Spatial Information Divergence: Using Global and Local Indices to Compare Geographical Masks Applied to Crime Data

Ourania Kounadi* and Michael Leitner†

*University of Salzburg
†Louisiana State University, Baton Rouge

Abstract
Advances in Geographic Information Science (GISc) and the increasing availability of location data have facilitated the dissemination of crime data and the abundance of crime mapping websites. However, data holders acknowledge that when releasing sensitive crime data there is a risk of compromising the victims' privacy. Hence, protection methodologies are primarily applied to the data to ensure that individual privacy is not violated. This article addresses one group of location protection methodologies, namely geographical masks that are applicable for crime data representations. The purpose is to identify which mask is the most appropriate for crime incident visualizations. A global divergence index (GDi) and a local divergence index (LDi) are developed to compare the effects that these masks have on the original crime point pattern. The indices calculate how dissimilar the spatial information of the masked data is from the spatial information of the original data in regards to the information obtained via spatial crime analysis. The results of the analysis show that the variable radius mask and the donut geomask should be primarily used for crime representations as they produce less spatial information divergence of the original crime point pattern than the alternative local random rotation mask and circular mask.

1 Background
Geoprivacy is the privacy of our location, for instance where we are right now, where we live, or where we perform our activities. According to Beresford and Stajano (2003) location privacy is “the ability to prevent other parties from learning ones current or past location”. Similarly, Duckham and Kulik (2006) describe location privacy, “as a special type of information privacy which concerns the claim of individuals to determine for themselves when, how, and to what extent location information about them is communicated to others”. Researchers have explored geoprivacy by mainly investigating two aspects: (1) the private information disclosure when location data are released; and (2) the development of location protection methodologies to avoid information disclosure. The first aspect reveals the problem, whereas the second aspect gives a solution to the problem.

1.1 Private Information Disclosure
When location data are released there is a risk of disclosing private information about those involved in the dataset. Social media location applications allow users to share their current...
location among their friends. But if this information is available to everyone without the
people’s knowledge, then their location has been compromised. A striking example occurred in
2008 in Japan when teachers unintentionally disclosed personal information about 980 stu-
dents. They used the Google My Maps application to geo-reference the students’ residential
addresses, names, and telephone numbers. The map’s access was set to public but when the
teachers tried to delete them, the data were still available. That happened because Google
stores My Maps’ information on two or more different servers. Even though information was
removed from one server the data still remained on another (Burdon 2010). Another potential
violation source are applications in location-based services. Smart phones enable the installa-
tion of applications from third parties, namely providers of geolocation applications, which
can then retrieve and process the location data from mobile phones (Article 29 2011). Krumm
(2007) showed that location data can be observed repeatedly to estimate home addresses and
also associate critical locations with identities. Furthermore, thematic maps are commonly
used to present confidential, sensitive, or private information like a distribution of crime inci-
dents. If the map’s locations are the victims’ residences then sensitive information about them
has been compromised. For instance, location protection techniques that are employed by
crime mapping websites may fail to protect victims’ privacy. Burdon (2010) identified a rape
case at the “SpotCrime” website (http://spotcrime.com/) showing clearly a small apartment
block highlighted by Google Street view. Even though the respective crime mapping website
aggregates the attribute address information, the exact location of attack, at least for this inci-
dent, was shown on the map.

1.2 Location Protection Methodologies

To avoid the risk of private geo-information disclosure scholars have proposed and developed
a plethora of location protection methodologies. A comprehensive summary and description of
these methodologies can be found in Bridwell (2007), Cottrill (2011), and Gambs et al.
(2010). Some of these methods were developed for location-based services applications. Data
from location detection devices (cellular phones, global positioning system (GPS), and radio-
frequency identification (RFID)) are subject to re-identification of movement patterns of indi-
vidual users. To prevent such re-identification, You et al. (2007) proposed two schemes that
alter the original movement by creating dummy trajectories. With respect to publishing indi-
vidual trajectories, Monreale et al. (2010) proposed a generalization method that takes into
account the original’s data clustering analysis results. According to Cohen et al. (2001), aggre-
gating high volumes of pervasive data to refined views is necessary for the needs of such appli-
cations but also to address privacy concerns regarding tracking people’s locations. Also,
“spatial and temporal cloaking” adapts the spatial and temporal resolution to achieve a differ-
ent level of anonymity constraints (Gruteser et al. 2003). However, in some cases the method-
ologies seem to overlap. For instance the “location cloaking” described by Cheng et al. (2006)
employs the “variable radius geographical mask” presented by Kwan et al. (2004).

Other methodologies modify the identifiers of the data (locations or attributes). For
instance, the “mix zone” method creates areas where user movements cannot be traced
because a user’s identity is mixed with all other users’ identities in that area (Beresford and
Stajano 2004). Another method is to retrieve data under pseudonyms which may also change
periodically so as to avoid re-identification using locations of unique identifiers (Buttyán et al.
2007). Furthermore, Puttaswamy and Zhao (2010) proposed an approach for location-based
social applications where the encryption key is managed by users to decrypt location data
stored on servers.
In some of these methods the user plays an integral role in deciding the degree of information disclosure. Myles et al. (2003) proposed a mechanism for location-based applications where users are able to control the distribution of their information. A similar approach by Hull et al. (2004) is a framework that integrates data from a variety of sources, gathers user privacy preferences, and finally builds an infrastructure to impose preferences for sharing these data. Furthermore, user preferences have been taken under consideration by Duckham and Kulik (2005) as a negotiation mechanism between users and location-based service providers. User preferences are used to define the degree of degrading the quality of spatial information (obfuscation). Last, in “fuzzy inference systems” rule bases are built upon user’s needs to define the spatial precision on five levels. These levels are specialized by location-based services and user groups (Hashemi and Malek 2012).

Furthermore, protection methodologies have been used for thematic location data. If location data contain confidential, sensitive, or private information about individuals, then disclosing those locations can compromise personal privacy. With regards to health related georeferenced data, Boulos et al. (2006) proposed a controlled access to confidential data where a software agent acts on behalf of the researcher and returns only aggregated, useful, and accurate results. Finally, geographical masks are developed to protect static and non-temporal confidential location data such as point map representations. Geographical masks introduce uncertainty to the dataset by altering original locations. Point distribution maps of confidential or sensitive themes, such as health and crime related information, are commonly used in scientific publications. Furthermore, authors of such publications seem to prefer geographical masks in order to protect their location data (LaBeaud et al. 2008, Zhou et al. 2012, Curtis and Mills 2011). Similar methods are being used in national crime mapping websites, such as the “anonymization methodology” (http://data.police.uk/about/#location-anonymisation) of the UK based “Police.uk” website and the “block generalization” (http://www.crimemapping.com/help/faq.aspx#whatis) of the US based “crimemapping” website. In this article, we investigate this type of protection methodologies because it is best applicable for crime incidents representations, like point maps.

1.3 Study Design

Our objective is to calculate how dissimilar the spatial information of the tested (masked) data is from the spatial information of the reference (original) data. First we explore all available geographical masks in the literature (Section 2). We analyze the masks’ differences, which is a necessary step before proceeding to a fair comparison. Then we present detailed information on the datasets acquisition and preprocessing steps (Section 3). Following the material description, we then demonstrate published effects’ detection techniques (Section 4). The effects’ detection techniques were used to compare the spatial analytical results or the spatial information dimensions of the masked and the original data. This section, along with spatial point pattern analysis methods that are commonly used in crime analysis, provided us with the essential background to develop two indices, the global divergence index (GDi) and the local divergence index (LDi). These two indices describe the spatial information divergence of a masked point pattern with an original point pattern (Section 5).

2 Selection of Geographical Masks

Different geographical masks groups and their categories are presented in Table 1. Armstrong et al. (1999) were the first to introduce the term “geographical masks” when they added the
| Masking group | Category | No | Description | Masking Degree |
|---------------|----------|----|-------------|----------------|
| Aggregation | Point Aggregation | 1 | a new symbolic point represents the location of several points | variable |
| | Areal Aggregation | 2 | a new symbolic area represents the location of several points | variable |
| | Street aggregation at the midpoint | 3 | point aggregation at the midpoints of the streets | variable |
| | Street aggregation at the intersection | 4 | point aggregation at the intersection of the streets | variable |
| Affine transformation | Translation | 5 | displacement of points by a fixed increment | constant |
| | Scale change | 6 | displacement of points by a scaling constant | constant |
| | Rotation | 7 | displacement of points by a fixed angle about a pivot point | constant |
| | Concatenation of isomasks | 8 | combination of translation, scale, and rotation | constant |
| | Stochastic mechanism in the affine transformation | 9 | random selection of the increment, scaling constant, or angle in a predefined range | constant |
| Random Perturbation | Circular mask | 10 | each point is randomly rotated on a circle with a fixed radius | constant |
| | Variable radius mask | 11 | each point is randomly rotated within a circle of variable radius | constant |
| | Weighted mask | 12 | each point is randomly rotated on a circle with a variable radius $r/\text{r}$ is population density dependent | variable |
| | Local random rotation | 13 | randomly rotating the centre of each grid cell | constant |
| | Local random translation | 14 | each point is displaced by a random increment within the grid cell | constant |
| | Donut geomasking | 15 | each point is randomly displaced by at least a minimum distance and by less than a maximum distance in a torus | variable |
| Flipping | Global horizontal flipping | 16 | flipping the horizontal central axis of the map | constant |
| | Global vertical flipping | 17 | flipping the vertical central axis of the map | constant |
| | Global axes flipping | 18 | flipping the horizontal and the vertical central axes of the map | constant |
| | Local random flipping | 19 | flipping the horizontal, vertical, or both axes of each grid cell randomly | constant |
| Neighbour Information | Nearest neighbour distance mask | 20 | provides the distance information of each point to the two nearest neighbours | not applicable |
| Ordered Nearest neighbours | 21 | mimics approximately the relative locations of the points | not applicable |
| Other | Population-density-based Gaussian spatial blurring | 22 | each point is displaced randomly based on a normal distribution whose standard deviations are inversely correlated to the surrounding population density | variable |
| Linear programming de-identification | 23 | original locations make up set A. Masked locations make up set B. Using a linear programming model the confidential locations of A are moved to some locations of B. | variable |
| Matrix masking based on spatial smoothing techniques | 24 | the form and the degree of masking are variable depending on the spatial pattern and the disclosure risk, thus resulting in a utility-risk profile | variable |
| Contextual Information | 25 | replaces geographic identifiers with contextual information | not applicable |
geographic dimension to previous masking methods (Duncan and Pearson 1991, Cox 1994),
which have originally been proposed for tabular data. The methods proposed by Armstrong
et al. (1999) (Table 1, No. 1, 2, 5–9, 21, 22, 25 and the general concept of random pertur-
bation) are not used in this study because they introduce general functions, which can poten-
tially form other masks when specific parameters are applied to them. These masking
concepts were later used and evolved by other scientists. For instance, Kwan et al. (2004)
used random perturbation and the circle as a masking area (Table 1, No. 10–12). Similarly,
the donut geomasking method (Table 1, No. 15) proposed by Hampton et al. (2010) is a
random perturbation technique, where the displacement is restricted by a minimum distance
so as to ensure privacy protection. Also, Leitner and Curtis (2004, 2006) used random per-
turbation, rotation, aggregation, and flipping techniques applied at a global or local level
(Table 1, No. 3, 4, 13, 14, 16–19). Furthermore, Cassa et al. (2006) (Table 1, No. 22),
Wieland et al. (2008) (Table 1, No. 23), and Kwan et al. (2004) (Table 1, No. 12) incorpo-
rated in their techniques the underlying population density as an additional parameter. Last,
the flexible model of Zhou et al. (2010) minimized the effects of the mask on particular
types of analysis, such as loglinear and logistic regression analysis (Table 1, No. 23).

In the majority of the masks, exact locations lie within an area that the mask’s
developer has predefined. The size of that area is called the “uncertainty area” and defines
the disclosure degree of the locations. For instance, if the original location is randomly
distributed over a block of 2.5 km² and 100 people reside in this block, the disclosure degree
will be 1%. The masking degree (last column in Table 1) is the parameter
that defines the uncertainty area. For instance, the masking degree for the circular
mask is defined by its radius. A radius of 10 m (masking degree is 10 m) will result
in an uncertainty area of 314 m². The neighbor information and contextual
information techniques of Table 1 either provide spatial information about the original
points or remove geographic identifiers from the dataset; therefore, they do not have
the masking degree parameter. Apart from these three all other masks do have a masking
degree.

The masking degree can be constant or variable. A constant masking degree means
that the size of the uncertainty area is equal for all points, whereas a variable masking
degree means that the size of the uncertainty area varies depending on a specific factor.
In most cases, this factor is the underlying population density. For example, if a point is
randomly displayed inside a square where there are no residents, then the mask’s developer
will have to increase the square size up to a level of the minimum number of residents
required to protect privacy. From the geographical masks presented in Table 1 only a
selection of masks with constant masking degrees are examined. Applying the same
constant masking degree to all masks ensures that all points are transferred to equally-sized
uncertainty areas. Making the masks’ uncertainty areas the same size is a necessary step
to comparing the effects that the masks have on the original pattern. From all methods
with a constant masking degree we selected those that had less effect on the spatial charac-
teristics of the original patterns than others, as reported in the respective literature.
Furthermore, we also included the donut geomasking method. Even though donut
geomasking was published with a variable masking degree, its methodology can be easily
applied with a fixed maximum distance, thus transforming the variable masking degree to a
constant masking degree. Finally, the geographical masks tested here are: the circular and
variable radius masks proposed by Kwan et al. (2004), the local random rotation proposed
by Leitner and Curtis (2006), and the donut geomasking proposed by Hampton et al.
(2010).
3 Reference and Masked Datasets

The original dataset consisted of burglaries in apartments in Vienna for 2007 and was pro-
vided by the Criminal Intelligence Service, Austria. We extracted a 500 point random sample
from the original dataset (10,854 incidents). To extract 500 points, we assigned random
numbers from a uniform distribution to each point, re-ordered the points in an increasing
order based on their random numbers and finally selected the first 500 point. This sample is
called the “original sample dataset” for the remainder of this article. However, solely one
dataset would not fulfill the scope of this study and therefore we created more datasets. The
advantage of several artificial datasets is that the comparison of masks is based on average
analytical results from different visualizations. Hence, the results are not biased towards the
analytical results of only one spatial point pattern, but are formed by several patterns. To
create more datasets, the characteristics of the original point pattern were identified and
datasets that simulate some of these characteristics were created. In particular, the original
sample dataset was the starting point from which we created 30 different visualizations, all of
which maintain the major spatial characteristic of crime phenomena, namely that they are
more clustered than randomly distributed (Weisburd et al. 2004, Sherman et al. 1989, Eck
et al. 2000).

First, we calculated the first-order nearest neighbour index ($\text{NNI} = 0.579133$) and the
one standard deviational ellipse (347 points inside the ellipse and 153 points fall outside) of
the original sample dataset. When one standard deviational ellipse contains approximately
68 percent of all input features then the features have a spatially normal distribution. This
means that they are densest in the center (the ellipse is centered on the mean center for all
features) and become less dense toward the periphery (Mitchell 2005). This also describes
the clustering pattern of the original sample dataset since 69.4 percent of all incidences
(347) were enclosed by the ellipse. To create reference datasets we assumed that if another
dataset contains 347 clustered points and 153 random points its NNI would be similar to
the original dataset.

We then created one dataset of 153 random points and 30 datasets of 347 clustered
points. Regarding the random dataset only one is required since the pattern for any addi-
tional random dataset would yield similar spatial analytical results. For clustered points we
manually created irregular polygons inside the study area that varied in size and location.
Each time we selected a different count and combination of these polygons and randomly
displayed the remaining 347 points, which acted as injected clusters. For each dataset
the clustered points (347) were merged with the random points’ dataset (153) thus result-
ing in 30 datasets. Each time a new dataset was created the NNI of it was calculated to
ensure that the point pattern of the artificial dataset was clustered at a similar degree to
the original one. From now on these are called “reference datasets”. The number of 30
“reference datasets” was selected as a balance to minimize the analytical computa-
tional time that is presented next (Sections 5 and 6) and, at the same time, to provide a
meaningful number of different visualizations. The reference datasets differ from the original
sample dataset in terms of the points’ locations, but simulate some of their spatial character-
istics. Hence, all reference datasets are located in the same area (Vienna), are 500 points
each, their first-order NNI is similar to the original sample dataset (range: 0.503–0.656),
and their mean NNI (0.594) is not statistically significantly different from the original
NNI (NNI=0.579, one sample T-test: $t = 2.325$, $p > 0.01$). Both the original as well as the
30 reference data sets are statistically significantly different from a random distribution at
$p < 0.001$. 

© 2014 The Authors. Transactions in GIS published by John Wiley & Sons Ltd Transactions in GIS, 2015, 19(5)
For each reference dataset we applied the four geographical masks (circular mask, variable radius mask, local random rotation, and donut geomasking). In total 120 masked datasets (four masks times 30 reference datasets) are compared with the 30 reference datasets. For the masking degree we selected an uncertainty area of 184,000 m², which means a cell size of 428.952 m for the local random rotation and a radius of 242.01 m for the other masks. The uncertainty area was defined as the midpoint between uncertainty areas that were reported in previous studies. Leitner and Curtis (2006) proposed a cell size smaller than 350 × 350 m (uncertainty area = 122,500 m²) for accurate results, whereas Kwan et al. (2004) argued that a masked pattern is not perceived differently from an original pattern for a radius of 915 ft (r = 278.892 m, uncertainty area = 244,355 m²). Also, for the minimum radius of donut geomasking we use the ratio “minimum radius = 0.1 * maximum radius” (24.201 m), as has been proposed by Allshouse et al. (2010).

4 Effects’ Detection Techniques

Previous studies investigated the masks’ effects on the spatial distribution of original points. Armstrong et al. (1999) compared the spatial information dimensions of the original point pattern with those of the masked point pattern. The information dimensions were grouped as: (1) pair-wise relations; (2) event-geography relations; (3) trends; and (4) anisotropies. The first group reveals whether a mask preserves the actual distances, the relative distances, the actual orientations, and the relative orientations for each pair of original and masked points. The second group measures distances and orientations between masked points and other geographic features of a study area compared with original measurements. For instance, the distance from a police station to a masked crime incident compared to the distance to an original crime incident. The last two groups show if a mask preserves the existence and directionality of trends and anisotropies (i.e. the change of properties in different directions). The trend of the point pattern was also examined by Kwan et al. (2004). Additionally, the same authors calculated the maximum difference in density. This is the maximum difference of the density of masked incidents per square mile and the density of original incidents per square mile. Lastly, Leitner and Curtis (2004) performed a comparison through visual observations of the points’ distribution similarity. The study’s participants ranked perceived similarities between pairs of masked-original point patterns. The results of these observations compared the effects of various masks.

Another effect of masks is the variations between hotspots derived from masked and original incident locations. This was an additional information dimension by Armstrong et al. (1999), where they examined the existence, the actual locations, and the relative locations of masked compared to original clusters. Similarly, Kwan et al. (2004) calculated the clustering distance using the cross K function analysis. The same authors also created two and three dimensional density surfaces so as to calculate the number and actual locations of the surfaces’ peaks. Leitner and Curtis (2004) compared the similarity of observed hotspots, drawn by participants, which were derived from masked and original points. Last, the clustering analysis results of original and masked points were examined extensively using the spatial scan statistic to detect the clusters’ sensitivity, specificity, detection rate, accuracy, and most significant cluster (Cassa et al. 2006, Olson et al. 2006, Wieland et al. 2008, Hampton et al. 2010).
5 Spatial Information Divergence

5.1 Method

Some of the aforementioned authors described or calculated the effects that geographical masks have on the distribution of original points. More importantly, in all previously mentioned studies the effects of various masks on original hotspots were examined. That is because clustering detection is an important aspect of spatial analysis. In crime analysis it is used to detect high crime concentrations and allocate police resources accordingly, or it can be used as a predictive tool for future crimes (Chainey et al. 2008, Levine 2008). In this section we present our methodology to estimate the Spatial Information Divergence of the masked compared to the original points. This is calculated by means of the Global Divergence Index (GDi) and the Local Divergence Index (LDi). For the demonstration of both indices we used one of the reference datasets (Section 3) and applied a circular mask with a masking degree of 1,000 m. This is an arbitrary number, which is used only in this section, in order to provide a more “extreme” and therefore clearer visualization of both divergences. The GDi represents the divergence from the centrographic analytical results, whereas the LDi represents the divergence from the hotspots analytical results. The GDi utilizes statistics such as the spatial mean that measures the spatial center of the point pattern and the one standard deviation ellipse that measures the spread or variability of the point pattern (Clarke and Eck 2005, Eck et al. 2005, Chainey and Ratcliffe 2005). On the other hand the LDi utilizes point based methods for clustering detection (spatial ellipses and LISA (local indicators of spatial autocorrelation)). In particular, we used the nearest neighbour hierarchical clustering (Nnh) (Everett 1974, D’andrade 1978) and the Getis-Ord Gi* statistic (Getis and Ord 1996), which are two of the most widely used methods in spatial crime analysis (Garth 2006, Ratcliffe and McCullagh 1999, Murray et al. 2001, Chainey et al. 2002, Leipnik and Albert 2003, Levine 1999, 2006).

The GDi (Equation 1) is an equally weighted composite indicator of three single indices, namely the Mdi (the index of the divergence from the spatial mean, the geographic centre of concentration), the Odi (the index of the divergence from the one standard deviational ellipses’ orientations), and the MAdi (the index of the divergence from the one standard deviational ellipses’ major axes). The GDi combines three dimensions of point patterns’ centrographic characteristics so as to provide the overall global divergence.

\[
GDi := (Mdi, Odi, MAdi)
\]

The LDi (equation 2) is also an equally weighted composite indicator of two single indices, namely the Nnh.di, the index of the hotspot areas’ divergence using nearest-neighbour hierarchical spatial clustering and the Gi*.di, the index of the hotspot areas’ divergence using the Getis-Ord Gi* statistic. Here, the two single indices calculate the same dimension (the point pattern’s local characteristics) by calculating the divergence of the hotspot areas. However, since the indices use different algorithms, they may produce slightly different hotspot areas. Hence, the LDi is the average of these two indices in order to balance their differences.

\[
LDi := (Nnh.di, Gi*.di)
\]
the reference dataset. What is “the most dissimilar” varies among the indices and will be discussed in detail in the next sub-section. Previous techniques applied spatial analysis to the masked and to the original data and interpreted variations of the results. Here, the spatial analytical results are not being interpreted but are incorporated within the indices. This has several advantages. For example, if a “translate affine mask” is applied to two original datasets, according to Armstrong et al. (1999) both masked datasets will preserve the relative locations of the clusters. However, the Gi*di can also indicate which original dataset will be less distorted than the other. With regards to the quantitative effects’ detection techniques (Section 4), results of different spatial statistics cannot be compared to each other. For instance, it is difficult to conclude whether the clustering distance or the maximum difference in density (Kwan et al. 2004) is more affected by a mask because the results are in different measurement units. In addition, results of the same spatial statistics can be compared with each other (e.g. clustering sensitivity for different geographical masks) but they cannot interpret the magnitude of the distortion as it can be interpreted with the GDi and LDi. Last, this method utilizes spatial statistics that are widely used in crime analysis and therefore it has a focus on future masked crime representations.

5.2 Calculation of Single Indices

5.2.1 Mean’s Divergence (Mdi)

The most dissimilar masked mean from the original mean is the one that is still located inside the study area, but is as far away as possible from the original mean. Hence, the divergence of the masked mean from the original mean is a function of the distance from the original mean to the farthest edge of the study area (Figure 1, Equation 3).

\[ M_{di} = \frac{\text{Distance of the original mean to the masked mean}}{\text{Distance of the original mean to the farthest point away, but still in the study area}} \times 100 \]  

(3)

5.2.2 Orientation’s Divergence (Odi)

The most dissimilar ellipse’s orientation is the one that has the opposite direction from the original ellipse’s orientation (if \( \theta \) is the corner from the vertical axis to the major axis of an ellipse then \( -\theta \) is the opposite one). Hence, the divergence of the masked ellipse’s orientation to the original ellipse’s orientation is a function of their rotation compared to the maximum rotation (Figure 2, Equation 4).

\[ O_{di} = \left| \frac{\text{Orientation of the original ellipse} - \text{orientation of the masked ellipse}}{180} \right| \times 100 \]  

(4)

5.2.3 Major Axis’ Divergence (MAdi)

The metric for this index is the length of the major axis of a one-standard deviational ellipse. Therefore, we must first determine the most dissimilar major axis from the original major axis in terms of length. The minimum length of an ellipse’s major axis will be close to zero, but cannot be zero. For example, if a point dataset consists of points with the same coordinates an ellipse cannot be created because at least two points with different X and Y coordinates are needed to create the length and the width of an ellipse. If those points are very close together, but not exactly on top of each other, then the ellipse’s major axis will be very close to zero.
Therefore, for practical purposes we assume that the minimum length of an ellipse’s major axis is zero. For the maximum length of an ellipse’s major axis one must calculate the length of the major axis of a point dataset where all points are located on the boundary of the study area, furthest away from each other (½ of all points on one edge and the other ½ of all points on the other edge). The maximum length thus changes according to the size and shape of the study area. For the study area used in this research, the maximum length of the major axis is 21,785.86 m. Hence, the divergence of the masked ellipse’s major axis to the original ellipse’s major axis is a function of their difference in lengths compared with the biggest length difference (Figure 3, Equations 5a–b).

\[
M_{Adi} = \frac{|\text{Masked.MA} - \text{Original.MA}|}{\text{Maximum.MA} - \text{Original.MA}} \times 100, \quad \text{if } \text{Original.MA} \leq \frac{\text{Maximum.MA}}{2} \tag{5a}
\]

\[
M_{Adi} = \frac{|\text{Masked.MA} - \text{Original.MA}|}{\text{Original.MA}} \times 100, \quad \text{if } \text{Original.MA} > \frac{\text{Maximum.MA}}{2} \tag{5b}
\]

where Masked.MA = the length of the masked ellipse’s major axis; Original.MA = the length of the original ellipse’s major axis; and Maximum.MA = the maximum length of an ellipse’s major axis in the study area.

Equation (5a) considers as the biggest length difference the difference of the maximum major axis from the original major axis and equation (5b) considers as the biggest length difference the difference of the original major axis from the minimum major axis (=0).
5.2.4 Nearest-Neighbour Hierarchical Spatial Clustering Divergence (Nnh.di) and Getis-Ord Gi* Statistic Divergence (Gi*.di)

To determine the Nnh.di the original hotspots and the masked hotspots are defined as set A and B, respectively. Then the symmetric difference (S.D.), the set of elements between two sets that do not overlap, is calculated. If the symmetric difference is equal to 0, then the masked hotspots are identical to the reference hotspots and the index is also equal to 0. The closer the set of the symmetric difference is to the set \( A + B \) the bigger the index will be. The index will be equal to 100 when the two sets are completely disjoined. Calculations are similarly performed for the Gi*.di, and the formulas to calculate both indices are:

\[
\text{Nnh}.di = \frac{\text{Symmetric difference of } A \text{ and } B}{A + B} \times 100
\]

\[
\text{Gi}^*.di = \frac{\text{Symmetric difference of } A \text{ and } C}{A + C} \times 100
\]

where A = the original hotspots; B = the masked hotspots using Nearest-Neighbour Hierarchical Spatial Clustering (first order clusters); and C = the masked hotspots using the Gi* statistic (significant hotspots: Z-score > 1.65).

A limitation of these formulas (Equations 6 and 7) is that if case sets A and B are disjoined the index will always have a maximum value of 100, regardless of the distance between the...
disjoined sets. However, if one pair of sets is disjoined at a larger degree than a second pair of sets, then the global divergence index will vary as well thus revealing these differences. To conclude, for hotspots’ indices the most dissimilar hotspot areas are the ones that share no common parts with the original hotspot areas. Figures 4 and 5 show the final maps of the Nnh.di and the Gi*.di.

6 Results

The results of the GDi and the LDi are shown in Table 2. Both indicators show the same trend among the respective masks. The variable radius mask has the lowest effect on the reference data (GDi = 0.038 and LDi = 25.587), followed by the donut geomask (GDi = 0.042 and LDi = 27.071). The local random rotation mask has the greatest global information loss (GDi = 0.062) and the circular mask has the greatest local information loss (LDi = 36.992) among all masks tested. The Kruskall-Wallis test was performed (Harper and Ryan 2001) to compare the GDi and the LDi results across the masks (Table 3). The Kruskall-Wallis test is a non-parametric test that compares independent samples (Kruskal and Wallis 1952). The result for the GDi shows that the various masks are statistically significantly different (H = 14.47, p < 0.01) from each other. Then, the Mann-Whitney U test was applied as a post-hoc test for pairwise comparisons, since this non-parametric test compares only two independent samples (Mann and Whitney 1947). Pair-wise comparisons indicate that results of the local random rotation mask are significantly different compared to all other masks. Hence, the local random
Figure 4  The divergence of the masked hotspots to the original hotspots using Nearest-Neighbour Hierarchical Spatial Clustering

Figure 5  The divergence of the masked hotspots to the original hotspots using the Gi* statistic
rotation mask leads to the greatest spatial information divergence of the reference dataset’s
global characteristics. Similarly, the Kruskall-Wallis test for the LDi results shows that masks
are also significantly different ($H = 57.39$, $p < 0.01$) from each other. According to the Mann-
Witney pair-wise comparisons, both circular and local random rotation masks lead to greater
spatial information divergence than the donut and the variable radius masks. According to the

| masked datasets | circular mask | local random rotation mask | donut geomask | variable radius mask |
|-----------------|--------------|---------------------------|---------------|---------------------|
|                 | $GD_i$       | $LD_i$                    | $GD_i$        | $LD_i$              | $GD_i$       | $LD_i$ |
| 1               | 0.072        | 40.638                    | 0.064         | 37.871              | 0.048        | 30.377 |
| 2               | 0.058        | 39.722                    | 0.083         | 39.164              | 0.036        | 30.801 |
| 3               | 0.070        | 48.794                    | 0.074         | 38.950              | 0.035        | 29.202 |
| 4               | 0.044        | 40.473                    | 0.139         | 37.637              | 0.021        | 28.030 |
| 5               | 0.051        | 48.107                    | 0.076         | 34.516              | 0.061        | 29.018 |
| 6               | 0.041        | 38.617                    | 0.034         | 38.275              | 0.015        | 28.215 |
| 7               | 0.066        | 39.216                    | 0.037         | 35.996              | 0.060        | 27.707 |
| 8               | 0.031        | 38.468                    | 0.076         | 36.143              | 0.042        | 29.187 |
| 9               | 0.021        | 25.634                    | 0.048         | 24.131              | 0.038        | 19.956 |
| 10              | 0.022        | 43.347                    | 0.049         | 37.635              | 0.025        | 33.320 |
| 11              | 0.067        | 41.104                    | 0.064         | 38.809              | 0.039        | 29.070 |
| 12              | 0.044        | 38.847                    | 0.074         | 36.740              | 0.055        | 29.352 |
| 13              | 0.045        | 33.191                    | 0.036         | 29.392              | 0.030        | 23.395 |
| 14              | 0.026        | 30.196                    | 0.032         | 25.760              | 0.043        | 23.771 |
| 15              | 0.044        | 41.013                    | 0.063         | 35.949              | 0.036        | 30.278 |
| 16              | 0.066        | 36.603                    | 0.015         | 34.114              | 0.022        | 29.070 |
| 17              | 0.060        | 34.128                    | 0.079         | 33.936              | 0.059        | 26.816 |
| 18              | 0.028        | 36.436                    | 0.039         | 36.210              | 0.020        | 28.769 |
| 19              | 0.036        | 32.044                    | 0.097         | 28.344              | 0.071        | 21.689 |
| 20              | 0.030        | 25.730                    | 0.068         | 25.338              | 0.027        | 20.981 |
| 21              | 0.024        | 20.714                    | 0.073         | 21.463              | 0.024        | 16.834 |
| 22              | 0.032        | 32.049                    | 0.011         | 30.995              | 0.014        | 25.983 |
| 23              | 0.051        | 32.364                    | 0.013         | 31.659              | 0.073        | 27.847 |
| 24              | 0.024        | 37.068                    | 0.074         | 33.690              | 0.015        | 26.391 |
| 25              | 0.029        | 30.340                    | 0.064         | 27.142              | 0.064        | 22.862 |
| 26              | 0.078        | 40.370                    | 0.162         | 39.299              | 0.084        | 30.523 |
| 27              | 0.046        | 37.766                    | 0.041         | 34.445              | 0.052        | 25.254 |
| 28              | 0.062        | 36.107                    | 0.081         | 33.591              | 0.037        | 25.558 |
| 29              | 0.036        | 39.376                    | 0.047         | 38.022              | 0.038        | 29.935 |
| 30              | 0.043        | 51.301                    | 0.050         | 39.161              | 0.070        | 31.941 |
| Range           | 0.057        | 30.586                    | 0.151         | 17.836              | 0.069        | 16.486 |
| SD              | 0.017        | 6.722                     | 0.033         | 5.057               | 0.019        | 3.810  |
| Mean            | 0.045        | 36.992                    | 0.062         | 33.813              | 0.042        | 27.071 |
tests’ results it can be concluded that by applying the donut mask or the variable radius mask the masked data would have less spatial information divergence than applying the other two masks.

This method reveals spatial variations that can be interpreted at a scale of none to a maximum divergence (0–100). As an additional statistical measure to validate our method, we calculated root mean squared errors (RMSE) of the dataset’s spatial characteristics. The masked spatial characteristics are used as observed values and the original spatial characteristics as expected values. This measure was also used for the assessment of varying parameters for the donut geomask by Clifton and Gehrke (2013). Here, the RMSE has been calculated for all global and local characteristics as follows: (1) spatial mean difference as the distance of the original mean to the masked mean in meters; (2) ellipse’s orientation difference as their rotation in degrees; (3) long axis difference as their length difference in meters; (4) Getis-Ord Gi* difference as their symmetrical difference in km²; and (5) Nearest-Neighbour Hierarchical Spatial Clustering difference as their symmetrical difference in km². Table 4 shows results of the RMSEs as well as results of all single indices. Only the comparison of the results is viable because the indices have no units and possess values from 0–100, although the RMSEs have the same units of the properties that they measure and of course they do not have range limits. For this reason, the smaller the values of RMSEs and divergence indices are, the better the performance of a mask is. As it was anticipated, the masks’ performance is the same for both divergence indices and RMSEs. The results of the local characteristics show that the circular mask performs better in terms of the spatial mean (RMSE: 3.2 m, Mdi: 0.017), the variable mask performs better in terms of the ellipses orientation (RMSE: 0.078 degrees, Odi: 0.033), and the donut geomask performs better in terms of the long axis (RMSE: 5.454 m, Ldi: 0.031). On the other hand, one trend is observed for the local characteristics. Regardless of the hotspot technique (Gi* or Nnh), the variable mask outperforms the others. Then, the performance is decreasing in the following order: the donut geomask, the local random rotation mask, and finally, the circular mask.

The trend observed in local results can be justified considering the methodology of each mask. In Figure 6 100 possible destinations (masked locations) for a single original location

| Gdi                          | Ldi                          |
|------------------------------|------------------------------|
| Kruswal-wallis test: H = 14.47, p = 0.002 | Kruswal-wallis test: H = 57.39, p < 0.001 |
| Mann-Whitney pairwise comparisons, 95% confidence interval, local random rotation mask | Mann-Whitney pairwise comparisons, 95% confidence interval, circular mask |
| Mann-Whitney pairwise comparisons, 95% confidence interval, local random rotation mask | Mann-Whitney pairwise comparisons, 95% confidence interval, local random rotation mask |
| Circular mask | 0.012 | – | 0.026 |
| Local random rotation mask | – | 0.156 | – |
| Donut geomask | 0.034 | <0.001 | <0.001 |
| Variable mask | 0.004 | <0.001 | <0.001 |
are shown for each mask. All masks of this example have an uncertainty area of approximately 39,000 m² (111.4 m radius and 197.5 m grid size). However, the possible destinations of the masked locations vary remarkably. The local random rotation mask (purple points in Figure 6) rotates the point around the centre of its grid cell. Therefore the point will be located along the circumference of a circle where its centre is the centre of the grid cell and its radius is the distance between the centre and the original location. The circular mask (blue points in Figure 6) will locate the masked location along the circumference of the circle whereas the variable radius mask (green points in Figure 6) may locate the masked point anywhere within the circle. Similarly, the donut geomask (yellow points in Figure 6) will locate the masked point anywhere within the circle except for the predefined small circle area near the original point. Looking at the last three masks, it is visually clear that the variable radius mask is more

Table 4 Performance of geographical masks by: (a) average results of single indices; and (b) RMSE of spatial characteristics

| Mask Type               | Circular Mask | Local Random Rotation Mask | Donut Geomask | Variable Mask |
|-------------------------|---------------|-----------------------------|---------------|---------------|
| Global characteristics  | Md_i          | 0.017                       | 0.033         | 0.018         | 0.041         |
|                         | mean RMSE     | 3.200                       | 6.812         | 3.370         | 7.460         |
|                         | Odi           | 0.061                       | 0.082         | 0.076         | 0.033         |
|                         | orientation RMSE | 0.138                    | 0.207         | 0.159         | 0.078         |
|                         | Ldi           | 0.057                       | 0.072         | 0.031         | 0.041         |
|                         | long axis RMSE | 8.994                      | 12.608        | 5.454         | 7.126         |
| Local characteristics   | Gi*.di        | 33.596                      | 28.024        | 21.948        | 20.539        |
|                         | Gi* RMSE      | 24.816                      | 15.834        | 12.384        | 11.627        |
|                         | Nnh.di        | 40.388                      | 39.601        | 32.194        | 30.636        |
|                         | Nnh RMSE      | 18.879                      | 18.188        | 14.705        | 13.909        |

Figure 6 Potential destinations of the original location by geographical mask
likely to have the least effect on the original location. However, the analysis that was carried out above is required in order to quantify and more importantly to differentiate the masks’ effect results in a statistically significant manner.

Also, the results show that the spatial information divergence of global characteristics is much smaller than the divergence of local characteristics \((GDi = 0.038 – 0.062 \text{ and } LDi = 25.587 – 36.992)\). In fact, for these datasets, geographical masks and masking degrees, the global information divergence is small, which means that the masked data do not notably alter the global characteristics of the original data. However, the considerable divergence of local characteristics is expected due to the fact that all masks act at a local level (each point is located somewhere near its original location). Nevertheless, the GDi is important, despite the fact that the index is rather low in the above analysis. The main reason is the map’s scale, which can be seen in Equations (2) and (4). For example, if the scale was larger (smaller study area), the reference mean’s distance to the farthest point in the study area would have been smaller, thus resulting in a bigger Mdi.

Additionally, there is a noteworthy variance across datasets that is particularly apparent for LDi results \((\text{Range} = 16.603–30.586 \text{ and } \text{SD} = 4.007 – 6.722)\). To evaluate the significance of the spread of results a Friedman test, which compares distributions across repeated measures (Friedman 1937), was utilized. The test reveals whether there are any reference datasets that consistently produce higher or lower local information divergence across the four geographical masks. The test rejected the null hypothesis that the masks’ distributions are the same, which further implies that the reference datasets are statistically significantly different from each other \((Q = 84.68, p < 0.01)\). The reference datasets have an equal number of points and similar (but not the same) clustering degrees. To examine if the clustering differences among the datasets affect the LDi results, the first-order NNI of each reference dataset was compared with the masks’ average LDi. Figure 7 shows a fitted linear model with an increasing trend between the average LDi and the NNI for each dataset. The Pearson product-moment correlation shows a statistically significant positive correlation between the two variables \((r = 0.572, n = 30, p = 0.001)\). This means that the distribution of the original points affects the spatial information divergence and the more clustered the original data are, the higher the local divergence will be. The fact that the distribution of reference points has an effect on the spatial information divergence of masked points has not been discussed previously in the literature. This finding should be taken into consideration when future researchers apply geographical masks before presenting their data. For instance, Kwan et al. (2004) propose an uncertainty area about the average size of census block groups in their study area.
as the size that achieves an optimum balance between privacy protection and accuracy of results. There are at least two factors that led to this proposed area in Kwan et al. (2004). Firstly, the underlying population density and secondly, the point pattern itself. Hence, the choice of the uncertainty area should not be employed without closer scrutiny but rather assessed individually on each occasion.

7 Concluding Remarks and Future Directions

The comparison of the geographical masks revealed that the variable radius mask has the least effect on the distribution of the original point pattern. However, it does not produce significantly different results from the donut geomask. Both resulted in statistically significantly less spatial information divergence than the local random rotation and the circular mask. Also for the masking degrees applied in this analysis, the global spatial characteristics of the reference datasets changed little. This reveals that at a global level the masked data do not significantly alter the original data. On the other hand, a notable change was observed for the local spatial characteristics, which implies that geographical masks affect the original data, mostly at a local level. Lastly, the results varied remarkably across the datasets revealing that for the same masking degree, the more clustered the data are, the bigger the spatial information divergence will be.

The indices that have been used to compare geographical masks could be further exploited for other GIS analytical scenarios. The authors can think of a total of four application areas, including the one presented in this research, where the GDi and the LDi could be used. The first application area is when an intended error is introduced into the original data using a geographical mask. This is done mostly for privacy preservation. In this application area, the spatial information divergence is the accuracy loss of the masked point dataset due to the intentional displacement of the original points. A second possible application area is related to the unintended error produced by the geocoding process. This error can be the result of the match rate (if during the geocoding process some points fail to be geocoded) where the spatial information divergence is the accuracy loss produced by a smaller point dataset compared to the original one. Additionally, the error can be the result of the positional error of the geocoded points. In this scenario, spatial information divergence is the accuracy loss produced when point locations are being displaced due to the positional error compared to original locations. The last application area is related to the description and quantification of point patterns’ variations over time. In this case spatial information divergence is not an accuracy loss but rather a change in data quality over time due to natural or human factors. An example of spatial change over time due to human factors might be the phenomenon of urban sprawl. Also, the expansion of invasive exotic plants is another example of spatial change due to natural factors.

Previously used effect detection techniques proposed threshold values for masks’ masking degrees that balance privacy with accuracy. However, these masking degrees are dependent on characteristics of their study areas such as the underlying population and the examined datasets. Consequently, searching for an acceptable threshold value for spatial accuracy loss with different geographical masks is still a research gap that needs further investigation. In this article we employed spatial divergence indices to compare the performance of different geographical masks. As a next step or direction for future research these indices or similar ones should be further explored so as to identify a value within their ranges from which the original data seem to alter significantly. Furthermore, the proposed threshold values should be in
compliance with acceptable visualisation methods for scientific research, information dissemination to the public, etc.

Note

All functions in Section 3 were performed in ArcGIS 10.0 (Spatial Statistics Tools), except for the Nnh, which was performed in CrimeStat 3.3 (Hot Spot Analysis I).

References

Allshouse W B, Fitch M K, Hampton K H, Gesink D C, Doherty I A, Leone P A, Serre M L, and Miller W C 2010 Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. Geocarto International 25: 443–52

Armstrong M P, Rushton G, and Zimmerman D L 1999 Geographically masking health data to preserve confidentiality. Statistics in Medicine 18: 497–525

Article 29 2011 ‘Opinion 13/2011 on Geolocation Services on Smart Mobile Devices’. Article 29 Data Protection Working Party, 881/11/EN WP 185, 16 May 2011 (available at http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp185_en.pdf)

Beresford A R and Stajano F 2003 Location privacy in pervasive computing. IEEE CS and IEEE Communications Society 1: 46–55

Beresford A R and Stajano F 2004 Mix zones: User privacy in location-aware services. In Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, Washington, DC

Boulos M N K, Cai Q, Padget J A, and Rushton G 2006 Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. Journal of Biomedical Informatics 39: 160–70

Bridwell S A 2007 The dimensions of locational privacy. In Miller H J (ed) Societies and Cities in the Age of Instant Access. Berlin, Springer: 209–25

Burdon M 2010 Privacy invasive geo-mashups: Privacy 2.0 and the limits of first generation information privacy laws. University of Illinois Journal of Law, Technology and Policy 1: 1–50

Buttyán L, Holczer T, and Vajda I 2007 On the effectiveness of changing pseudonyms to provide location privacy in VANETs. In Molva R, Tsudik G, and Westhoff D (eds) Security and Privacy in Ad-hoc and Sensor Networks. Berlin, Springer-Verlag Lecture Notes in Computer Science Vol. 3813: 129–41

Cassa C A, Grannis S J, Overhage J M, and Mandl K D 2006 A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. Journal of the American Medical Informatics Association 13: 160–65

Chainey S and Ratcliffe J 2005 GIS and Crime Mapping. London, John Wiley and Sons

Chainey S, Reid S, Stuart N 2002 When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime. In Kidner D, Higgs G, and White S (eds) Socio-economic Applications of Geographic Information Science. London, Taylor and Francis: 21–36

Chainey S, Tompson L, and Uhlig S 2008 The utility of hotspot mapping for predicting spatial patterns of crime. Security Journal 21: 4–28

Cheng R, Zhang Y, Bertino E, and Prabhakar S 2006 Preserving user location privacy in mobile data management infrastructures. In Danesiz G and Golle P (eds) Privacy Enhancing Technologies. Berlin, Springer-Verlag Lecture Notes in Computer Science Vol. 4258: 393–412

Clarke R V and Eck J E 2005 Crime Analysis for Problem Solvers: In 60 Small Steps. Washington, DC, Office of Community Oriented Policing

Clifton K J and Gehrke S R 2013 Application of geographic perturbation methods to residential locations in the Oregon Household Activity Survey. Transportation Research Record 2354: 40–50

Cohen N H, Purakayastha A, Turek J, Wong L, and Yeh D 2001 Challenges in Flexible Aggregation of Pervasive Data. Redmond, WA, IBM Research Division 1

Cottrill C D 2011 Location privacy: Who protects? URISA Journal 23(2): 49–61

Cox L H 1994 Matrix masking methods for disclosure limitation in microdata. Survey Methodology 20: 165–69

Curtis A and Mills J W 2011 Crime in urban post-disaster environments: A methodological framework from New Orleans. Urban Geography 32: 488–510
D’andrade R 1978 U-statistic hierarchical clustering. *Psychometrika* 4: 58–67

Duckham M and Kulik L 2003 Simulation of obfuscation and negotiation for location privacy. In Cohn A G and Mark D M (eds) *Spatial Information Theory*. Berlin, Springer-Verlag Lecture Notes in Computer Science Vol. 3693: 31–48

Duckham M and Kulik L 2006 Location privacy and location-aware computing. In Drummond J, Billen R, Forrest D, and Joao E (eds) *Dynamic and Mobile GIS: Investigating Change in Space and Time*. Boca Raton, FL, CRC Press: 35–51

Duncan G T and Pearson R W 1991 Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science* 6: 219–39

Eck J E, Chainey S, Cameron J G, Leitner M, and Wilson R E 2005 *Mapping Crime: Understanding Hot Spots*. Washington, DC, National Institute of Justice

Eck J E, Gersh J S, and Taylor C 2000 Finding crime hot spots through repeat address mapping. In Goldsmith V, McGuire P,ollenkopf J H, and Ross T A (eds) *Analyzing Crime Patterns: Frontiers of Practice*. Thousand Oaks, CA, Sage Publications: 49–64

Everett B 1974 *Cluster Analysis*. London, Heinemann

Friedman M 1937 The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675–701

Gambs S, Killijian M O, and del Prado Cortez M N 2010 Show me how you move and I will tell you who you are. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on Privacy and Security in GIS and LBS*, San Jose, California: 34–41

Garth D 2006 *Crime, Neighborhood, and Public Housing*. New York, LFB Scholarly Publishing

Getis A and Ord J K 1996 Local spatial statistics: An overview. In Longley P and Batty M (eds) *Spatial Analysis: Modelling in a GIS Environment*. Cambridge, UK, Geoinformation International.: 261–77

Gruteser M and Grunwald D 2003 Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the First International Conference on Mobile Systems, Applications and Services*, Ambleside, United Kingdom: 31–42

Hampton K H, Fitch M K, Allshouse W B, Doherty I A, Gesink D C, Leone P A, Serre M L, and Miller W C 2010 Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology* 172: 1062–69

Harper D and Ryan P 2001 PAST: Paleontological statistics software package for education and data analysis. *Palaeontology Electronica* 4: 1–9

Hashemi M and Malek M R 2012 Protecting location privacy in mobile geoservices using fuzzy inference systems. *Computers, Environment and Urban Systems* 36: 311–20

Hull R, Kumar B, Lieuwen D, Patel-Schneider P F, Sahuguet A, Varadarajan S, and Vyas A 2004 Enabling context-aware and privacy-conscious user data sharing. In *Proceedings of the Fifth IEEE International Conference on Mobile Data Management*, Berkeley, California: 187–98

Krumm J 2007 Inference attacks on location tracks. *Pervasive Computing* 6: 127–43

Kruskal W H and Wallis W A 1952 Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47: 583–621

Kwan M P, Casas I, and Schmitz B C 2004 Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica* 39(2): 13–28

LaBeaud A D, Gorman A-M, Koonce J, Kippes C, McLeod J, Lynch J, Gallagher T, King C H, and Mandalakas A M 2008 Rapid GIS-based profiling of West Nile virus transmission: Defining environmental factors associated with an urban-suburban outbreak in Northeast Ohio, USA. *Geospatial Health* 2: 215

Leipnik M R and Albert D P 2003 *GIS in Law Enforcement: Implementation Issues and Case Studies*. London, Taylor and Francis

Leitner M and Curtis A 2004 Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives* 49: 22–39

Leitner M and Curtis A 2006 A first step towards a framework for presenting the location of confidential point data on maps: Results of an empirical perceptual study. *International Journal of Geographical Information Science* 20: 813–22

Levine N 1999 CrimeStat: A spatial statistics program for the analysis of crime incident locations. In *Proceedings of the Fourth International Conference on Geocomputation*, Fredericksburg, Virginia

Levine N 2006 Crime mapping and the CrimeStat program. *Geographical Analysis* 38: 41–56

Levine N 2008 The “hottest” part of a hotspot: Comment on “The utility of hotspot mapping for predicting spatial patterns of crime”. *Security Journal* 21: 295–302

Mann H B and Whitney D R 1947 On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18: 50–60

Mitchell A 2005 *The ESRI Guide to GIS Analysis*, Volume 2. Redlands, CA, Esri Press

© 2014 The Authors. Transactions in GIS published by John Wiley & Sons Ltd
Monreale A, Andrienko G, Andrienko N, Giannotti F, Pedreschi D, Rinzivillo S, and Wrobel S 2010 Movement data anonymity through generalization. *Transactions on Data Privacy* 3: 91–121
Murray A T, McGuffog I, Western J S, and Mullins P 2001 Exploratory spatial data analysis techniques for examining urban crime implications for evaluating treatment. *British Journal of Criminology* 41: 309–29
Myles G, Friday A, and Davies N 2003 Preserving privacy in environments with location-based applications. *Pervasive Computing* 2: 56–64
Olson K L, Grannis S J, and Mandl K D 2006 Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health* 96: 2002–08
Puttaswamy K P N and Zhao B Y 2010 Preserving privacy in location-based mobile social applications. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems and Applications*, Annapolis, Maryland
Ratcliffe J H and McCullagh M J 1999 Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems* 1: 385–98
Sherman L W, Gartin P, and Buerger M E 1989 Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology* 27: 27–55
Wieland S C, Cassa C A, Mandl K D, and Berger B 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: 17608–13
Weisburd D, Bushway S, Lum C, and Yang S M 2004 Trajectories of crime at places: A longitudinal study of street segments in the city of Seattle. *Criminology* 42: 283–322
You T H, Peng W C, and Lee W C 2007 Protecting moving trajectories with dummies. In *Proceedings of the Eighth International Conference on Mobile Data Management*, Mannheim, Germany: 278–82
Zhou Y J, Dominici F, and Louis T A 2010 A smoothing approach for masking spatial data. *Annals of Applied Statistics* 4: 1451–75
Zhou J, Lurie M N, Bärnighausen T, McGarvey S T, Newell M-L, and Tanser F 2012 Determinants and spatial patterns of adult overweight and hypertension in a high HIV prevalence rural South African population. *Health and Place* 18: 1300–06