Adaptive Voronoi Masking: A Method to Protect Confidential Discrete Spatial Data

Fiona Polzin
ITC-Faculty of Geoinformation and Earth Observation,
University of Twente, Enschede, The Netherlands

Ourania Kounadi[1]
Department of Geography and Regional Research, University of Vienna, Austria

Abstract

Geomasks assure the protection of individuals in a discrete spatial point data set by aggregating, transferring or altering original points. This study develops an alternative approach, referred to as Adaptive Voronoi Masking (AVM), which is based on the concepts of Adaptive Aerial Elimination (AAE) and Voronoi Masking (VM). It considers the underlying population density by establishing areas of K-anonymity in which Voronoi polygons are created. Contrary to other geomasks, AVM considers the underlying topography and displaces data points to street intersections thus decreasing the risk of false-identification since residences are not endowed with a data point.

The geomasking effects of AVM are examined by various spatial analytical results and are compared with the outputs of AAE, VM, and Donut Masking (DM). VM attains the best efficiency for the mean centres whereas DM does for the median centres. Regarding the Nearest Neighbour Hierarchical Cluster Analysis and Ripley's K-function, DM demonstrates the strongest performance since its cluster ellipsoids and clustering distance are the most similar to those of the original data. The extend of the original data is preserved the most by VM, while AVM retains the topology of the point pattern. Overall, AVM was ranked as 2nd in terms of *data utility* (i) and also outperforms all methods regarding the *risk of false re-identification* (ii) because no data point is moved to a residence. Furthermore, AVM maintains the *Spatial K-anonymity* (iii) which is also done by AAE and partly by DM. Based on the performance combination of these factors, AVM is an advantageous technique to mask geodata.

2012 ACM Subject Classification Security and privacy → Privacy protections; Security and privacy → Data anonymization and sanitization; Information systems → Geographic information systems; Mathematics of computing → Exploratory data analysis

Keywords and phrases Geoprivacy, location privacy, geomasking, Adaptive Voronoi Masking, Voronoi Masking, Adaptive Aerial Elimination, Donut Geomasking, ESDA

Digital Object Identifier 10.4230/LIPIcs.GIScience.2021.II.1

Supplementary Material The Geoprivacy Github repository contains the scripts to run AVM (Also AAE) as well as one of the six area data sets used in this study:
Software: `https://github.com/okounadi/Geoprivacy`

1 Introduction

1.1 Background

The advances of GIS and the interest in spatial analysis have led to an increase of thematic maps in research and online platforms visualizing point data. However, several studies in health geography, reproductive and sexual health did not anonymize or aggregate data; instead, the original data were used [4, 14, 17]. Publishing an individual's location either in

[1] Corresponding author

paper or digital form - knowingly or unknowingly - increases the risk of re-identification by and violates individual privacy. How simple the re-identification of individuals is, was already demonstrated by Brownstein et al. [4]: By applying the reverse-identification method, 7% of the spatially coded addresses were accurately identified while all 550 of the plotted address points were disclosed within 14 m of the right address. Furthermore, Kounadi and Leitner [17] exposed that within an eight-year duration, almost 70,000 home addresses had been disclosed in academic research.

The consequences of disclosure are vast; an individual being identified as an HIV-patient - correctly or wrongly - can affect him or her by discrimination or social stigmatization [27]. Identifications may cause harassment [9], unwanted advertisement or humiliation [26, 25]. Kounadi and Leitner [19] criticize that general rules on privacy do not include details of the spatial re-identification risk notwithstanding the fact that relevant research and reports on geodata exist [12, 20]. Consequently, confidential spatial data sets do not only have to be preserved but also need to comply with present-day restrictions and regulations on the right to privacy [19]. However, Ajayakumar et al. [1] criticized that geomasks are still unavailable for many institutions due to the lack of expertise in geospatial proficiency although the awareness of the power of mapping has grown particularly in health organizations and clinics which have become spatially literate lately. The authors stress that geomasks need to become more of a real-world requirement.

1.2 Problem statement

Some geomasks displace the points a specific distance aside from its original location (e.g., *local random rotation* by Leitner and Curtis [22] and *Voronoi masking* (VM) by Seidl et al. [29], while others aggregate points (e.g., *spatial and point aggregation* by Armstrong et al. [3]). Other geomasks consider the underlying population density adapting the displacement error such as the *Donut Geomasking* (DM) [15] and the *Adaptive Areal Elimination* (AAE) [19]. By considering the population density, the "masker" is able to determine a level of *K-anonymity* in which each record (i.e. person) within a masked data set cannot be identified from at least K-1 records [24]. Regarding geodata, K-anonymity assures that every location such as household, address or an individual's location cannot be differentiated from minimum K-1 locations. This means, that *spatial K-anonymity* (SKA) describes the probability of identifying a location that can be linked to an individual by reverse geocoding. This is needed to evaluate the degree of privacy and when measuring the degree of displacement.

A possible solution to prevent re-engineering of original locations could be points' aggregation. However, when doing so, the ability to distinguish spatial relations or clusters and deriving persuasive information is decreased [30, 3, 21]. Obviously, the data becomes less useful for research purposes [30, 13]. Contrary to aggregation, geomasks that modify the locations are preferred for analytical purposes. Nevertheless, the transferred points can be moved to a position which has real observations [23] or where they cannot exist [7], resulting in *false identification* [29]. False identification represents the incorrect linking of a household or person to a data point. Contrary to that, correct identification is the correct linkage of a household or person to a data point [29]. The consequences of identification can result in negative effects impeding an individual's social prominence [13]. Besides, it can unintentionally involve individuals, who were not part of the research [7]. Such limitations influence both the disclosure risk and a successful investigation of spatial patterns.

Generally, there is neither a recommended nor approved geomask technique [30, 13] and each method has disadvantages and advantages. Zandbergen [30] suggests counterbalancing data utility and confidentiality protection. Also, not a lot of geomasks consider the underlying

topography except for the *Street aggregation at intersection* or at *midpoint* [22] or the *Location Swapping* method [31]. Yet these that do consider the underlying topography do not offer a predefined level of SKA. It is evident that existing techniques must be improved to overcome such shortcomings and also become widely accessible.

1.3 Study scope and design

Our alternative approach, referred to as Adaptive Voronoi Masking (AVM), is based on the concepts of AAE and VM. AVM shall protect the individual's privacy based on SKA while also decreasing the false re-identification risk. We evaluate known geomasks, namely the VM, the AAE, and the DM, and compare them with the proposed AVM in terms of three key aspects: a) SKA, b) false re-identification, and c) data utility.

In the next section (Methodology) we explain the two geomasks that AVM is based on (VM and AAE) and then describe the functionality of AVM. Next, we present the exploratory spatial data analysis (ESDA) methods that are used to compare and evaluate the original data points with the outcome of the geomasks (i.e. masked data points). Last, we introduce the study area, the software, and data used. In section 3 (Results), we report the ESDA results and finally discuss and conclude our findings in section 4 (Conclusion). Apart from the AVM, DM, and AAE, we also evaluate the DM geomask. DM was chosen as a comparative geomask since it is a popular technique and it has a small effect on the geographical characteristics of the original point pattern as highlighted in academic literature [15, 30, 2]. The algorithm for this method was retrieved online[2].

2 Methodology

2.1 Adaptive Areal Elimination (AAE)

AAE assures privacy by moving the original locations within uncertainty areas. The so-called uncertainty areas describe an area, where the masked points are displaced in, e.g. torus or circle [19]. For instance, DM moves the original data within an uncertainty area selected from a uniform distribution [15] while the population-density-based Gaussian spatial blurring dislocates points within a circle based on a normal distribution [5]. However, these geomasks assume that population is homogeneously distributed - which is not the case in most instances. This assumption can result in masked data points with a lower actual K-anonymity than the estimated K-anonymity [2]. Hence, AAE is aiming to ensure K-anonymity even when the geomasking method and its parameters are known. K-anonymity can be measured precisely when uncertainty areas do not overlap and when it is applied at a lower or equal level of the available resolution [19].

To execute the AAE algorithm, two data sets are needed: a) a point file and b) a spatial data set that either includes an attribute with discrete information (e.g., as administrative units containing an attribute field with the total households in each unit) or represents discrete information (e.g., point data representing households). This attribute is called RoRi (risk of re-identification). Generally, risk of re-identification can contain information such as addresses, households, or population. A disclosure value for this field predefined to describe the minimum K-anonymity which is used to obscure confidential information. In the next step, the process of merging polygons starts: depending on the disclosure value,

[2] `https://mserre.sph.unc.edu/BMElab_web/donutGeomask/donutGeomask.htm` *(Last accessed on January 22th*, 2021)

every polygon containing a lower risk of re-identification value than the disclosure value, is merged with its neighbouring polygon or polygons until each polygon has values that are either greater than or equivalent with the disclosure value to create the K-anonymized areas. Next, original data are aggregated to the centroids of the merged polygons or randomly displaced within the merged polygons. Random displacement can be performed by a random perturbation to the coordinates of each data point by a random distance and at a random direction (equation 1).

$$Xm = Xo + D * cosine(\Theta)$$
$$Ym = Yo + D * sine(\Theta) \quad (1)$$

where, Xo, Yo are the original coordinates of a point, Xm, Ym are the resulting masked coordinates, D is a random value within a predefined range, and Θ *is a random angle.*

In the AAE each masked point shall lie within its k-anonymized polygon. Thus, the displaced masked point/s has to be conditioned on the boundaries of each polygon. In this study, we implement the random displacement that yielded better performance results in the study by Kounadi and Leitner (2016). When studying the outputs of AAE more closely, some masked data are moved further distances than necessary. This can be explained by the process of merging polygons that selects the neighbour with the longest boundary, which may result in K-anonymized areas that are larger than needed to ensure SKA.

2.2 Voronoi Masking (VM)

VM creates Voronoi polygons around the original data points are displaced to the closest segment (edge) of its corresponding polygon [29, 13]. The theoretical basis of creating Voronoi polygons starts with the triangulation of the original points into an irregular network that meets the *Delaunay criterion* (i.e. no point is inside the circumcircle of any triangle). Then, the perpendicular bisectors for each triangle edge are generated. These are the edges of the Voronoi polygons while the locations of the bisector's intersections determine the vertices. Every point within each polygon is closer to the original point of its creation than to other original points.

Advantages of VM is that points in neighbouring polygons are displaced to the same position, enhancing their K-anonymity and that a higher point density results in smaller distances between the original data and masked data thus giving a pattern that is similar to the original one [29, 13]. For a small scale area or an area with a minimum of two households, VM dislocates the original data a lesser distance than compared to other geomasks that do not consider the underlying settlement patterns. VM is an efficient approach regarding the preservation of the spatial point pattern as it has been proved by Seidl et al. [29] who implemented various methods to evaluate its performance. Finally, Seidl et al. [29] praise that in case of applying a data set that is including all residences within the area of interest, no displaced point will be located on an actual residence and thus false identification of residences is not possible. Points are typically located in the centre of a parcel or at the street segment. VM will definitely move points away from these locations. However, this process does not guarantee that segments of Voronoi polygons will not cross residential parcels and therefore VM cannot decrease the risk of false identification in this regard.

When applying VM in areas with scattered residences some data points will be dislocated at large distances, which affect spatial patterns. Also, a smaller amount of masked data will be depicted on the map than the original data due to the overlapping of points at the

displaced locations. Although this assures a higher K-anonymity, the map viewer may not be aware that some points represent at least two addresses, increasing the risk of spatially analysing or perceiving the output differently. Last, although K-anonymity is increased, compared to other geomasks, a predefined level of SKA cannot be guaranteed.

2.3 Adaptive Voronoi Masking (AVM)

AVM extracts the asset of considering the underlying population density by joining polygons as AAE does and displaces the original data based on the concept of VM. In respect thereof, the original data are moved to the closest segment of their corresponding Voronoi polygon which lies within their merged AAE-polygon. In case a Voronoi segment lies outside its dissolved polygon, the point is transferred to the boundary of the merged polygon and not to the edge of the Voronoi polygon. Through that, AVM intends to circumvent the predicament of moving points to a polygon containing a different population threshold thus preserving the *predefined SKA*. Further, the underlying topography is considered by moving points to the closest street intersection that has a higher amount of surrounding buildings than if moved to the nearest segment. Through that, AVM avoids shifting the points directly to another residence causing *false re-identification* but it also prevents the displacement to *invalid locations* such as water bodies or forests.

To execute AVM, the following data sets are required: a) a point file (as needed in VM and AAE), b) a polygon file including risk of re-identification information (as required in AAE), and c) a line file depicting the street network. Firstly, the data is pre-processed as done for AAE. Subsequently, a disclosure threshold for the risk of re-identification field is selected and polygons with a smaller value than the chosen disclosure value are merged with its adjacent polygon until all polygons receive a value that is greater or equal to the set disclosure value. Here, the general spatial rule is applied defining that every polygon is combined with the bordering polygon that has the longest shared border [19].

Secondly, every data point that is lying within a polygon with at least two data points is transferred by the concept of the VM technique. It is guaranteed that the data points are replaced to the closest segment of their corresponding Voronoi polygon within their dissolved polygon. Thirdly, the polygons containing only one data point randomly transfer the data point within their merged polygon. Afterwards, all newly displaced points are shifted to the closest street intersection inside their K-anonymized polygon. Figure 1 shows the outputs of the steps using as an example the city centre of Dresden, Germany.

2.4 ESDA for evaluating geomasking performance

ESDA identifies and characterizes locations, shapes, and magnitudes of statistically substantial patterns within an area of interest [10]. Studies on geoprivacy implemented ESDA methods on original data and masked data to investigate and compare the performance of geomasks. Armstrong et al. [3] scrutinized the effect of geomasks by exploring pair-wise relations, event-geography relations, anisotropies, and trends. Seidl et al. [29] applied the kernel density estimation, global Moran's I, distance to K-nearest neighbour, the cross K-function - also known as Ripley's K-function, and the nearest neighbour hierarchical cluster analysis. Kwan et al. [21] also applied the Ripley's K-function, the kernel density estimation, and examined the visualisation of the point pattern. Leitner and Curtis [22] analysed the visualisation of the point pattern as well. Several approaches exist to analyse the efficiency of geomasks. Here, we use four methods that were already used in previous studies and are described in the next subsections.

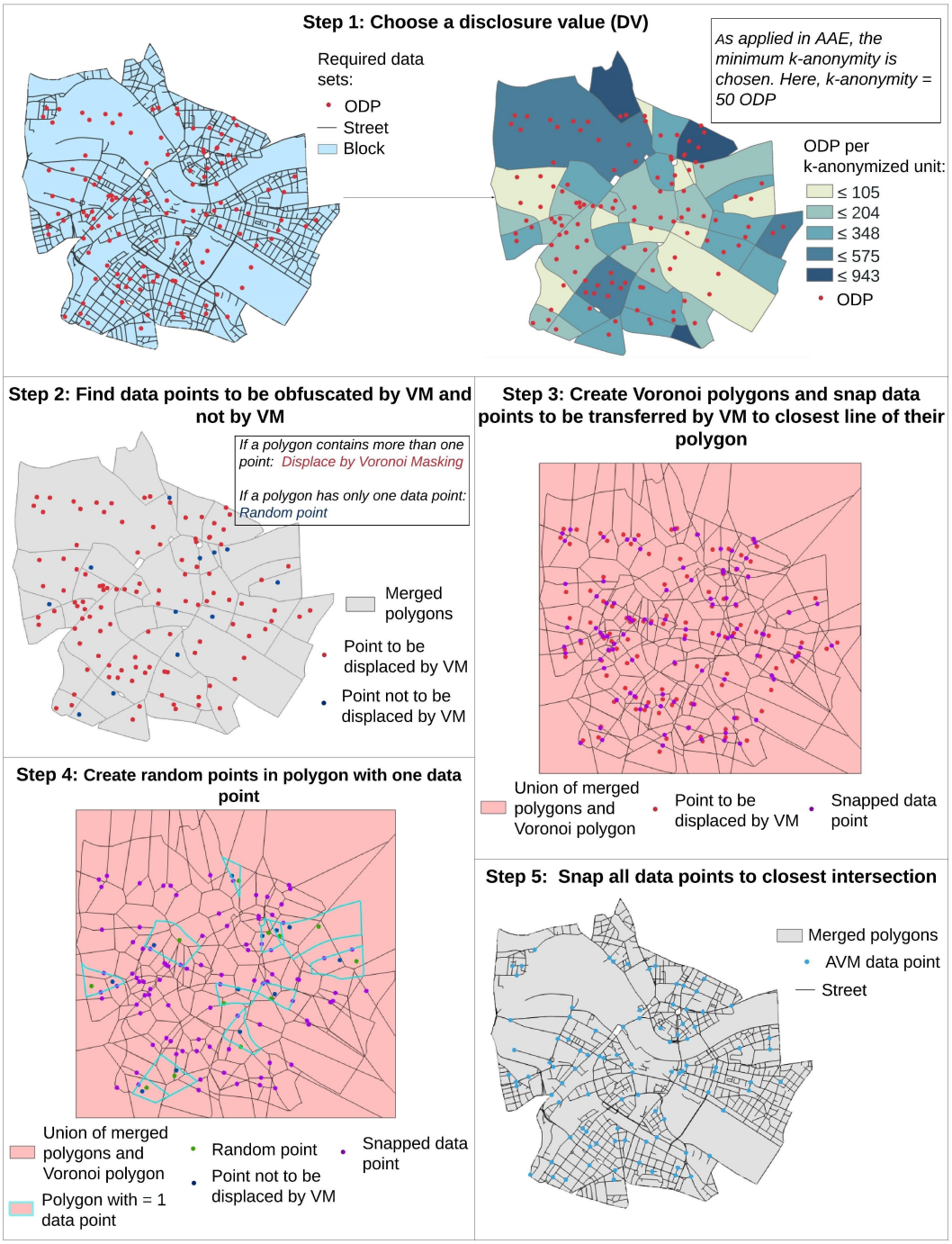


Figure 1 A visualisation of the AVM outputs in the city centre of Dresden.

2.4.1 Visualisation of point pattern

This technique is used to a) scrutinize the extent of the original data and compare it with that of the masked data and b) to investigate whether the masked data are displaced on other residencies increasing the risk of false re-identification or are transferred to void locations such as forests or lakes.

2.4.2 Central tendency

The mean and median centres of the original data and the masked data are compared through their distance's divergence. This has been applied by Seidl et al. [29] and Gupta and Rao [13].

2.4.3 Ripley's K function

Ripley's K-function identifies whether the masked points are clustered, dispersed, or randomly distributed and whether the point distribution between original data and masked data remains linked or not. In the case of linked point distribution, the geomasks perform spatially dependent on the original data. Ripley's K-function conflates spatial dependence regarding point feature scattering or aggregation over a variety of distances [8], which returns a more detailed output than other ESDA pattern detection techniques. By analysing the spatial patterns over several lengths as well as spatial scales, the point patterns alter. Thus, it can reflect how the scattering or aggregating of points centroids shifts when the size of the neighbourhood varies.

2.4.4 Nearest neighbour hierarchical cluster analysis

In most previous studies, the impact of geomasks on original hot spots has been probed. This is important since clustering detection plays a vital role in spatial analysis. For instance, by detecting hot spots, high concentrations of crime incidents can be explored and predicted for future scenarios [6]. Nearest neighbour hierarchical cluster analysis allows examining and comparing the clustering pattern of the original data with the pattern of the masked data regarding amount of clusters, size, orientation, density.

3 Experiments' settings

3.1 Study area

The choice of the study area is based on the availability of processed and free data. Moreover, area data sets must allow different levels of spatial granularity and population density. The chosen area is the Free State of Saxony in Eastern Germany. Saxony has 13 districts containing more than 4 million inhabitants[3], of which more than 563,000 were registered in the state capital Dresden. Yet, the highest population and population density are found in the city of Leipzig with a total of 587,857 people and 1,974 inhabitants per km^2. Contrary to that, the district Nordsachsen has only 97 inhabitants per km^2 - the lowest in Saxony. Hence, the State of Saxony is an explicit choice to investigate the performance of geomasks because it has highly populated as well as rural areas. The geomasks are applied on the State of Saxony, the city of Leipzig because it has the most inhabitants and the highest population density, and the district of Zwickau. Zwickau was chosen because when calculating the average inhabitants (ca. 313,685) and population density (ca. 493/km^2) per district in Saxony, Zwickau has the closest values (inhabitants: 317,531; population density: 334/km^2).

[3] `https://www.statistik.sachsen.de/` *(Last accessed on January 22th*, 2021)

3.2 Data

For the polygon file, a line shapefile representing the street network in Saxony was derived online[4], and used to create streets blocks. It was aimed to develop blocks which are not too coarse but also not too small. Since the original road network file also included several street classes such as "footway", "path", or "cycleway", all street duplicates, as well as all street classes except for "primary", "secondary", and "tertiary", were deleted off the shapefile to generalize the road network. Thus, an enormous amount of small blocks is avoided resulting in shorter processing times of the geomasks. Also, only single-line road features in place of matched pairs of divided road lanes were maintained. Small and open configurations of roads were removed. Next, polygon features were created from the remaining polylines. Any polygons outside the study area were removed.

The same street network dataset was also used for intersection displacement step of AVM. However, in this case, it was more detailed containing also the "residential class". Thus, a smaller but yet meaningful displacement of the data points can be achieved. More street classes such as "footway" or "cycleway" were not included to prevent false re-identification.

Also, additional attributes were included. First, risk of re-identification information needs to be added within the polygons. Hence, a point data set with addresses in Saxony from 2018 was chosen and can be downloaded directly from ESRI[5]. Originally, the point data set consists of 947,164 data points. The points were counted per polygon as the risk of re-identification information. Second, two "sensitive" data sets were created as random subsets from the addresses in Saxony consisting of 2,000, and 200 points for each and thus mimicking different population densities to examine the performance of AVM at different situations. These data sets simulate potential confidential or private discrete data. Third, attributes such as "id" (unique identifier) and "area" (size of a polygon) were added as they are necessary for the algorithms. Finally, the polygons were clipped based on the boundaries of the three study areas. The boundaries were obtained from the *Federal Agency for Cartography and Geodesy*[6]. The clipped study areas do not completely correspond in size with the original district of Zwickau, City of Leipzig, or the Free State of Saxony. This is due to the removal of smaller streets creating somehow different sizes and shapes of the study areas. In the case of referring to a specific study area with a certain number of points, the data sample is named *study area + number of points* (i.e., Saxony 200). Figure 2 shows the resulting area data sets.

3.3 Software

We used ArcGIS Pro 2.5 by the international GIS-software developer ESRI and CrimeStat 3.3. by Levine & Associates (2020). ArcGIS Pro is used for data exploration, visualisation, for running the AAE and DM algorithms, and for the creation of the AVM and VM algorithms. Hereby, the embedded ArcPy Python package was used. The ESDA evaluation methods, with the exception of nearest neighbour hierarchical cluster analysis, were operated in ArcGIS Pro. Nearest neighbour hierarchical cluster analysis was performed in CrimeStat 3.3 (i.e. a program of spatial statistics for exploring locations of crime incidents). CrimeStat can be downloaded for free online[7].

---

[4] `https://download.geofabrik.de/europe/germany/sachsen.html` *(Last accessed on January 22th*, 2021)

[5] `https://opendata-esri-de.opendata.arcgis.com/datasets/esri-de-content::adressen-sachsen` *(Last accessed on January 22th*, 2021)

[6] `https://www.bkg.bund.de/DE/Home/home.html` *(Last accessed on January 22th*, 2021)

[7] `https://www.icpsr.umich.edu/CrimeStat/` *(Last accessed on January 22th*, 2021)

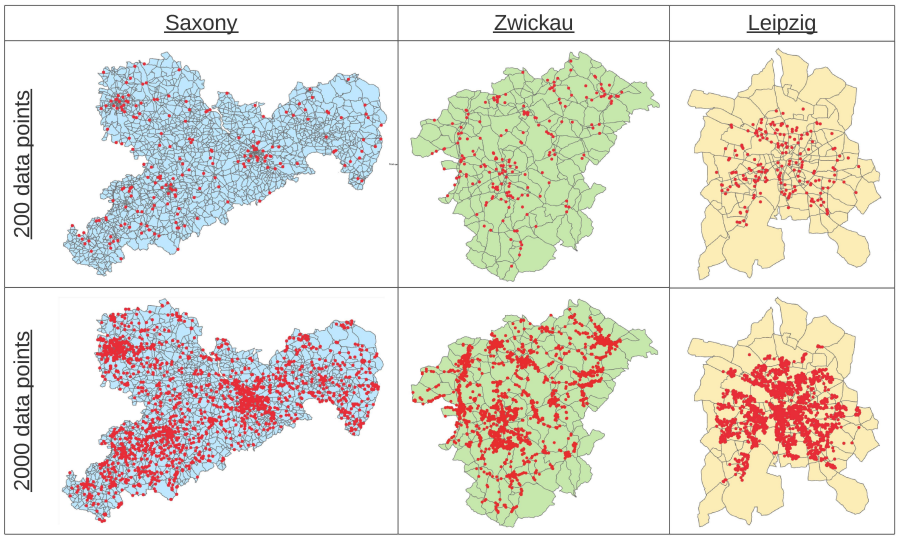


Figure 2 The six area data sets that are used in the study.

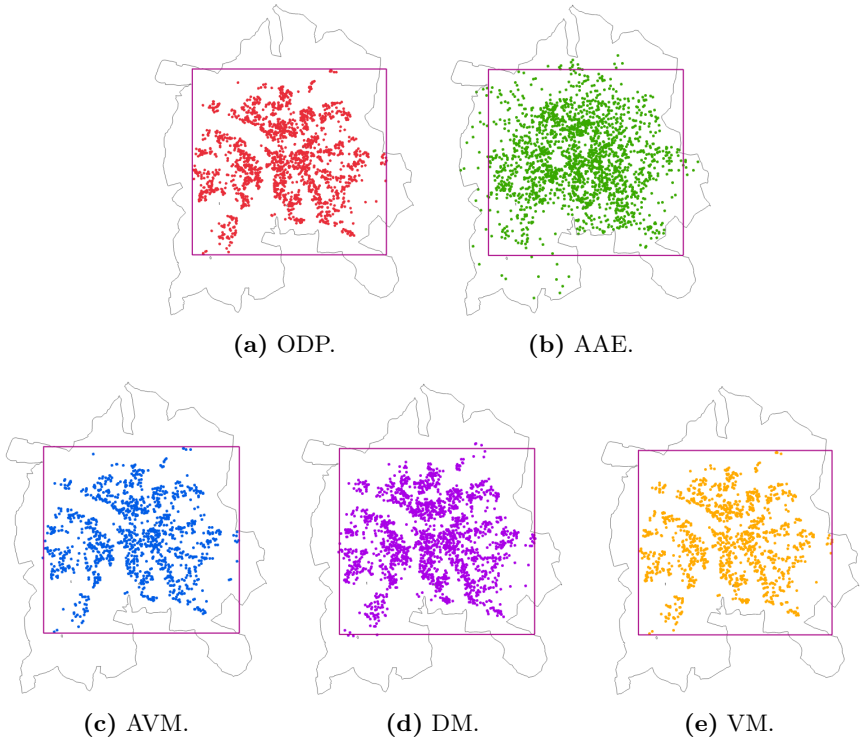
4 Results

AVM, AAE, and DM were applied with a SKA level of 50 addresses. VM is not an adaptive geomask and thus a SKA level cannot be predefined and guaranteed. The four ESDA methods are applied to the original data as well as on the masked data to examine the effects of the geomasks on the original data. [18] and detect dissimilarities of spatial information loss and the preservation of original data granularity [28]. In the ideal case, the spatial analysis of the AVM masked data will be equal to that of the original data.

4.1 Visualisation

Figure 3 shows the extent of the original data for Leipzig 2000. AAE preserves the spatial extent of the original data the least. For Leipzig 2000, 53 points were dislocated outside of the original extent. DM performs more successfully than AAE with only five points not being located within the extent of the original data. The new technique AVM retains the spatial extent more effectively than AAE and DM. For Leipzig 2000 all points were within the extent while in the other area data sets only one to three points were located outside the extent. VM is outperforming the other geomasks regarding the preservation of the original extent. Only one data set (Leipzig 200) has one data point outside the original extent. We created the same maps for each area data set but since the observations are similar other maps are not presented. By ranking the performance of the geomasks, we can see that VM preserves the extend of original data the most (1^{st}), followed by AVM (2^{nd}), then DM (3^{rd}), and last is AAE (4^{th}).

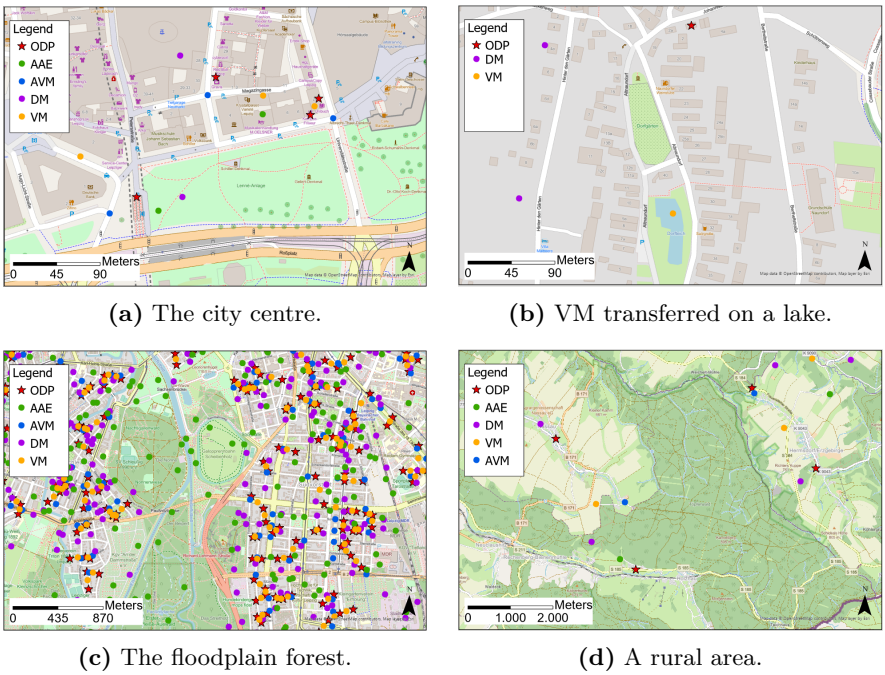
Regarding the preservation of the point pattern, it is perceivable that AAE seems to abandon the pattern the most: Particularly Figure 4b shows that the strongly visible floodplain forest, which intersects the city of Leipzig from northwest to southwest, is not kept by AAE. Unlike that, AVM, DM, and VM maintain this meandering space in most regard while AAE blurs this region completely. Figure 4a is illustrating a park in the city centre of Leipzig. All original data originate on buildings. Every geomask except for AVM displaced data points to either buildings or an uninhabited area as here, the park. Only one



(a) ODP. (b) AAE.

(c) AVM. (d) DM. (e) VM.

Figure 3 The bounding box shows the extent of the original data in Leipzig, while inner points are the masked data.



(a) The city centre. (b) VM transferred on a lake.

(c) The floodplain forest. (d) A rural area.

Figure 4 The results of the geomasks (masked data) compared to the original data in various locations.

VM point had been transferred to a street. Contrary to that, AVM moved all data points to a street intersection. Figure 4b is depicting an area with family homes, a central park, and a lake with a displaced VM data point. Another DM data point had been moved on a house. Figure 4c portrays the floodplain forest in Leipzig. It is noticeable that many points transferred by AAE are laying within this natural habitat which is invalid for a displacement location. Finally, the last Figure 4d show rural areas within Saxony with a low point density. As in the previous examples, AVM points remain on street intersections decreasing the risk of false identification. Again, AAE, DM, and VM moved data points to uninhabited areas, i.e. a forest or grassland. In this analysis, AVM is ranked as 1st in not translating points to illogical locations or false residencies, while all other geomasks follow.

4.2 Central tendency

The displacement distance between the masked and original mean and median centres are shown in table 1. Regarding the mean centres, VM and DM outperform AAE and AVM in most obfuscation settings. DM has the smallest displacement distances for Leipzig 200 (12.71m), Zwickau 200 (4.53m), Zwickau 2000 (3.97m), and Saxony 200 (98.91m). VM has the smallest displacement distances for Leipzig 2000 (1.18m) and Saxony 2000 (9.61 m). The AAE encompasses the furthest displacement distances for all data sets, while AVM performs better yet worse than DM and VM.

Regarding the median centres, VM surpasses the other geomasks. It has a displacement distance of 1.23m to the original median centre of Leipzig 2000, 42.33m to the one of Zwickau 200, and 13.93m to the one of Saxony 2000. DM presents the lowest displacement distances for Leipzig 200 (1.63m) and Saxony 200 (90.82m), while AVM has the smallest displacement for Zwickau 2000 (1.61m). Again, AAE has much greater displacement distances than the other three geomasks (except for Saxony 200).

Finally, it can be seen that the lower the point density, the stronger the variation of the geomasks' results. For instance, the 200 data points in Saxony represent the strongest variations and the highest displacement distances between the masked and original mean and median centres. Contrary to that, Leipzig 2000 has the highest population density and demonstrates the smallest displacement distances and lowest variations between the mean and median centres of the masked data and original data. The geomasks' performance was ranked for each area data set and then the mode of the rank was derived. For the mean displaced distance, DM yields the closest value to the original data mean (1st), followed by VM (2nd), then AVM (3rd), and last is AAE (4th). For the median displaced distance VM yields the closest value to the original data median (1st), followed by DM (2nd), then AVM (3rd), and last is AAE (4th).

4.3 Ripley's K-function

We calculated the expK-values and obsK-values of the original data and masked data of every five bands. The dissimilarity here is calculated as the clustering distance divergence of the masked data from the original data and it is shown in Table 2. For all area data sets and bands, the original data and masked data attain higher obsK-values than the expK-value indicating a strong clustering pattern. A second observation is that dissimilarities are more distinct in smaller bands than in larger bands. The third observation is that VM and AVM create a more clustered pattern than the original (shown by mostly negative divergence values), while AAE and DM tend to create a less clustered pattern than the original (indicated by mostly positive values).

Table 1 Displacement distances (in meters) from the mean/ median centre of the original data to the mean/ median centres of the masked data. The lowest displacement distance per data set is depicted in bold. The method that scores most times the best is DM, followed by VM.

Area data set	AVM		VM		AAE		DM	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Leipzig 200	28,86	28,11	12,72	40,3	129,72	158,17	**12,71**	**1,63**
Leipzig 2000	3,57	5,51	**1,18**	**1,23**	44,24	31,58	3,02	3,12
Zwickau 200	85,71	55,22	23,48	**42,33**	177,18	562,56	**4,53**	100,68
Zwickau 2000	10,11	**1,61**	6,14	3,39	39,46	241,07	**3,97**	8,27
Saxony 200	616,41	1528,4	182,45	231,68	766,97	1357,64	**98,91**	**90,82**
Saxony 2000	27,46	65,52	**9,61**	**13,93**	39,83	66,85	14,39	29,08

Table 2 Ripley's K results on the divergence of the clustering distance (in meters) of the masked data from the original data, from 99 simulations for each area data set and in three bands. The smallest divergence is marked in bold. DM yielded the smallest values most times.

Area data-set	Leipzig 200			Leipzig 2000			Zwickau 200		
Geomask /bands	*1-5*	*6-10*	*11-15*	*1-5*	*6-10*	*11-15*	*1-5*	*6-10*	*11-15*
AVM	-847	-230	64	-361	**-61**	-47	-1650	-750	-342
VM	-1857	-1519	-680	-351	451	**-39**	-3279	-2296	-1775
AAE	1113	664	695	543	858	977	1271	2226	2593
DM	**525**	**77**	**32**	**183**	132	116	**371**	**443**	**272**
Area data-set	**Zwickau 2000**			**Saxony 200**			**Saxony 2000**		
Geomask /bands	*1-5*	*6-10*	*11-15*	*1-5*	*6-10*	*11-15*	*1-5*	*6-10*	*11-15*
AVM	-466	-198	-134	-454	642	1529	**-197**	127	136
VM	**9**	**22**	**-43**	-13443	-10122	-6429	-2138	-1091	-538
AAE	1413	1495	1854	2722	-276	**301**	1835	1383	1090
DM	753	410	379	**-251**	**202**	399	375	**53**	**34**

DM has the most similar obsK-values for Leipzig 200 (all bands), Saxony 200 (bands one to ten), Saxony 2000 (bands six to fifteen), and Zwickau 200 (all bands). AVM scores the most alike obsK-values for Leipzig 2000 (bands six to ten) and Saxony 2000 (bands one to five). VM displays a stronger point accumulation than the original data for data sets with a low point density (Leipzig 200, Zwickau 200, and Saxony 200) and demonstrates the most dissimilar obsK-values than the original data which can be elucidated by the technique's character to alter the point pattern in scattered areas due to greater displacement distances of the established Voronoi polygon. However, VM reaches very similar obsK-values to those of the original data for Leipzig 2000 (bands eleven to fifteen), and Zwickau 2000 (all bands). Again, AAE fares the weakest results.

We derived the geomasks' ranks for each area data set/ band and then the mode rank. DM is ranked 1st, followed by AVM (2nd), then VM (3rd), and last is again AAE (4th).

4.4 Nearest neighbour hierarchical cluster analysis

Table 3 shows the number of clusters, the mean points per cluster, and mean cluster density per m^2 for each original data and masked data and in each area dataset. For Leipzig 200, the original data produced six clusters, with a mean of 7.33 points per cluster. Regarding

the number of clusters, AAE, AVM, and DM are the nearest to the original data value with seven clusters. AVM has the closest mean points per cluster at 7.50 and the most similar cluster density. For Leipzig 2000, VM has the closest number of clusters as the original data (VM: 139; original data: 136). Regarding the mean points, original data contains 7.73 points and DM reaches the nearest measure at 7.83. Also, DM has the most alike density at 0.0000979 m^2 (original data: 0.0001144 m^2).

About Zwickau 200, the original data yielded eight clusters, mean points at 7.75, and a density of 0.0000055 per m^2. Regarding the first metric, DM has the same value. Regarding mean points, VM outperforms the other geomasks at 7.82, whilst DM shows the nearest value for the cluster density at 0.0000060 per m^2. For Zwickau 2000, the original data yielded 110 clusters, mean points of 8.70, and a density of 0.0000655 per m^2. VM has 111 clusters followed by AVM with 112 outperforming the other geomasks. The closest mean point value was obtained by DM at 8.66 as well as for the mean cluster density at 0.0000598 per m^2.

In Saxony 200, four clusters were generated by the original data with mean points at 6.5 and a mean cluster density of zero. DM succeeded the same values as original data whereas VM indicates the most different values. For Saxony 2000, the original data demonstrates 66 clusters, mean points at 7.86, and a mean cluster density at 0.0000042 per m^2. Regarding the first parameter, AVM outperforms the other methods at 63 clusters. Concerning the second parameter, DM reaches the closest mean points at 7.80. Finally, the most alike cluster density was obtained by AVM at 0.0000044 per m^2. AAE fares the worst with regard to the number of clusters and the mean cluster density while VM demonstrates the least efficiency for mean points.

Last, we derived the geomasks' ranks for each area data set/ metric and then the mode rank based on the divergence value (the closer to the original data value the higher is the rank). For both the mean points per cluster and mean cluster density, DM is ranked 1^{st}, followed by AVM (2^{nd}), then VM (3^{rd}), and last is AAE (4^{th}). For the cluster density, both DM and AVM are ranked as 1^{st}, followed by AAE (2^{nd}), then VM (3^{rd}).

4.5 Evaluation and comparison of geomasks

In the ESDA results subsections, we stated the ranking of each geomask. The final ranks are shown in table 4 to indicate the performance regarding data utility. AVM is ranked first for not displacing points to illogical locations or other residencies while VM is ranked first for retaining the extend of the original data. Hence, both geomasks are ranked as first for the visualization ESDA method because there are only these two metrics. The same applies to the central tendency (two metrics: mean and median), while for the nearest neighbour hierarchical cluster analysis we calculated the mode of the three metrics. DM is clearly the geomask that retains the pattern of the masked data the closest to the original one, while AAE distorts the pattern the most. Our proposed AVM method performs also very well and it is ranked as second regarding data utility.

Apart from data utility, this paper discussed the importance of preserving a level of SKA for the derived masked data. Unfortunately, trying to anonymize data sufficiently will eventually decrease their data utility. Also, displacing points to other domiciles should be avoided to prevent false re-identification. Hence, the optimal masking solution is to find the golden mean between these three aspects. These aspects are summarized in table 5, and compared across the geomasks. As stated before, the only method that prevents false re-identification is AVM. DM offers the best data utility, however, it only partially preserves a certain level of SKA because it assumes that the underlying population is homogeneously distributed. Both AVM and AAE retain a certain level of SKA while VM performs the

worst considering all three aspects. By comparison, AVM is the optimal solution because it prevents false re-identification, offers a certain level of SKA, and is ranked second in terms of data utility.

Table 3 Nearest neighbour hierarchical cluster analysis results for each geomask and area data set. The specific metrics are the number of clusters (number), mean points per cluster (points), and mean cluster density (density). Values closer to the original data values are marked in bold. DM has the closest values followed by the AVM.

Geomasks /Metric		**Area data set**					
		Leipzig 200	Leipzig 2000	Zwickau 200	Zwickau 2000	Saxony 200	Saxony 2000
	Number	*6*	*136*	*8*	*110*	*4*	*66*
Original data	*Points*	*7 .33*	*7 .73*	*7 .75*	*8 .7*	*6 .5*	*7 .86*
	Density	*0 .000013*	*0 .0001144*	*0 .0000055*	*0 .0000655*	*0*	*0 .0000042*
	Number	**7**	140	9	112	5	**63**
AVM	Points	**7 .5**	8 .08	7 .44	8 .98	6 .4	7 .68
	Density	**0 .0000113**	0 .0008718	0 .0000177	0 .0004652	**0**	**0.0000044**
	Number	9	**139**	11	**111**	6	71
VM	Points	6 .67	8 .02	**7 .82**	8 .94	6 .7	8 .07
	Density	0 .0000221	0 .0001964	0 .0000077	0 .00010002	0 .0000002	0 .0000055
	Number	**7**	57	5	44	3	36
AAE	Points	6 .43	7 .02	6 .6	8 .25	7	7 .58
	Density	0 .0000169	0 .0000835	0 .0000028	0 .0000511	**0**	0 .000003
	Number	**7**	109	**8**	89	**4**	61
DM	Points	6 .86	**7 .83**	7 .63	**8 .66**	**6 .5**	**7 .8**
	Density	0 .0000207	**0 .0000979**	**0 .000006**	**0 .0000598**	**0**	0 .0000037

Table 4 Ranking of geomasking techniques based on their performance on four ESDA methods (visualisation of point pattern, central tendency, Ripley's K-function, and nearest neighbour hierarchical cluster analysis). DM retains the masked data pattern the most similar to the original data, followed by AVM.

Evaluation Method	**Geomask (rank)**			
	AVM	**VM**	**AAE**	**DM**
Visualization	*1st*	*1st*	4th	3rd
Central tendency	3rd	*1st*	4th	*1st*
Ripley's K-function	2nd	3rd	4th	1st
Nearest neighbour hierarchical cluster analysis	2nd	3rd	4th	1st
Mode Rank	**2nd**	**3rd**	**4th**	**1st**

Table 5 Evaluation of geomasking techniques based on the ability to: a) prevent the risk of false re-identification, b) to ensure spatial K-anonymity, and c) to preserve original point pattern (data utility ranking). AVM offers the best combination of these three aspects (marked in bold).

Geomasks	**False re-identification**	**Spatial K-anonymity**	**Data Utility**
AAE	yes	yes	4th
AVM	**no**	**yes**	**2nd**
DM	yes	partly	1st
VM	yes	no	3rd

5 Conclusion

This study presented a new geographical masking method. AVM (i) considers the underlying population density by defining a level of K-anonymity, as AAE does, (ii) displaces a part of the original data based on the concept of VM, and (iii) by considering the underlying geography transfers points to the closest street intersection. Thus, it decreases the risk of false re-identification immensely and does not relocate data points to illogical positions.

The statistical analyses evidenced that AVM did not perform as well as DM regarding data utility, yet it was ranked as second among the four examined geomasks. Adding to that, it preserves the SKA accurately (AAE does this as well) and is the only method that does not dislocate points to illogical locations and minimizes the risk of false re-identification. However, it can be argued that a map viewer will view fewer data points (due to the street intersection aggregation) influencing the spatial perception of a phenomenon. Contrary to that, DM and VM, as well as AAE, can transfer data points to other residences or parcels increasing the risk of false re-identification. Based on three key factors (spatial K-anonymity, false re-identification, and data utility), it can be concluded that AVM is the most encouraging method in terms of the preservation of data utility and decreasing the risk of false re-identification to protect the individual's privacy.

Still, our method is not free of constraints (just like any geomasking method). For example, it might be a better approach to visualize a protected version of the distribution of a point pattern, but it will be less accurate in detecting local patterns compared to DM. Even more, it is a technique that can be successfully applied to confidential spatial data points but not to other geodata types. Location-enabled technologies capture geodata that are more complex and have to be treated/protected by different methods and privacy metrics [16, 20]. For instance, social media data capture, among other attributes, the spatiotemporal stamps of a user, which could be further processed to infer more than one type of spatial information (e.g., home or work locations). The evaluation of a method's efficiency regarding protection for this type of geodata should involve other measures and possibly be diversified by types of spatial information [11].

For the quality or information loss of masked data we applied four ESDA methods. Still, more methods can be implemented such as the global Moran's I for spatial autocorrelation or distance to K-nearest neighbour, as well as Local Indicators of Spatial Association. In addition, it is of great interest to examine the performance of AVM on national data sets. Furthermore, it is recommended to juxtapose AVM with more geomasks that were not applied here to gather more knowledge about the new approach.

Researchers and the public are becoming more aware of the privacy risks related to geodata. However, privacy guidelines as established by Kounadi and Resch [20] as well as the existence of geomasks have to become more well-known to researchers, institutions, companies, or the public sector. A first step to reach this goal is to make geomasks accessible and reproducible. During this research, it was discovered that only the geomask DM is retrievable online for free. This is confounding considering the fact that many researchers stress to mask confidential discrete spatial data. Our method is available for free via the Github repository "Geoprivacy"[8]. A further step is to employ geomasks for open-source software. Through that, companies, researchers, and institutions can share their data and findings with the public without jeopardizing individual privacy.

[8] `https://github.com/okounadi/Geoprivacy`

References

1 Jayakrishnan Ajayakumar, Andrew J Curtis, and Jacqueline Curtis. Addressing the data guardian and geospatial scientist collaborator dilemma: how to share health records for spatial analysis while maintaining patient confidentiality. *International Journal of Health Geographics*, 18(1):1–12, 2019.

2 William B Allshouse, Molly K Fitch, Kristen H Hampton, Dionne C Gesink, Irene A Doherty, Peter A Leone, Marc L Serre, and William C Miller. Geomasking sensitive health data and privacy protection: an evaluation using an e911 database. *Geocarto international*, 25(6):443–452, 2010.

3 Marc P Armstrong, Gerard Rushton, and Dale L Zimmerman. Geographically masking health data to preserve confidentiality. *Statistics in medicine*, 18(5):497–525, 1999.

4 John S Brownstein, Christopher A Cassa, and Kenneth D Mandl. No place to hide—reverse identification of patients from published maps. *New England Journal of Medicine*, 355(16):1741–1742, 2006.

5 Christopher A Cassa, Shaun J Grannis, J Marc Overhage, and Kenneth D Mandl. A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *Journal of the American Medical Informatics Association*, 13(2):160–165, 2006.

6 Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal*, 21(1-2):4–28, 2008.

7 National Research Council et al. *Putting people on the map: Protecting confidentiality with linked social-spatial data*. National Academies Press, 2007.

8 Philip M Dixon. R ipley's k function. *Wiley StatsRef: Statistics Reference Online*, 2014.

9 Matt Duckham and Lars Kulik. Location privacy and location-aware computing. *Dynamic & mobile GIS: investigating change in space and time*, 3:35–51, 2006.

10 Weijung J Fu, Peikun K Jiang, Guomo M Zhou, and Keli L Zhao. Using moran's i and gis to study the spatial pattern of forest litter carbon density in a subtropical region of southeastern china. *Biogeosciences*, 11(8):2401, 2014.

11 Song Gao, Jinmeng Rao, Xinyi Liu, Yuhao Kang, Qunying Huang, and Joseph App. Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of twitter users. *Journal of Spatial Information Science*, 2019(19):105–129, 2019.

12 Christopher Graham. Anonymisation: managing data protection risk code of practice. *Information Commissioner's Office*, 2012.

13 Ruchika Gupta and Udai Pratap Rao. Preserving location privacy using three layer rdv masking in geocoded published discrete point data. *World Wide Web*, 23(1):175–206, 2020.

14 Danielle F Haley, Stephen A Matthews, Hannah LF Cooper, Regine Haardörfer, Adaora A Adimora, Gina M Wingood, and Michael R Kramer. Confidentiality considerations for use of social-spatial data on the social determinants of health: Sexual and reproductive health case study. *Social Science & Medicine*, 166:49–56, 2016.

15 Kristen H Hampton, Molly K Fitch, William B Allshouse, Irene A Doherty, Dionne C Gesink, Peter A Leone, Marc L Serre, and William C Miller. Mapping health data: improved privacy protection with donut method geomasking. *American journal of epidemiology*, 172(9):1062–1069, 2010.

16 Carsten Keßler and Grant McKenzie. A geoprivacy manifesto. *Transactions in GIS*, 22(1):3–19, 2018.

17 Ourania Kounadi and Michael Leitner. Why does geoprivacy matter? the scientific publication of confidential data presented on maps. *Journal of Empirical Research on Human Research Ethics*, 9(4):34–45, 2014.

18 Ourania Kounadi and Michael Leitner. Spatial information divergence: Using global and local indices to compare geographical masks applied to crime data. *Transactions in GIS*, 19(5):737–757, 2015.

19 Ourania Kounadi and Michael Leitner. Adaptive areal elimination (aae): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, 57:59–67, 2016.

20 Ourania Kounadi and Bernd Resch. A geoprivacy by design guideline for research campaigns that use participatory sensing data. *Journal of Empirical Research on Human Research Ethics*, 13(3):203–222, 2018.

21 Mei-Po Kwan, Irene Casas, and Ben Schmitz. Protection of geoprivacy and accuracy of spatial information: How effective are geographical masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*, 39(2):15–28, 2004.

22 Michael Leitner and Andrew Curtis. Cartographic guidelines for geographically masking the locations of confidential point data. *Cartographic Perspectives*, (49):22–39, 2004.

23 Gerard Rushton, Marc P Armstrong, Josephine Gittler, Barry R Greene, Claire E Pavlik, Michele M West, and Dale L Zimmerman. *Geocoding health data: the use of geographic codes in cancer prevention and control, research and practice*. CRC Press, 2007.

24 Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

25 Bill Schilit, Jason Hong, and Marco Gruteser. Wireless location privacy protection. *Computer*, 36(12):135–137, 2003.

26 Klaus Schwab, Alan Marcus, JO Oyola, William Hoffman, and Michele Luzi. Personal data: The emergence of a new asset class. In *An Initiative of the World Economic Forum*, 2011.

27 Dara E Seidl, Piotr Jankowski, and Keith C Clarke. Privacy and false identification risk in geomasking techniques. *Geographical Analysis*, 50(3):280–297, 2018.

28 Dara E Seidl, Piotr Jankowski, and Atsushi Nara. An empirical test of household identification risk in geomasked maps. *Cartography and Geographic Information Science*, 46(6):475–488, 2019.

29 Dara E Seidl, Gernot Paulus, Piotr Jankowski, and Melanie Regenfelder. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, 63:253–263, 2015.

30 Paul A Zandbergen. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in medicine*, 2014, 2014.

31 Su Zhang, Scott M Freundschuh, Kate Lenzer, and Paul A Zandbergen. The location swapping method for geomasking. *Cartography and Geographic Information Science*, 44(1):22–34, 2017.