and W. C. Miller. 2010. "Mapping Health Data: Improved Privacy Protection With Donut Method Geomasking." *American Journal of Epidemiology* 172 (9): 1062–1069.

Hipp, J. R. 2007. "Block, Tract, and Level of Aggregation: Neighbourhood Structure and Crime and Disorder as a Case in Point." *American Sociological Review* 72: 659–680.

Hohl, K., B. Bradford, and E. Stanko. 2010. "Influencing Trust and Confidence in the London Metropolitan Police: Results from an Experiment Testing the Effect of Leaflet Drops on Public Opinion." *British Journal of Criminology* 50 (3): 491–513.

Home Office. 2008. *From the Neighbourhood to the National: Policing our Communities Together*. London: Home Office.

Home Office. 2010. *Policing in the 21st Century: Reconnecting Police and the People*. London: Home Office.

ICO. 2012. *Crime-Mapping and Geo-Spatial Crime Data: Privacy and Transparency Principles*. Wilmslow: Information Commissioner's Office.

Johnson, S. D., and K. J. Bowers. 2010. "Permeability and Crime Risk: Are Cul-De-Sacs Safer?" *Journal of Quantitative Criminology* 26 (1): 89–111.

Kounadi, O., K. Bowers, and M. Leitner. 2014. "Crime Mapping On-line: Public Perception of Privacy Issues." *European Journal of Criminal Policy and Research*. doi:10.1007/s10610-014-9248-4.

Kubrin, C., and R. Weitzer. 2003. "New Directions in Social Disorganization Theory." *Journal of Research in Crime and Delinquency* 40 (4): 374–402.

Kwan, M., I. Casas, and B. C. Schmitz. 2004. "Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks?" *Cartographica* 39 (2): 15–28.

Leitner, M., and A. Curtis. 2004. "Cartographic Guidelines for Geographically Masking the Locations of Confidential Point Data." *Cartographic Perspectives* 49: 22–39.

Openshaw, S. 1984. *The Modifiable Areal Unit Problem*. Norwich: GeoBooks.

Ouimet, M. 2000. "Aggregation Bias in Ecological Research: How Social Disorganization and Criminal Opportunities Shape the Spatial Distribution of Juvenile Delinquency in Montreal." *Canadian Journal of Criminology* 42: 135–156.

Police.uk. 2014. "Anonymisation Methodology." Accessed June 23. http://data.police.uk/about/#anonymisation.

Polvi, N., T. Looman, C. Humphries, and K. Pease. 1991. "The Time Course of Repeat Burglary Victimisation." *British Journal of Criminology* 31: 411–414.

Ratcliffe, J. H. 2001. "Policing Urban Burglary." Trends and Issues in Crime and Criminal Justice. Paper, 213. Canberra: Australian Institute of Criminology.

Rengert, G. F., and B. Lockwood. 2009. "Geographical Units of Analysis and the Analysis of Crime." In *Putting Crime in Its Place: Units of Analysis in Spatial Crime Research*, edited by D. Weisburd, W. Bernasco, and G. Bruinsma, 3–33. New York: Springer.

Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15: 351–357.

Sampson, R. J., and W. B. Groves. 1989. "Community Structure and Crime: Testing Social-Disorganization Theory." *American Journal of Sociology* 94: 774–802.

Sampson, R. J., S. W. Raudenbush, and F. Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy." *Science* 277: 918–924.

Sherman, L. W., P. R. Gartin, and M. E. Buerger. 1989. "Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place." *Criminology* 27: 27–56.

Smith, W. R., S. G. Frazee, and E. L. Davison. 2000. "Furthering the Integration of Routine Activity and Social Disorganization Theories: Small Units of Analysis and the Study of Street Robbery as a Diffusion Process." *Criminology* 38: 489–524.

Wieland, S. C., C. A. Cassa, K. D. Mandl, and B. Berger. 2008. "Revealing the Spatial Distribution of a Disease while Preserving Privacy." *Proceedings of the National Academy of Sciences of the United States of America* 105 (46): 17608–17613.

Weisburd, D., G. Bruinsma, and W. Bernasco. 2009. "Units of Analysis in Geographic Criminology: Historical Development, Critical Issues, and Open Questions." In *Putting Crime in Its Place: Units of Analysis in Spatial Crime Research*, edited by D. Weisburd, W. Bernasco, and G. Bruinsma, 3–33. New York: Springer.

Weisburd, D. L., S. Bushway, C. Lum, and S.-M. Yang. 2004. "Trajectories of Crime at Places: A Longitudinal Study of Street Segments in the City of Seattle." *Criminology* 42 (2): 283–321.

Wikström, P.-O. 2006. "Individuals, Settings, and Acts of Crime: Situational Mechanisms and the Explanation of Crime." In *The Explanation of Crime: Context, Mechanisms, and Development*, edited by P.-O. Wikström and R. J. Sampson. Cambridge: Cambridge University Press.

Wooldredge, J. 2002. "Examining the (Ir)Relevance of Aggregation Bias for Multilevel Studies of Neighbourhoods and Crime with an Example Comparing Census Tracts to Official Neighbourhoods in Cincinnati." *Criminology* 40 (3): 681–709.