Computers, Environment and Urban Systems 57 (2016) 59–67

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets

Ourania Kounadi [a,*], Michael Leitner [a,b]

[a] Doctoral College GIscience, University of Salzburg, Department of Geoinformatics – Z_GIS, Schillerstraße 30, 5020 Salzburg, Austria
[b] Department of Geography and Anthropology, Louisiana State University, E-104 Howe-Russell-Kniffen Geoscience Complex, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 10 March 2015
Received in revised form 12 January 2016
Accepted 16 January 2016
Available online 4 February 2016

Keywords:
Geographical masks
Confidentiality
k-anonymity
Location privacy
Crime data
Spatial error

ABSTRACT

Geographical masking is the conventional solution to protect the privacy of individuals involved in confidential spatial point datasets. The masking process displaces confidential locations to protect individual privacy while maintaining a fine level of spatial resolution. The adaptive form of this process aims to further minimize the displacement error by taking into account the underlying population density. We describe an alternative adaptive geomasking method, referred to as Adaptive Areal Elimination (AAE). AAE creates areas of a minimum K-anonymity and then original points are either randomly perturbed within the areas or aggregated to the median centers of the areas. In addition to the masked points, K-anonymized areas can be safely disclosed as well without increasing the risk of re-identification. Using a burglary dataset from Vienna, AAE is compared with an existing adaptive geographical mask, the donut mask. The masking methods are evaluated for preserving a predefined K-anonymity and the spatial characteristics of the original points. The spatial characteristics are assessed with four measures of spatial error: displaced distance, correlation coefficient of density surfaces, hotspots' divergence, and clusters' specificity. Masked points from point aggregation of AAE have the highest spatial error in all the measures but the displaced distance. In contrast, masked points from the donut mask are displaced the least, preserve the original spatial clusters better, have the highest clusters' specificity and correlation coefficient of density surfaces. However, when the donut mask is adapted to achieve an actual K-anonymity, the random perturbation of AAE introduces less spatial error than the donut mask for all the measures of spatial error.

1. Introduction

The growing tendency of releasing data on the Internet, technological developments, and open data policies may be in conflict with the individuals' right to information privacy (Kulk & Van Loenen, 2012). For example, revealing users' locations through Geo-social networks (GeoSNs) (e.g. Twitter) may cause privacy threats and requires privacy-protection methods (Vicente, Freni, Bettini, & Jensen, 2011). Geo-social networks reveal current or past locations of individuals, location based services (LBS) collect spatial trajectories of users, and Web-services map confidential locations of health or crime related information. As the availability and use of confidential and sensitive data increases, new methods are being developed to allow the dissemination of information while protecting the privacy of individuals.

This paper deals with confidential discrete location data. This type of spatial datasets appear in scientific publications as thematic point maps (Kounadi & Leitner, 2014) and as dynamic maps on the internet to disseminate information to the public (police.uk, 2015). Furthermore, confidential location data are released for research studies (dhsprogramm, 2015). Hampton et al. (2010) argue that health-related thematic point maps are used extensively because they allow the assessment of spatial heterogeneity in one area, they lead decisions regarding policy prevention and intervention programs, and provide information to allocate resources. Nevertheless, these datasets have to be protected and comply with current regulations and restrictions on the individuals' right to privacy. For example, the Health Insurance Portability and Accountability Act (HIPAA) requires the de-identification of geographic units that do not meet specific population thresholds (Nass, Levit, & Gostin, 2009). Also, several reports provide guidelines on the issues of sharing or publishing spatial crime data (Graham, 2012; ICO, 2012; Wartell & McEwen, 2001). However, general regulations on privacy, such as acts and directives, do not address the particulars of the spatial re-identification risk. Cottrill (2011) reveals the absence of location specific directives in four international privacy regulations for Europe, Australia, Canada, and Japan.

Despite the absence of location specific directives, research findings revealed that published locations on maps can be re-engineered back to identify their actual locations (Brownstein, Cassa, Kohane, & Mandl,

* Corresponding author.
E-mail addresses: ourania.kounadi@sbg.ac.at (O. Kounadi), mleitne@lsu.edu (M. Leitner).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.01.004

2006; Cassa, Wieland, & Mandl, 2008; Leitner, Mills, & Curtis, 2007). Furthermore, re-engineered locations from maps can be linked with other sources to collect additional private information (Kounadi, Lampoltshammer, Leitner, & Heistracher, 2013; Krumm, 2007). Hence, a plethora of location protection methods has been developed by the scientific community to prevent the spatial re-identification risk that arises with such types of disclosures. This paper continues on the existing literature of location protection methods and proposes an adaptive geographical masking technique that is best applicable for confidential discrete location data. The method allows masked confidential points to be disclosed, yet the actual locations cannot be de-identified among k-cases even after a successful reengineering process.

The following sections discuss the current literature on the location protection methods for discrete point data (Section 2) and the adaptive geographical masking techniques that this study is built upon (Section 3). Next, we present an alternative protection method, which is called "Adaptive Areal Elimination" (Section 4), and by using a real-world dataset we apply and evaluate its effectiveness (Section 5).

2. Location protection methods for discrete point data

Masking such as record transforming, attribute transforming, or displacing are disclosure limitation techniques that were initially created to protect the confidentiality of tabular and non-spatial databases (Duncan & Pearson, 1991). These techniques were later on developed to include individual level location data (Armstrong, Rushton, & Zimmerman, 1999). Armstrong et al. (1999) introduced the term "geographical masking", which was further adopted by other scholars to develop several location protection methods for discrete point data. Geographical masking is based on concepts of imprecision (lack of specificity in information) and inaccuracy (lack of correspondence between information and reality). An example of imprecision is spatial aggregation where confidential locations are aggregated into areal units. Methods based on inaccuracy introduce an error to the original locations by displacing them within a study area (Kwan, Casas, & Schmitz, 2004; Leitner & Curtis, 2004, 2006). Moreover, these types of geographical masking techniques (also referred to as geographical isomasks) seem to be preferred by scientists when confidential data are visualized in scientific publications (Kounadi & Leitner, 2014).

Recent geographical isomasks adapted the displacement error by taking into account the underlying population density (Cassa, Grannis, Overhage, & Mandl, 2006; Gruteser & Grunwald, 2003; Hampton et al., 2010; Wieland, Cassa, Mandl, & Berger, 2008). The advantage of the adaptive geomasking is that it enables "maskers" to define a desirable level of K-anonymity. A K-anonymity protection means that each person or record in a masked dataset cannot be distinguished from at least k-1 persons or records whose information also appears in the dataset (Sweeney, 2002). For spatial datasets K-anonymity ensures that each location (e.g. locations of individuals, households, addresses) cannot be distinguished from at least k-1 locations.

3. Adaptive geographical masking: the current approach

Geographical isomasks ensure privacy protection by displacing original points within uncertainty areas produced by the masks. In particular, uncertainty area is the area where a masked point may lie within (e.g. circle or torus). For instance, "donut geomasking" randomly displaces the coordinate of each data point within a torus based on a uniform distribution (Hampton et al., 2010). "Population-density-based Gaussian spatial blurring" displaces a point within a circle in a direction and distance that is randomly selected from a normal distribution (Cassa et al., 2006). To illustrate how K-anonymity is achieved with adaptive geographical masking we use a simpler approach. In this example a point is displaced within a circle in a direction and distance that is randomly selected from a uniform distribution (adaptive variable radius mask). Fig. 1a shows a hypothetical area and the locations of ten households (A to K). Point C and E are also household locations of a confidential dataset that need to be masked. To simplify the example, locations are plotted on a squared grid, where the distance from node to node equals to 1 unit and K-anonymity equals to three households. Fig. 1b shows the uncertainty areas of confidential locations C and E. Point C can be displaced anywhere within the small circle of a 2 units radius, whereas point E can be displaced anywhere within the larger circle of a 5 units radius. The size of the radius has been adjusted so as to contain at least three households including the households of points C and E.

A limitation of current methods is the resolution that was used for the disclosure information (e.g. population). For instance, Hampton et al. (2010) and Cassa et al. (2006) used population information within administrative boundaries such as census block groups. Hence, in order to calculate the displacement error they relied on the assumption of a homogeneously distributed population within the areas. This approach can lead to masked points having smaller actual k-anonymity (*Kact*) than estimated k-anonymity (*Kest*), especially in areas with high population distribution heterogeneity (Allshouse et al., 2010).

As illustrated in Fig. 1 an accurate K-anonymity can be achieved if disclosure information is available at a point level (e.g. locations of households). However, there is still a re-engineering possibility associated with current methods, which may lead to the *Kact* be less than what was originally aimed for (*Kest*). Fig. 1c shows the masked locations of C and E (green dots). If the masking method and the K-anonymity is

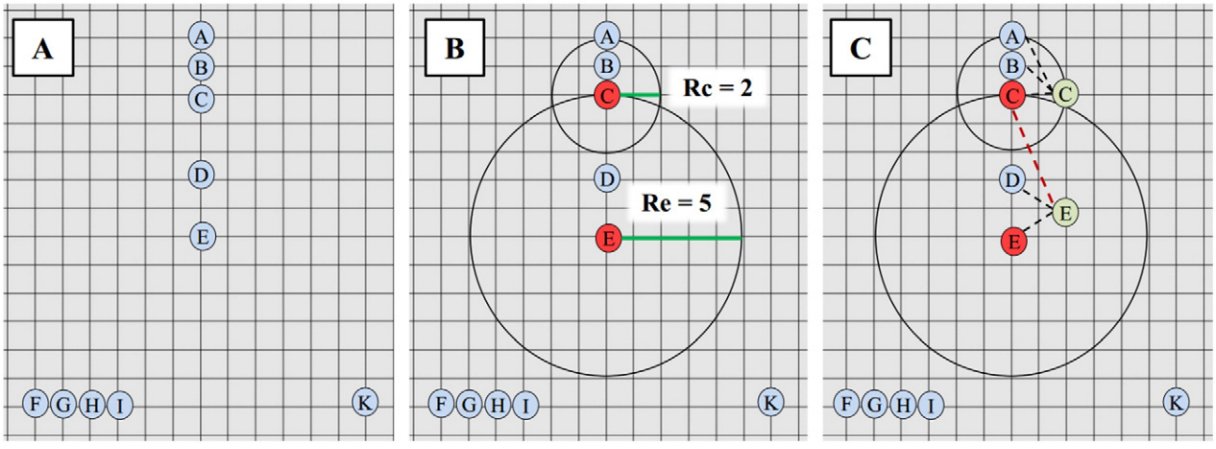

Fig. 1. A) hypothetical area and locations of households, B) uncertainty areas of confidential locations C and E based on a K-anonymity equal to three, and C) original locations of C and E represented by red dots are being displaced to masked locations C and E represented by green dots using the adaptive variable radius mask.

disclosed, then someone that possesses a households' point file could possibly try to estimate the original locations (red dots) from the masked locations (green dots). This can be done by calculating the nearest neighbors of the masked locations. For example, if a masked point is not moved into households' locations (as it happens in this example), then the first three household nearest neighbors could be a possible set of the original point. If a masked point is moved into households' locations, then the first two household nearest neighbors and the masked point could be a possible set of the original point. Based on this logic, the three nearest neighbors of the masked location C are A, B, and C and therefore masked location C cannot be re-identified among three households. The three nearest neighbors of the masked location E are C, D, and E. However C cannot be the original location of masked location E because C will be displaced only within the small circle. Hence masked location E cannot be re-identified among two households, D and E, and the actual K-anonymity for this point is smaller than three. The example shows that with overlapping uncertainty areas there is a possibility of the *Kact* being smaller than the *Kest*.

Our position to the requirements of adaptive geomasking is that a) K-anonymity should be accurate (estimated anonymity = actual anonymity) and b) the masking method and its parameters should be disclosed. This means that K-anonymity is ensured when the protection method and its parameters are disclosed, yet the actual locations cannot be de-identified among k-cases even after a successful reengineering process. We also support the disclosure of the masks' parameters as a proof of a transparent policy and to ensure that privacy is adequately protected. Examples of transparent policies for privacy protection are the Police.UK website (Data.police.uk, 2015) and the Demographic and Health Survey (DHS) Program (Burgert, Colston, Roy, & Zachary, 2013). Both examples release masked confidential information and explain in detail the masking methods they employ.

4. Adaptive areal elimination (AAE)

K-anonymity can be calculated accurately when a predefined K-anonymity is processed at an equal or lower level of the resolution than it is available and when uncertainty areas do not overlap. To achieve this we use existing techniques in an adaptive form to minimize the displacement error. The first technique is point aggregation (Armstrong et al., 1999) where a new symbolic point represents the location of several original points. This technique has been used by snapping each latitude and longitude of confidential points to the nearest point on a square grid (Krumm, 2007) to the nearest street intersection (Leitner & Curtis, 2004), or at the midpoint of a street segment (Leitner & Curtis, 2004). The second technique is random perturbation of confidential points within non-overlapping areas. This technique has been used by randomly translating original points within grid cells (Leitner & Curtis, 2006).

More specifically, this adaptive masking technique can either a) displace confidential points to points that are the centroids of K-anonymized areas, or b) randomly displace confidential points within K-anonymized areas. Each centroid (i.e. centre of gravity) is calculated by the polygon's n vertices (set of geographic coordinates X_i, Y_i) as shown below:

$$\begin{aligned}&\text{Centroid of X coordinate}: Cx = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_{i+1})\,(x_i y_{i+1} - x_{i+1} y_i)\\ &\text{Centroid of Y coordinate}: C_y = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_{i+1})\,(x_i y_{i+1} - x_{i+1} y_i)\end{aligned} \quad (1)$$

where A is the area of the polygon (Bourke, 1988; Waller & Gotway, 2004).

Random displacement means that the confidential points have an equal chance of being relocated at any set of coordinates within their K-anonymized areas. This was accomplished by using the "random point" tool of the ArcGIS 10.0 program. The tool selects two random numbers from a uniform distribution with a minimum and maximum X value and a minimum and maximum Y value, in order to restrict the random set of coordinates to a specified extent (the extent here is the K-anonymized area). To create the K-anonymized areas we first define an attribute of a spatial database named "RoRi" (risk of re-identification) that is used to calculate and ensure a minimum K-anonymity for a masked dataset. RoRi can include information such as: population, households, or addresses. Second, spatial features containing the RoRi attribute can be points, administrative units, and grid cells (i.e. points or polygons). Third, the masking process operates on areas (any type of polygon) and is called "Adaptive Areal Elimination" (AAE).

4.1. Method: Data and steps

To perform the AAE masking method two datasets are required: a) a spatial feature that either represents RoRi (e.g. addresses point file), or contains RoRi information (e.g. administrative units with an attribute field of the number of addresses), and b) a confidential point file. The steps of the method are presented in Fig. 2. The first step is data preprocessing. If an address point file is available, the points have to be aggregated into polygons. The next step is to define a disclosure value for the RoRi field. Disclosure value is the minimum K-anonymity that confidential information can be disclosed. To choose an appropriate disclosure value the current practices or regulations about a specific type of confidential information have to be considered. For example the Police.UK website presents crime information aggregated into "anonymous" map points that have a catchment area, which contains at least eight postal addresses (Data.police.uk, 2015). A minimum of eight addresses may not be a sufficient K-anonymity for health data or even for crime data in countries outside of the UK where such publications

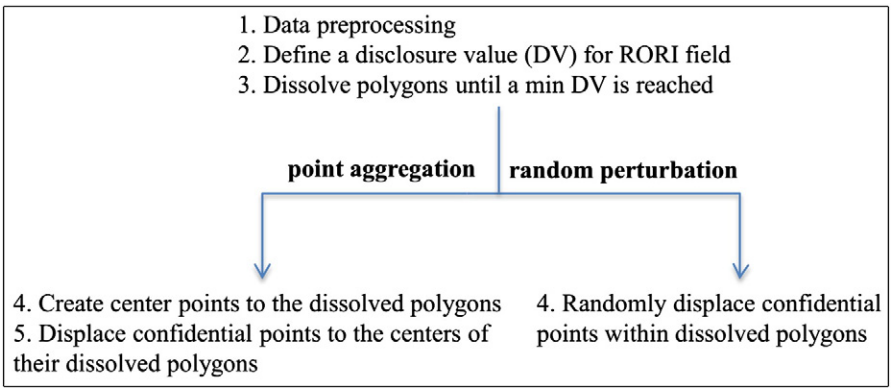

Fig. 2. Adaptive areal elimination (AAE): the steps of the method.

are not allowed. In the third step the dissolving process starts. Based on a predefined disclosure value every polygon that has a RoRi value less than the disclosure value is being dissolved with its neighbor or neighbors until all polygons have RoRi values that are equal or greater than the disclosure value. The general spatial rule is that each polygon is merged with the neighboring polygon that has the longest shared border, unless a polygon is equilateral. In this case, the dissolving process occurs with the equilateral polygon and all adjacent polygons. This step creates the K-anonymized areas (dissolved polygons). The process ensures that the areas are reproducible for the same dataset and disclosure value. As a result of this process the minimum K-anonymity as well as the K-anonymized areas can be disclosed. In the last step the "masker" can either aggregate the confidential points to the centroids of the dissolved polygons or randomly displace the confidential points within the dissolved polygons. Hereafter, point aggregation based on AAE is abbreviated as "APA" (adaptive point aggregation), and similarly random perturbation based on AAE is abbreviated as "ARP" (adaptive random perturbation).

We created an example of a hypothetical confidential dataset to visually illustrate the outputs of the method's steps. The example is located in a small area in Vienna of a 22.4 km^2 size (the area is provided as a polyline file). We also use addresses and street network files from Open Street Map (OpenStreetMap, 2015; OpensStreetMap, 2015). For this example, the address point file (in total 1231 addresses) is the spatial feature that represents RoRi. The disclosure value is 20 addresses (K-anonymity = 20) and the confidential dataset consists of 50 randomly distributed points within the area. Fig. 3 — step 1 shows the pre-processing steps for this example. First, the clipped streets enclosed in the area were merged with the outline of the area (a periphery of street segments). This results in the final polylines that represent the street blocks of the area. The polylines are then converted into polygons that define the boundary of each block. These steps are specific to the input data and would be redundant if an alternative polygon was used (e.g. squared grid). Then, the sum of the addresses in each polygon feature was calculated in the RoRi field and the disclosure value was set to 20 addresses (Fig. 3 — step 2). The addresses' calculation in each feature will also be redundant if the "masker" decides to use a polygon spatial feature that already contains RoRi information (e.g. administrative units). The input and output of the dissolving process for a disclosure value of 20 addresses is shown in Fig. 3 — step 3. The input is the blocks polygon file and the output the final dissolved polygons. The map at the bottom of Fig. 3 — step 3 shows that the process returned 28 K-anonymized areas. Each area (dissolved polygon) contains 21 to 162 addresses. Last, the hypothetical confidential locations and the masked confidential locations were displaced with the APA and the ARP methods, and are shown in Fig. 3 — step 4.

4.2. Adjusting RORI and K-anonymity based on the confidential dataset

For certain masking scenarios RoRi is the K-anonymity of a confidential dataset. This may not be the case for every type of information. To explain how RoRi can be used to ensure a minimum K-anonymity we provide several examples, without excluding the possibility that other types of information, not listed here, may require a different masking treatment.

A first group of datasets are those that pinpoint individuals. Examples are: patients' residencies or individuals with particular private attributes and their residencies (e.g. residencies of drug users). In this case RoRi can be a population, households or residential addresses. Let us assume that a confidential dataset contains locations of breast cancer patients and RoRi is based on a grid file with population information. If one aggregation point represents the location of 100 people and two patients are aggregated to this point, then the disclosure risk is 2%. The two patients are not identified among 100 people and K-anonymity equals to 50 people. Since we want a minimum K-anonymity to be defined from the beginning of the masking process, the RoRi attribute has to be divided by the number of patients in each polygon.

A different approach is required for residential burglaries or domestic violence. For these data RoRi should be either households or residential addresses because the disclosed information pinpoints, or indirectly affects, all members of a family and not only one person. Furthermore, RoRi equals to K-anonymity because the same location can be burglarized multiple times. Hence, if one aggregation point represents the location of 100 households and two burglaries are aggregated to this point, the disclosure risk is still 1%. The two burglaries are not identified among 100 households and the K-anonymity equals to 100 households.

For an "all burglaries" dataset an entire address point file is needed since burglaries cannot be distinguished between commercial or residential locations. As before, RoRi equals to K-anonymity. Last, some crime types can occur in any geographical location (including parks, public buildings, residencies, and streets). Examples of these types are: violence and sexual offenses, homicide, and arson. Nevertheless, these datasets still need to be protected since some locations may be mapped to commercial or residential buildings. Therefore we can treat these datasets as in the previous example and use an address point file for calculating the RoRi.

5. Application and comparison of the method

We use a burglaries dataset from the city of Vienna in Austria to apply the AAE method. The burglaries dataset was provided by the Criminal Intelligence Service Austria (Federal Criminal Police Office). Reported crime data in Austria are stored in the so-called "Security Monitor" database (SIMO) since 2004. Each reported crime incident is geocoded in a local Austrian projection system (MGI Austria Lambert). The dataset was available for the first trimester of 2009 (January to March), and consists of 5806 incidents. The incidents include all types of burglaries (i.e. single family houses, apartments, basements, and businesses). Since this type of information requires an entire address point file for the RoRi attribute, we used the Open Street Map addresses as previously (77,862 points). Visual observations of the study area showed that Open Street Map addresses are somehow incomplete for some suburbs of Vienna. This issue is not in conflict with preserving the privacy in the dataset because in these neighborhoods the "real" K-anonymity will be higher due to the missing addresses. However, a complete as possible dataset would have been preferred, if it were available, because it would yield smaller spatial errors (more addresses implies more and smaller K-anonymized areas). Also, we used the blocks of the entire city of Vienna as the polygon units for the method (22,139 polygons). The pre-processing of the data was performed similarly to the method's example in the previous section.

The burglaries were masked using the APA and the ARP masking methods. Additionally, we used the donut geomasking method (Hampton et al., 2010), which is available as a Python code (DonutGeomask, 2015). In order to select a K-anonymity we looked at current practices and guidelines with regards to crime data. First, a report by the Information Commissioner's Office (ICO, 2012) visualizes in a scale bar the number of households' as disclosure thresholds for privacy protection of spatial crime data. The report suggests that a number of households from 1 to 10 impose high re-identification risk, 10 to 30 a medium re-identification risk, and more than 30 a lower risk. Second, as already mentioned, the "Location Anonymization" technique that is currently employed by the Police.uk website uses a minimum of eight addresses as a disclosure threshold value (Data.police.uk, 2015). Last, the respondents of a survey, which was carried out in London with regard to privacy and crime mapping, would opt for a medium-risk protection method in terms of risk of privacy violation (Kounadi, Bowers, & Leitner, 2014). More specifically, the majority of the respondents preferred the "Location Anonymisation" technique that is currently employed by the Police.uk website or a protection method with a disclosure threshold

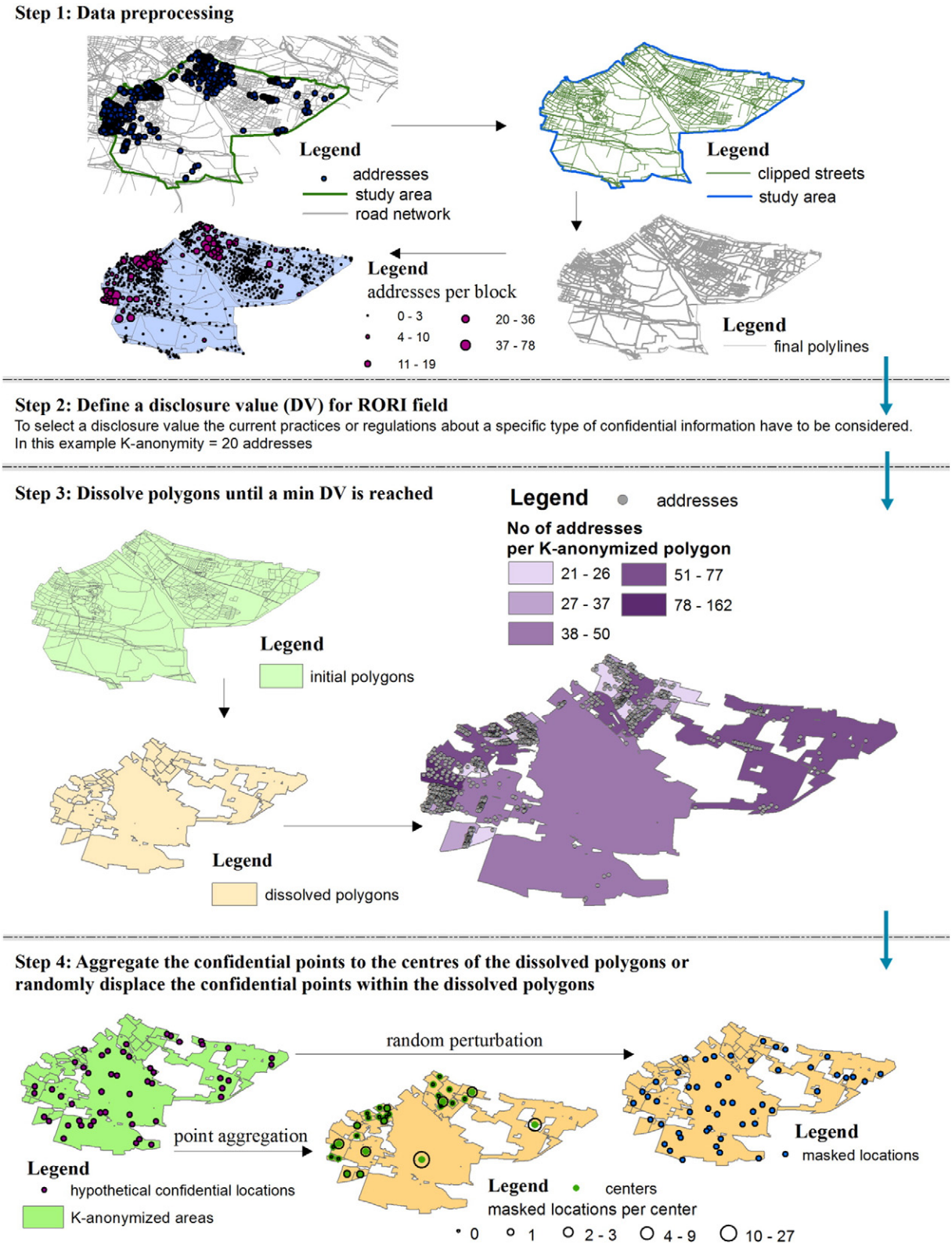

Fig. 3. Example of a hypothetical masking scenario. Step 1: initial processing of input data. Step 2: selection of the disclosure value. Step 3: the dissolving process of the AAE masking method. Step 4: masked points from adaptive point aggregation (APA) and masked points from adaptive random perturbation (ARP) within K-anonymized areas.

value that ranges from 8 to 20 postal addresses. Hence, for all the masking methods (APA, ARP, and donut) a disclosure value of 20 addresses was selected. An additional minimum disclosure value was required for the donut mask, which defines the inner radius of the torus. This was set to 2 addresses (10% of the maximum disclosure value).

The uncertainty areas for the APA and ARP masks are dissolved polygons (K-anonymized areas). The uncertainty areas for the donut mask are donuts (toruses) that may or may not be "eaten". A donut appears "eaten" because a part of it fell outside its polygon file and it was excluded. Polygons can be administrative areas. Hence, with donut

geomasking original points are relocated in a random direction by at least a minimum distance and less than a maximum distance, while retaining the original points within their geopolitical boundaries. The geopolitical boundary that is selected here is the "Zählbezirk", which is similar to a registration district. In total there are 245 "Zählbezirk" regions in Vienna with an average size of 1.69 km^2. Different types of uncertainty areas are shown in Fig. 4. The figure depicts a small area in Vienna. For privacy reasons the area has been rotated and does not include the "Zählbezirk" regions, which would allow localizing the area in Vienna. Also, since address-level crime data are not currently published in Austria, the actual crime locations are not presented here. To account for limitations of the donut mask, we calculated the number of overlapping donuts and the number of original burglaries that achieved an actual K-anonymity lower than 20 addresses. Actual K-anonymity was defined as the number of points that were closer to the original location than the maximum distance of displacement (outer radius). The vast majority of the donuts overlapped (98.5%) and 33.9% of the burglaries achieved a K-anonymity that ranges from 0 to 19 addresses.

As a next step the donut mask was modified to achieve an actual K-anonymity of at least 20 addresses for all points. First, the restriction that the masked points should retain their geopolitical boundaries was removed. Then further runs of the donut mask algorithm were performed by increasing every time the maximum K-anonymity until the desired actual K-anonymity (*Kact*) was reached. Similarly to Allshouse et al. (2010) the parameter for the minimum K-anonymity (*Kmin*) was adjusted accordingly to the 10% of the maximum K-anonymity (*Kmax*). The results of Table 1 show that with a *Kmax* of 550 addresses all burglaries achieved a *Kact* equal or greater than 20 addresses. Furthermore, with a *Kmax* of 150 addresses or more the vast majority of the burglaries achieved the desired K-anonymity with less than 1% of burglaries having lower *Kact* than 20 addresses.

5.1. Measures of spatial error

The final part of this analysis is to compare the spatial error that is introduced to masked datasets with the use of the geographical masks. To calculate the error we use four measures. The first measure is the *displaced distance* in meters, which is the distance from the original point to the masked point. The second measure is the *correlation coefficient between density surfaces* created using original and masked points. Shi, Alford-Teaster, and Onega (2009) performed a kernel density estimation (KDE) using a series of bandwidths to create density surfaces from original points and masked points. The points were masked using a random translation geographical mask with a fixed maximum distance and the bandwidths were chosen as portions and multiples of that distance. Then the Pearson's correlation coefficients between the density surfaces were calculated. The statistic shows the effect of the geomasking process on the KDE analysis and with which parameters masked points will provide similar KDE results to the KDE results of the original points. In our application there is no fixed maximum displaced distance and it can also vary by method. Therefore, we used the average displaced distance for all pairs of original and masked points in all methods. Then we selected one bandwidth that is equal to the average displaced distance, one that is 0.25 times the distance, and one that is four times the distance. The KDE was calculated using a normal distribution function as shown below:

$$g(\boldsymbol{x}_j) = \Sigma\left\{[W_i \times I_i] \times \frac{1}{h^2 \times 2\pi} \times e^{-\left[\frac{d_{ij}^2}{2\times h^2}\right]}\right\} \quad (2)$$

where d_{ij} is the distance between a point and any reference point in the area, h is the bandwidth, W_i is a weight at the point and I_i the intensity (equation adopted by Levine (2004)).

The third measure is the Hotspots' Divergence (Kounadi & Leitner, 2015a). Hotspots' Divergence is an index that calculates how much dissimilar are the spatial clusters that are created from masked points compared to spatial clusters that are created from original points. The index ranges from 0 to 100 and calculates the non-overlapping areas (symmetrical difference) of masked and original spatial clusters. A value of 0 means that masked clusters are identical to original clusters and a value of 100 means that masked and original clusters are disjointed (their areas do not overlap). The index can be calculated using the following formula:

$$\text{Local divergence} = \frac{\text{Symmetric difference of A and B}}{\text{A} + \text{B}} \times 100 \quad (3)$$

where A = area of original hotspots and B = area of masked hotspots.

This index was employed with the nearest-neighbor hierarchical clustering method (Everett, 1974) and is denoted as *Nnh.di*. Kounadi and Leitner (2015b) compared the Nnh.di with hotspots divergence indexes that use other spatial clustering techniques such as the Getis-Ord Gi* statistic (Gi*.di) and the Anselin Local Moran's I statistic (Ans.di). The Nnh.di had the highest correlation with the perceived similarity of a pair of point pattern. Thus, it is preferred as a measure that can predict how dissimilar a masked point pattern would be perceived compared to

Table 1
The percentage of original points that achieve an actual k-anonymity (*kact*) equal or greater than 20 addresses for several values of minimum K-anonymity (*Kmin*) and maximum K-anonymity (*Kmax*) using the donut mask.

Kmin	Kmax	% Kact ≥20
2	20	66.10%
5	50	93.70%
10	100	98.05%
15	150	99.30%
20	200	99.60%
25	250	99.65%
30	300	99.74%
35	350	99.79%
40	400	99.86%
45	450	99.93%
50	500	99.98%
55	550	100.00%

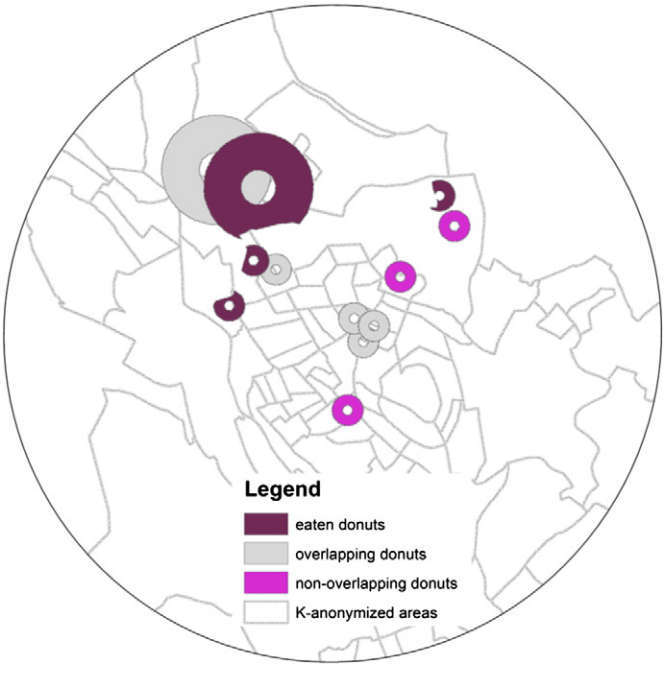

Fig. 4. Uncertainty areas for the APA and ARP masks are the K-anonymized areas and the uncertainty areas for the donut mask are different types of donuts.

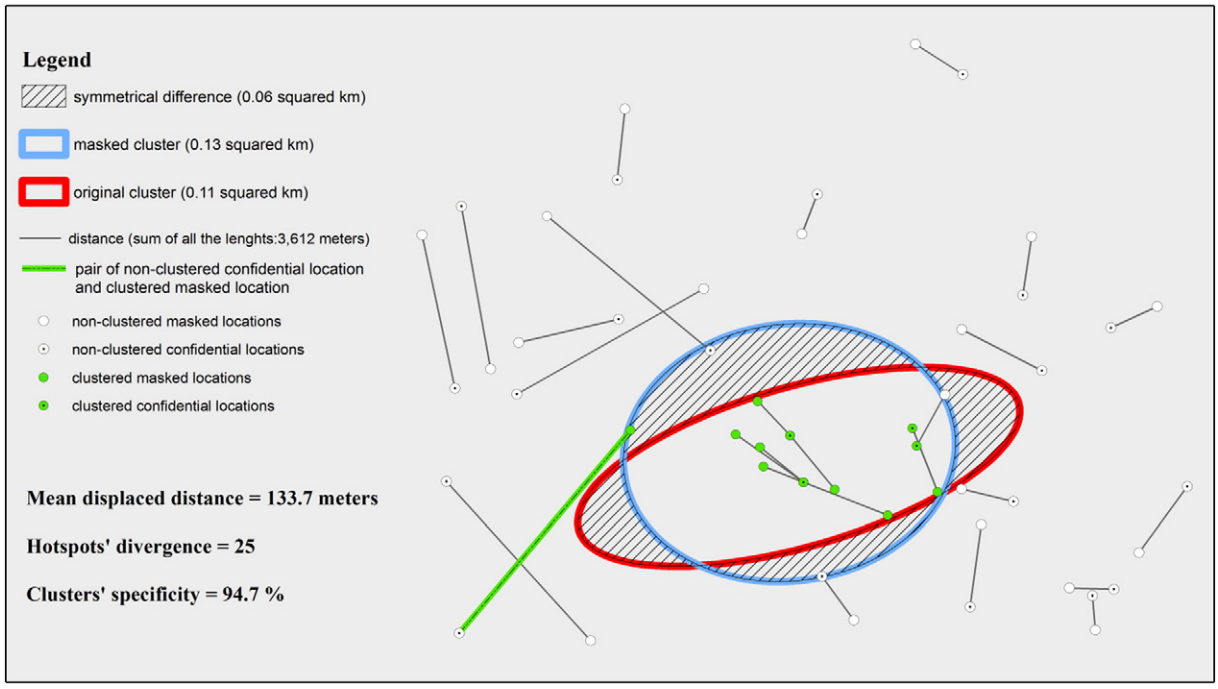

Fig. 5. A snapshot of the study area that illustrates the three measures of spatial error: a) displaced distance, b) hotspot's divergence, and c) clusters' specificity.

the original one. To allow further comparisons and the predictability of the perceived similarity, we used the parameters that were used in the aforementioned study, including two standard deviational ellipses to outline clusters, a minimum of five points per cluster, only first-order clusters, and a search radius based on the random nearest neighbor distance.

The fourth measure is the *clusters' specificity*, which calculates the percentage of masked points that originate from non-clustered original points and are still non-clustered. This measure was also used previously to evaluate the error of adaptive geographical masks (Cassa et al., 2006; Hampton et al., 2010). The four measures were calculated for the three initial masks (APA, ARP, donut mask with $Kmax = 20$). Since the donut mask with $Kmax = 20$ does not protect satisfactorily the vast majority of the original points, we also included in the analysis the donut mask with $Kmax = 150$ where 99.3% of the points have a $Kact \geq 20$ (less than 1% error rate). The first, third, and fourth measures are visualized in a small snapshot of the study area in Fig. 5.

The results of the measures are presented in Table 2. The mask that outperforms all others, for all measures of spatial error, is the donut mask with $Kmax = 20$. The mask with the second smallest mean displaced distance is APA (244.11 m), followed by the ARP (316.94 m), and the donut mask with $Kmax = 150$ (365.18 m). The average displaced distance for all pairs of original and masked points for all methods was found to be 264 m and was used to select the bandwidths of 66 m, 264 m, and 1056 m for calculating the correlation coefficient between density surfaces. For a bandwidth of 1056 m the surfaces of masked and original points were almost identical for all masks. For the smaller bandwidths, the ARP had the second highest correlation coefficient (0.76 and 0.98), followed by the donut mask with $Kmax = 150$ (0.62 and 0.97). Our results show that by using bandwidths that are close to or higher than the displaced distance the produced masked density surface will be highly correlated to the original density surface. Also, for both clusters' specificity and Nnh.di the mask with the second smallest error is ARP (Nnh.di = 51.55 and clusters' specificity = 84.26%) and is followed by the donut mask with $Kmax = 150$ (Nnh.di = 53.70 and clusters' specificity = 78.85%). The APA performs the worst in terms of the KDE (correlation coefficients are 0.46 and 0.94), retaining original locations of spatial clusters (Nnh.di = 63.60), as well as misclassifying non-clustered points as clustered points (clusters' specificity = 35.90%). Furthermore, Kruskal–Wallis tests were applied to the results of the first and second measures to find out whether the displaced distances and density surfaces are statistically different by masking method. The results of the test show that displaced distances are statistically different among the masks ($H = 4092$, $p < 0.001$), that density surfaces produced with a 66 m bandwidth are statistically different among the masks ($H = 311.1$, $p < 0.001$), and that density surfaces produced with a 264 m bandwidth are statistically different among the masks ($H = 53.43$, $p < 0.001$). However, density surfaces produced with a 1056 m bandwidth are not statistically different among the masks ($H = 3.34$, $p = 0.34$).

Table 2
The spatial error of masked datasets for four masks (APA, ARP, donut — $Kmax = 20$, and donut — $Kmax = 150$) and four measures (displaced distance, correlation coefficient between density surfaces, divergence of nearest neighbor hierarchical clusters — Nnh.di, and clusters' specificity).

	Displaced distance (meters)				Correlation coefficient between density surfaces by bandwidth			Nnh.di (0–100)	Clusters' specificity (%)
	Mean	Min	Max	RMSE	66 m	264 m	1056 m		
APA	244.11	1.85	6436.06	430.15	0.46	0.94	1.00	63.60	35.90%
ARP	316.94	1.83	6056.73	562.40	0.76	0.98	1.00	51.55	84.26%
Donut (Kmax = 20)	130.40	18.88	1858.97	190.12	0.85	0.99	1.00	42.45	85.46%
Donut (Kmax = 150)	365.18	51.48	4908.61	539.22	0.62	0.97	1.00	53.70	78.85%

6. Discussion

The preferment and use of geographical masking techniques like AAE against other techniques such as spatial aggregation lies in the fact that the latter yields a substantial spatial information loss compared to geographical masking. Although privacy can be guaranteed through aggregation there is little potential for further quantitative analysis (Fefferman, O'Neil, & Naumova, 2005). For example, Luo, McLafferty, and Wang (2010) disaggregated cancer cases from zip code to census block to examine the relationship between late-stage breast cancer and risk factors using logistic regression. The results of the coefficients showed that the disaggregation of spatially aggregated units may lead to inaccurate findings on health risk factors. Further drawbacks of aggregation compared to geomasking are the reduced sensitivity to cluster detection and that clusters crossing administrative boundaries cannot be identified (Cassa et al., 2006).

This paper presented the AAE geographical masking method as an alternative approach to protect spatial datasets with a user-defined level of k-anonymity. Additionally, the K-anonymity quantifies the maximum level of disclosure risk accurately, while allowing the parameters of the method to be disclosed (e.g. K-anonymized areas, point aggregation or random perturbation). Our approach differs from previous techniques since it is adapted based on the resolution that the disclosure information (RoRi) is available and thus, subject to the accuracy of the RoRi information, reports the risk of re-identification accurately. To analyze the effectiveness of the method a real dataset was used (burglaries in Vienna). However the proposed method is generally applicable to other types of confidential datasets such as health and demographic data. Furthermore, the Adaptive Areal Elimination approach has been automated with a Python code for the ArcGIS program. Further queries about the code may be addressed directly to the corresponding author of this paper.

AAE overcomes the main limitation of other adaptive masking techniques, which is the assumption of a homogeneous region for the calculation of the masking displacements. When a donut mask with a $Kmax = 20$ addresses is used, a considerable amount of original points (33.9%) have a *Kact* between 0 and 19 addresses. By increasing the K-anonymity from 20 to 150 less than 1% of the original points (0.7%) have a *Kact* between 0 and 19 addresses. Nevertheless, our results differ from the results of Allshouse et al. (2010) that also examined the *Kest* (*Kmin* or *Kmax*) against the *Kact* of the donut mask. Allshouse et al. (2010) suggested that for heterogeneous population the *Kmin* should be tripled to protect privacy with less than 1% error rate. According to our results a $Kmin = 15$ (5 times higher than 3) is required to achieve the desired level of anonymity (20 addresses). The results seem to be highly dependent on the study area and the distribution of the population. Therefore a single correction factor (such as tripling the parameters) may not guarantee a required level of privacy.

Using four measures of spatial error and by comparing the results of the AAE with those of the donut mask, we evaluated the effectiveness of the masks in terms of preserving the original spatial characteristics of the point pattern. Apart from the displaced distance, the APA performs considerably worse than the ARP and the donut mask. The correlation coefficients of density surfaces are the smallest among the masks, the clusters' specificity is only 35.9% while the divergence of the hotspots is much higher than the divergence of the other masks (63.6). Our results agree with the results of previous studies suggesting that aggregation results in higher errors compared to other geographical masks (Hampton et al., 2010; Wieland et al., 2008). On the other hand, for an equal user-defined level of K-anonymity (20 addresses) the donut mask yielded the smallest spatial error for all three measures. However, as already mentioned, the donut mask with $Kmax = 20$ preserves the desired anonymity only for 66.1% of the original points. Hence, a fair comparison should include the APA, the ARP, and the donut mask with $Kmax = 150$, all of which preserve the desired anonymity for more than 99% of the original points. The results demonstrate that masked points from the ARP preserve the original spatial clusters better and have the highest clusters' specificity compared to the donut mask with $Kmax = 150$ or the APA. The APA retains the smallest displaced distance but the divergence of the ARP is 2.15 units lower than the divergence of the donut mask and 12.05 units lower than the divergence of the APA. Furthermore, considering the prediction models of the perceived similarity of point patterns by Kounadi and Leitner (2015b), for a K-anonymity of 20, the masked pattern of the APA will be perceived as less similar to the original one compared to the patterns produced by the ARP and the donut masks. Additionally, the masked pattern of the APA is the only pattern that is not perceived as similar or very similar to the original pattern for all predication models. Also the ARP increases the clusters' specificity, which is 5.41% higher than the donut mask and 48.36% higher than the APA.

A possible limitation of our approach is that it is examined for one study area. Nonetheless, Vienna is characterized by a variety of population densities (0–1416 addresses per km^2) and seems to be a good example for heterogeneous areas. For more homogenous areas a *Kest* of the donut mask will be closer to a *Kact* than the *Kact* in the analysis of this study. Thus, the donut mask could lead to less spatial information loss than ARP. Since this conclusion requires further investigation, we suggest that ARP performs better than the donut mask in heterogeneous areas. Also, an assumption of the method is that location information (e.g. geographical coordinates) is the only available information within the confidential dataset. If other variables are included in the dataset (i.e. age, ethnicity, sex) and the data "masker" wants to disclose them as well, a different de-identification strategy has to be developed. For example if a K-anonymity of 20 is required for a dataset that consists of ethnicity and location information, an original point has to be displaced within a K-anonymized area that contains at least 20 or more people of the same ethnicity. This moves the spatial K-anonymity closer to its original K-anonymity approach (Sweeney, 2002) and all quasi-identifiers have to be examined for each original point.

Last, two additional steps are required before applying the AAE masking method to a confidential dataset. The first is to select an appropriate RoRi file. The finer the resolution of the RoRi file, the more K-anonymized areas will be created, thus the overall spatial error of the masking process will be decreased. In respect to the types of RoRi information, several examples are mentioned in Section 4.2 that cover the majority of confidential discrete location datasets for which a masking process may be required. The second issue considered is the selection of the disclosure value (a minimum of K-anonymity). Different thresholds should be applied for different types of confidential data according to the regulations or guidelines posed by respective agencies (e.g. health or crime related organizations). Finally, we suggest using the random perturbation form of AAE that involves less spatial error than the point aggregation. The application of the method could be to either anonymize publicly available data on Web-services or single case anonymization such as a map presented on a scientific publication.

Author contributions

O.K and M.L conceived of the study, wrote the paper and interpreted the results. O.K designed the study, developed the method, and analyzed the data. M.L. coordinated the study and consulted on the geographic techniques to be used.

Acknowledgments

This research was funded by the Austrian Science Fund (FWF) through the Doctoral College GIScience at the University of Salzburg (DK W 1237-N23). Also, we thank Thomas J. Lampoltshammer (Salzburg University of Applied Sciences) and Dražen Odobašić (University of Zagreb) for their technical support.

References

dhsprogramm (2015). Available online: http://dhsprogram.com/data/ (accessed on 02 March 2015)

police.uk (2015). Available online: http://www.police.uk/ (accessed on 02 March 2015)

Allshouse, W.B., Fitch, M.K., Hampton, K.H., Gesink, D.C., Doherty, I.A., Leone, P.A., ... Miller, W.C. (2010). Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. *Geocarto International*, *25*, 443–452.

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

Bourke, P. (1988). *Calculating the area and centroid of a polygon.*

Brownstein, J.S., Cassa, C.A., Kohane, I.S., & Mandl, K.D. (2006). An unsupervised classification method for inferring original case locations from low-resolution disease maps. *International Journal of Health Geographics*, *5*, 56.

Burgert, C.R., Colston, J., Roy, T., & Zachary, B. (2013). *Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys.*

Cassa, C.A., Grannis, S.J., Overhage, J.M., & 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*, 160–165.

Cassa, C.A., Wieland, S.C., & Mandl, K.D. (2008). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics*, *7*.

Cottrill, C.D. (2011). Location privacy: Who protects? *URISA Journal-Urban and Regional Information Systems Association*, *23*, 49.

Data.police.uk (2015). Available online: http://data.police.uk/about/#location-anonymisation (accessed on 12 February 2015)

DonutGeomask (2015). Available online: http://www.unc.edu/depts/case/BMElab/donutGeomask/pyDonutGeomask1.0.htm (accessed on 16 February 2015)

Duncan, G.T., & Pearson, R.W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, *6*, 219–232.

Everett, B. (1974). *Cluster analysis.* London: Heinemann Educational Books Ltd.

Fefferman, N.H., O'Neil, E.A., & Naumova, E.N. (2005). Confidentiality and confidence: Is data aggregation a means to achieve both? *Journal of Public Health Policy*, *26*, 430–449.

Graham, C. (2012). *Anonymisation: managing data protection risk code of practice.* Information Commissioner's Office.

Gruteser, M., & Grunwald, D. (2003). Anonymous usage of location-based services through spatial and temporal cloaking. *Proceedings of the 1st international conference on mobile systems, applications and services* (pp. 31–42). San Francisco, California: ACM.

Hampton, K.H., Fitch, M.K., Allshouse, W.B., Doherty, I.A., Gesink, D.C., Leone, P.A., ... Miller, W.C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, *172*, 1062–1069.

ICO (2012). *Crime-mapping and geo-spatial crime data: Privacy and transparency principles.*

Kounadi, O., & Leitner, M. (2014). Why does geoprivacy matter? The scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics* (1556264614544103).

Kounadi, O., & Leitner, M. (2015a). Spatial information divergence: Using global and local indices to compare geographical masks applied to crime data. *Transactions in GIS*, *19*, 737–757.

Kounadi, O., & Leitner, M. (2015b). Defining a threshold value for maximum spatial information loss of masked geo-data. *ISPRS International Journal of Geo-Information*, *4*, 572–590.

Kounadi, O., Bowers, K., & Leitner, M. (2014). Crime mapping on-line: Public perception of privacy issues. *European Journal on Criminal Policy and Research*, *1-24*.

Kounadi, O., Lampoltshammer, T.J., Leitner, M., & Heistracher, T. (2013). Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic Information Science*, *40*, 140–153.

Krumm, J., & Truong, K. (2007). Inference attacks on location tracks. In A. LaMarca, & M. Langheinrich (Eds.), *Pervasive computing. Vol. 4480*. (pp. 127–143). Berlin Heidelberg: Springer.

Kulk, S., & Van Loenen, B. (2012). *Brave new open data world?* .

Kwan, M.P., Casas, I., & Schmitz, B.C. (2004). Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, *39*, 15–28.

Leitner, M., & Curtis, A. (2004). Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives*, *49*, 22–39.

Leitner, M., & 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*, 813–822.

Leitner, M.; Mills, J. W.; Curtis, A., Can novices to geospatial technology compromise spatial confidentially? Kartographische Nachrichten ('Cartographic News') 2007, 57, 78–84.

Levine, N. (2004). Chapter 8 kernel density interpolation. In Levine, N.; Associates (Ed.), *CrimeStat III: A spatial statistics program for the analysis of crime incident locations (version 3.0)*. Washington: DC: National Institute of Justice.

Luo, L., McLafferty, S., & Wang, F.H. (2010). Analyzing spatial aggregation error in statistical models of late-stage cancer risk: A Monte Carlo simulation approach. *International Journal of Health Geographics*, *9*.

Nass, S.J., Levit, L.A., & Gostin, L.O. (2009). *Beyond the HIPAA privacy rule: Enhancing privacy, improving health through research.* Washington, DC: National Academies Press.

OpensStreetMap (2015b). Available online: http://wiki.openstreetmap.org/wiki/Key:addr (accessed on 13 February 2015)

OpenStreetMap (2015a). Available online: http://wiki.openstreetmap.org/wiki/Highways (accessed on 13 February 2015)

Shi, X., Alford-Teaster, J., & Onega, T. (2009). *Kernel density estimation with geographically masked points. 17th international conference on geoinformatics*, *Vols. 1 and 2*. (pp. 1153–1156), 1153–1156 (2009).

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*, 557–570.

Vicente, C.R., Freni, D., Bettini, C., & Jensen, C.S. (2011). Location-related privacy in geo-social networks. *IEEE Internet Computing*, *15*, 20–27.

Waller, L.A., & Gotway, C.A. (2004). *Applied spatial statistics for public health data. Vol. 368*. (pp. 48–49). John Wiley & Sons, 48–49.

Wartell, J., & McEwen, J.T. (2001). *Privacy in the information age: A guide for sharing crime maps and spatial data series: Research report.* Institute for Law and Justice.

Wieland, S.C., Cassa, C.A., Mandl, K.D., & 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*, 17608–17613.