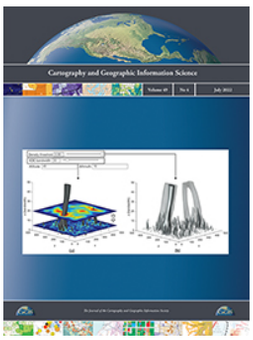

Taylor & Francis
Taylor & Francis Group

Cartography and Geographic Information Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tcag20

An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data

Jue Wang, Junghwan Kim & Mei-Po Kwan

To cite this article: Jue Wang, Junghwan Kim & Mei-Po Kwan (2022): An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data, Cartography and Geographic Information Science, DOI: 10.1080/15230406.2022.2056510

To link to this article: https://doi.org/10.1080/15230406.2022.2056510

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE
https://doi.org/10.1080/15230406.2022.2056510

Taylor & Francis
Taylor & Francis Group

An exploratory assessment of the effectiveness of geomasking methods on privacy protection and analytical accuracy for individual-level geospatial data

Jue Wang [a], Junghwan Kim [b] and Mei-Po Kwan [c,d]

[a]Department of Geography, Geomatics and Environment, University of Toronto – Mississauga, ON, Canada; [b]Center for Geographic Analysis, Institute for Quantitative Social Science, Harvard University, MA, USA; [c]Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China; [d]Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, China

ABSTRACT

The widespread use of personal geospatial data raises serious geoprivacy concerns for sharing these data, which may limit the reproducibility of research findings. One widely used method for securely sharing confidential geospatial information is applying geomasking techniques before sharing. Geomasking may reduce the usability of the data. Thus, researchers need to strike a balance between privacy protection and analytical accuracy. Although many geomasking methods have been proposed, there is no systematic evaluation of these methods or guidance on which method to use and how to apply it properly. To address this gap, we evaluate eight geomasking methods with simulated geospatial data with various spatial patterns and investigate their performance on privacy protection and analytical accuracy. We propose not only a set of preliminary guidelines for applying the proper geomasking methods when using different spatial analysis methods but also an evaluation framework for assessing geomasking methods for other spatial analysis methods. The findings will help researchers to properly apply geomasking for sensitive geospatial data and thus promote data sharing and interdisciplinary collaboration while protecting personal geoprivacy.

ARTICLE HISTORY
Received 26 August 2021
Accepted 18 March 2022

KEYWORDS
Geomasking; geoprivacy; data sharing; data confidentiality; analytical accuracy

1. Background

With the widespread use of personal geospatial data in GIScience and other fields (e.g. health geography) in recent decades, geoprivacy has become a major concern (Sherman & Fetters, 2007). Personal geospatial data were collected and analyzed with advanced geospatial methods in many research projects (Chaix et al., 2016; Wang & Kwan, 2018; Yoo et al., 2015). These data provide detailed information on the private locations of individuals, including where people live, work, and undertake other daily activities. Moreover, recent advances in geospatial technologies, such as wearable sensors integrated with global positioning systems (GPS), enable researchers to collect richer and more detailed personal location information (Boulos et al., 2019; Fuller et al., 2017; Kwan, 2012). The detailed geospatial information allowed for exciting new research findings, but the use of personal geospatial information also puts data contributors (e.g. research participants) at risk of being identified, especially when sensitive location data (e.g. patients' home locations) is involved (Brownstein et al., 2006; Curtis et al., 2006; Kim et al., 2021). In addition, when linked to other data sources via location details, personal information may be misused and individual privacy may be breached, which poses potential risks to data contributors (Kounadi & Leitner, 2014; Kounadi et al., 2018; VanWey et al., 2005).

In the context of public health, geospatial data related to personal health information is normally protected by government regulations (e.g. the Health Insurance Portability and Accountability Act [HIPAA] in the US). To avoid the potential leak of protected health information (PHI), the HIPAA regulates 18 identifiers of PHI for geographic scales smaller than the state level, where data cannot be shared or published if not de-identified (Delmelle et al., 2022; Tellman et al., 2010). Additionally, processing PHI may involve geocoding addresses or performing spatial analysis via online platforms. The sensitive data uploaded to the online geocoding servers may breach the confidentiality rules (Duncan et al., 2012). Although the geocoder in use may be inherently poor in accuracy and introduce error in the geocoded addresses (Owusu et al., 2020, 2017) and thus hide the true location, the errors are systematic and not random, which means reverse engineering could potentially re-identify the original address.

CONTACT Jue Wang gis.wang@utoronto.ca
Supplemental data for this article can be accessed here.

To protect the geoprivacy of data contributors, personal geospatial information collected in one research project may not be shared with others. This impedes data sharing in the research community and consumes invaluable resources for repetitive data collection (Wang & Kwan, 2020). Further, the ability to reproduce research outcomes – reproducibility, which is the cornerstone of the scientific paradigm (McNutt, 2014) – is limited by the difficulties of sharing personal geospatial information in geographic studies. Reproducibility, defined as "obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis" (National Academies of Sciences, 2019), promotes robustness as well as the generalizability of research results (McNutt, 2014; Richardson et al., 2015). In this light, for researchers who utilize sensitive geospatial data, sharing geospatial datasets faces more challenges because of geoprivacy (Boulos et al., 2019; Curtis et al., 2011; Fuller et al., 2017; Gutmann et al., 2008; VanWey et al., 2005).

Significant efforts have been made in securely sharing personal geospatial information (Armstrong & Ruggles, 2005; Duncan & Pearson, 1991; Kwan et al., 2004; Richardson et al., 2015). In GIScience and related fields, spatial anonymization of address points is an essential way to share sensitive geospatial data while protecting geoprivacy (Charleux & Schofield, 2020). One widely recognized method is to apply geographic masking (or geomasking) techniques on the geospatial data before sharing (Allshouse et al., 2010; Armstrong et al., 1999). Geomasking methods relocate individual geographic locations in the original data to other locations by adding a controlled level of noise, which masked the personal geospatial information by reducing the accuracy to a certain degree while retaining the usability of the data for research. "The goal is to provide individuals, institutions, and public health authorities a comfort level with the sharing of skewed, and hence, anonymized data, rather than using raw, fully identifiable data" (Cassa et al., 2006). Although many geomasking methods have been proposed (Hampton et al., 2010; Lu et al., 2012; Zhang et al., 2017), no previous study has evaluated these methods or provided guidance on which method to use and how to apply them properly.

Further, applying geomasking may reduce the usability of the data for research purposes. Thus, researchers need to strike a balance between privacy protection and analytical accuracy (Carr et al., 2014; Kwan et al., 2004; Nissenbaum, 2009). For example, when applying random perturbation geomasking methods, the level of confidentiality or protection increases when a larger radius is applied. At the same time, however, the analytical accuracy decreases because the spatial pattern of the original points can be distorted due to the error introduced, which implies that analytical accuracy has a negative relationship with the level of data confidentiality. Moreover, since different geomasking methods operate in different ways, there may be particular patterns of the trade-off relationship for specific geomasking methods, which is worth further investigation. There are currently two major hurdles to getting scholars or data managers to actually use geomasking methods: the difficulty or burden of applying geomasks, and the lack of guidance on which method to use. Some scholars have recently tried to mitigate the first hurdle by developing more accessible tools (e.g. Charleux & Schofield, 2020; Swanlund, Schuurman, et al., 2020a) and providing practical privacy-preserving steps for the collection, storage, analysis, and dissemination of spatiotemporal participatory sensing data (Kounadi & Resch, 2018), but there is no guidance to date that could inform people outside the field to make appropriate geomasking decisions.

To address this research gap, this study addresses the second hurdle and seeks to provide guidelines for applying geomasking by evaluating the performance of several widely used geomasking methods while considering the balance between data confidentiality and analytical accuracy. Specifically, this research aims to 1) understand the effectiveness of geomasking methods, and 2) explore the trade-off patterns between privacy protection and analytical accuracy, thus 3) provide preliminary guidelines on applying geomasking methods. In this study, we propose not only a set of guidelines for choosing proper geomasking methods for five widely used spatial analysis methods but also an evaluation framework for assessing geomasking methods for other spatial analysis methods. It will help researchers in different fields to properly apply geomasking for sensitive geospatial data and thus promote data sharing and collaboration among the disciplines while protecting personal privacy.

2. Geomasking methods

This section introduces the existing geomasking methods to protect geoprivacy when using spatial datasets collected at the individual level. Geomasking methods can be classified into three broad categories according to two characteristics (i.e. whether the method preserves the number of records and whether the method randomly relocates records): aggregation, affine transformation, and random perturbation (Armstrong et al., 1999; Kwan et al., 2004; Zandbergen, 2014). Figure 1 illustrates how existing geomasking methods can be classified into these three categories. Affine transformation and random perturbation preserve both the number of records and the data type (e.g. point), but aggregation preserves only one of these because it either aggregates the records of the original

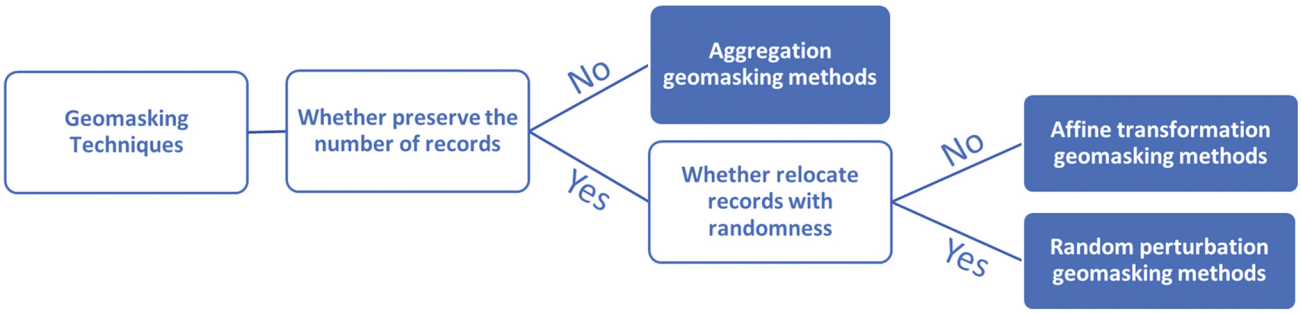

Figure 1. Classification of geomasking methods for geospatial datasets that are measured at an individual level.

geospatial data into a smaller number of records in the same data type or uses a lower spatial resolution and turn the original data into a different data type (e.g. from points to polygons or raster cells). Considering whether a method randomly relocates records, geomasking methods can be further distinguished: affine transformation deterministically relocates original points, while random perturbation does so stochastically. For affine transformation, the coordinates of the original records can be easily recovered from the geomasked dataset if the intruder knows the transformation matrix. Thus, this method is not widely used. Please also note other recently developed geomasking algorithms, such as the Voronoi masking (Seidl et al., 2015), adaptive areal elimination (Kounadi & Leitner, 2016), adaptive areal masking (Charleux & Schofield, 2020), street masking (Swanlund, Schuurman, et al., 2020b), and the Military Grid Reference System (MGRS) masking (Clarke, 2016). As the first exploratory study seeking to develop guidance on geomasking, this research focuses on relatively conventional geomasking methods, such as aggregation and random perturbation methods, which have been utilized by application studies (e.g. Clifton & Gehrke, 2013; Curtis et al., 2011; Kim & Kwan, 2021). However, it is worth mentioning that traditional geomasking methods do not always result in optimized results. Thus, future studies can benefit from examining the performance of new geomasking methods and comparing it with that of traditional geomasking methods.

2.1. Aggregation geomasking methods

Aggregation combines location records and assigns the aggregated attributes to a unit at a lower spatial resolution, such as the census unit or administrative area (Armstrong et al., 1999). It aggregates individual points into a smaller number of points or polygons so that the true locations of the original geographic records are hidden. Those polygons can also be the unit cell of a uniform raster (square or hexagonal grid cells) covering the study area. In the tests of this study, we implement the point aggregation geomasking method by aggregating points to the nearest centroids of grid cells since the simulated dataset does not contain any administrative boundaries. Aggregation geomasking methods thus reduce disclosure risk by summarizing or averaging individual records into a coarser resolution. However, the application of these methods leads to the loss of the precise location of the original data and thus may reduce the accuracy of analytical methods or even diminish the ability to detect spatial clusters (Cassa et al., 2006).

2.2. Random perturbation geomasking methods

Random perturbation relocates each geospatial record in a dataset to a new location by introducing a random spatial displacement (Armstrong et al., 1999). The random spatial displacement is calculated one by one for each record independently. Assume that one original record is located at (x_{old}, y_{old}) in a planar coordinate system, the relocated new point (x_{new}, y_{new}) can be calculated by Equation 1:

$$(x_{new}, y_{new}) = (x_{old} + d_x, y_{old} + d_y) \quad (1)$$

where d_x and d_y denote a spatial displacement introduced by a random perturbation method in the x- and y-axes. There are several types of random perturbation methods that use different approaches to determine d_x and d_y. In general, these methods can be classified into three categories depending on how the spatial displacements (d_x and d_y) are generated.

(a) Naïve random perturbation: This type of method relocates a geospatial record to a random location with the same probability to any location within a region or on a circle around the original location. These methods (see, Figure A1 in Supplementary Materials) include random perturbation within a circle (Armstrong et al., 1999), random direction with a fixed radius (Zandbergen, 2014), and random perturbation within an annulus (donut masking; Hampton et al., 2010; Stinchcomb, 2004) centered at the original location.

(b) Random perturbation with distribution functions: Instead of calculating random spatial displacements with the same probability in the naïve random perturbation,

distribution functions can be used to determine the random spatial displacement. Gaussian displacement and bimodal Gaussian displacement are two examples of random perturbation with distribution function methods (Zandbergen, 2014). Gaussian displacement employs a Gaussian distribution function to calculate the random spatial displacement, while bimodal Gaussian displacement uses the bimodal Gaussian distribution function instead of the Gaussian distribution to calculate the random spatial displacement (see, Figure A2 in Supplementary Materials). Compared to Gaussian distribution, the bimodal Gaussian distribution shows a more complex distribution pattern of probability density.

(c) Random perturbation with pre-set potential locations (also called location swapping): This type of geomasking method considers the land-use patterns of the study area (Zhang et al., 2017). Different from the methods mentioned above, which may relocate the original records (subjects' home addresses) to unreasonable locations (e.g. a random location in the middle of a lake or other non-residential area), location swapping methods relocates the original records to other potential locations that share similar geographic characteristics (i.e. nearby residential locations). In summary, the location swapping methods relocate an original point to a randomly selected location out of all the potential locations of a similar type within a region around the original location (see, Figure A3 in Supplementary Materials). The region can be defined as a circle (location swapping within a circle) or an annulus (location swapping within an annulus). One of the major advantages of both methods is the potential that they can adaptively adjust the radii necessary to guarantee a minimum spatial k-anonymity if needed.

3. Methods

To assess the effectiveness of the geomasking methods described in the previous section and provide guidelines on using them to protect geoprivacy, we evaluate their performance with regard to analytical accuracy, data confidentiality, and their trade-off relationships. In this exploratory research, we evaluated geomasking methods applied to geospatial datasets that are measured at the individual level (e.g. residential location of patience).

3.1. Analytical accuracy and confidentiality measurement

Analytical accuracy represents the accuracy of the spatial information (e.g. spatial pattern) in the geomasked dataset compared to the original dataset. Since geomasking methods introduce spatial error (noise) into the geospatial dataset to reduce disclosure risk, it is unavoidable that the analytical accuracy of the geomasked data decreases, and the results generated from geomasked data may be different from those generated from the unmasked data. Thus, it is critical to understand how different geomasking methods with various settings affect the accuracy of the results (compared to the results generated using unmasked data).

In this study, the data accuracy of various geomasking methods is assessed by comparing the degree of difference in the results of five widely-used spatial analysis methods based on original and geomasked data. These methods include the average nearest neighbor index (ANN), the minimum convex polygon (MCP), the standard deviation ellipse (a directional distribution with 1 standard deviation; SDE), kernel density estimation (KDE), and point density estimation (PDE). The ANN calculates the average distance of all points in a dataset to their nearest neighbor, which is used to measure the spatial pattern of a dataset. The MCP and SDE are widely used to represent the spatial distribution of features in a dataset or to generate activity space in human mobility studies. The KDE and the PDE are ways to create density surfaces or maps based on point feature datasets to represent spatial patterns.

Data confidentiality assesses how well the original geospatial records are effectively protected from re-identification risk. Applying a suitable and effective geomasking method can reduce the risk that intruders re-identify the true location of geomasked records. Spatial k-anonymity is a widely used metric for measuring the disclosure risk of a geomasked dataset (Zhang et al., 2017). It is an extension of k-anonymity, which measures the possibility of a certain record being uniquely identified among all the records in a dataset (Sweeney, 2002). The value k indicates the number of records that share similar attributes in a dataset. The smaller the k value, the more likely a record can be distinguished among all other records in the dataset. Similarly, spatial k-anonymity calculates the number of potential geospatial records in the anonymizing spatial region around one record (Ghinita et al., 2010). The anonymizing region is generated as a buffer with a radius of r and centered at the location to be masked or relocated. The radius is the distance between the original location and the relocated location (see, Figure A4 in Supplementary Materials). Thus, spatial k-anonymity can be used to evaluate the performance of geomasking methods: a larger k value indicates a lower probability of re-identifying the original location, implying a better performance of the geomasking method. It is worth noting that calculating spatial k-anonymity needs the residential locations of the general population in the study area. However, there are many ways to generate such dataset with publicly available data

(e.g. local government open data, OpenStreetMap by overlapping the point of interests data or building footprint data with a land-use map to abstract the residential locations in the study area). If the detailed residential location data is unavailable in anyway, though not ideal, aggregated reference data (e.g. census units/administration areas with population information) can also be used to calculate the spatial k-anonymity (Allshouse et al., 2010; Kounadi & Leitner, 2016).

We acknowledge there are other recently developed data confidentiality assessment methods, such as l-diversity (Machanavajjhala et al., 2007), t-closeness (Li et al., 2007). Though these methods may have advantages in privacy assessment, spatial k-anonymity, considering the spatial component in geospatial data, is still one of the widely used measurements for geoprivacy (Charleux & Schofield, 2020). Consistent with previous research, in this exploratory study, we evaluate data confidentiality by spatial k-anonymity (Ghinita et al., 2010; Sweeney, 2002).

3.2. Data Simulation

A large volume of geospatial data was collected in urban areas. Taking the United States as an example, 80.7% of the population is living within urban areas (United States Census Bureau, 2021). Thus, in the setting of an urban area, this exploratory study generates 100 points of *sensitive locations* that need to be geomasked (simulating the subjects' home locations from a sensitive dataset, such as AIDS patients) and 1,000 points of *simulated residential locations* (simulating other home locations besides the sensitive ones in the same hypothetical study area). These two sets are named the sensitive locations and residential locations respectively hereafter. It is noteworthy to mention that the arbitrary ratio of 1/10 between sensitive and residential locations used in the study is because the ratio in the real world is uncertain. Moreover, if this ratio in the real world is much smaller (e.g. 1/100 – 1/1,000), our study thus used a ratio that represents a much greater risk of disclosure than what is commonly encountered in real-world situations. This is meaningful because our focus would thus be on situations with higher risks of disclosure. The simulations are conducted within a hypothetical study area with a size of 5 km by 4 km, with an area of 20 km^2 (about the area of a circular buffer region of 30-minutes walking distance).

The geomasking methods relocate individual geographic locations in the original data to other locations while spatial k-anonymity calculates the number of potential geospatial records around each geomasked record, so the performance of the geomasking methods may be influenced by the specific spatial patterns of both the sensitive and residential locations. Three types of point patterns are considered for the simulated sensitive and residential locations: random, regular, and clustered (Figure 2), following a general approach of point pattern and urban form analysis (e.g. Bivand et al., 2008; Lu et al., 2008; Marshall & Garrick, 2010). For example, study areas with regular residential location patterns include high-density urban areas, such as Downtown Chicago and Manhattan Island in New York City, where the urban form has a regular pattern, such as rectangular-shaped blocks. Additionally, to understand how geomasking methods perform with clustered or random point patterns (that may capture most of the real-world scenarios, such as suburban areas in the U.S. context), we also analyze the performance with a regular point pattern as a baseline.

As a result of different combinations of random, regular, and clustered residential or sensitive locations, there are 9 different simulation scenarios (Table 1). For example, rdR-rdS means the scenario with random residential locations and random sensitive locations, while rgR-ctS stands for the scenario with regular residential locations and clustered sensitive locations.

Figure 2 shows the simulated point datasets of residential locations and sensitive locations in the random, regular, and clustered spatial patterns. To generate the points with a random pattern, the "Create Random Points" tool in ArcGIS 10.7 was employed. This tool creates the specified number of random points within the defined rectangular area of the study (Environmental Systems Research Institute, 2022). For the clustered patterns, the points were generated in ArcGIS with statistically significant clusters, as shown in Figures 2 (c) and 2 (f). The regular pattern point datasets are generated so that all the points are evenly located in the study area. The statistical significance of the clustered and random datasets was assessed by calculating the K-function value and conducting Monte-Carlo simulations (999 permutations). The results indicate that both the residential locations and sensitive locations have significant clustered patterns (p <0.05) for the clustered point datasets and significant random patterns (complete spatial randomness; p <0.05) for the random point datasets.

3.3. Evaluation of geomasking methods

Based on these simulated point datasets, the performance of different geomasking methods is evaluated with respect to both data confidentiality and data accuracy. We test eight geomasking methods, including random perturbation within a circle (RPC), random direction with a fixed radius (RDF), random perturbation within an annulus (RPA), Gaussian displacement (GD), bimodal Gaussian

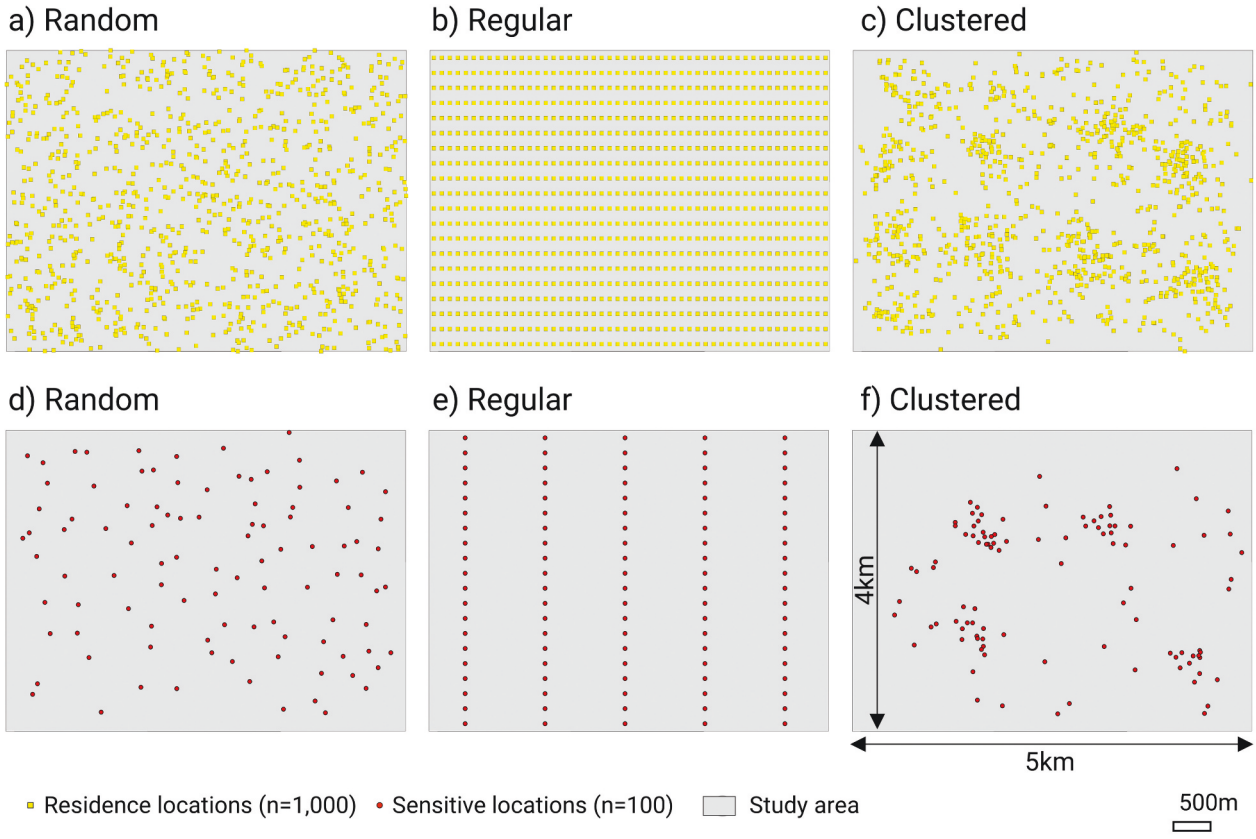

Figure 2. The simulated point datasets: a) random residential locations, b) regular residential locations, c) clustered residential locations, d) random sensitive locations, e) regular sensitive locations, f) clustered sensitive locations.

Table 1. The 9 different simulation scenarios with the residential and sensitive locations in various spatial patterns.

		Sensitive location pattern (S)		
		Random (rdS)	Regular (rgS)	Clustered (ctS)
Residential location pattern (R)	Random (rdR)	rdR-rdS	rdR-rgS	rdR-ctS
	Regular (rgR)	rgR-rdS	rgR-rgS	rgR-ctS
	Clustered (ctR)	ctR-rdS	ctR-rgS	ctR-ctS

displacement (BGD), location swapping within a circle (LSC), location swapping within an annulus (LSA), and point aggregation (PA). Each of these geomasking methods is evaluated using ten different radii: 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 meters.

To assess the levels of data confidentiality achieved by different geomasking methods, we calculate and compare the average spatial k-anonymity of all geomasked points for a given geomasking method and its parameter setting (radius). $K_{m,r}$, the level of confidentiality achieved by applying geomasking method m with a radius of r is calculated by Equation 2.

$$K_{m,r} = \frac{\sum_{p=1}^{n} k_p}{n} \quad (2)$$

where n denotes the number of geomasked points. For the simulated point datasets, n equals 100, which is the number of sensitive locations. However, n can be different when the point aggregation method is applied because it aggregates (summarizes) multiple points into one point. p indicates each geomasked sensitive location in the dataset, and k_p indicates the spatial k-anonymity value calculated for point p. k_p is estimated by counting the number of potential residential locations that are closer to the masked location p than the distance between the masked and the original locations (also illustrated in Figure A4). A higher value of $K_{m,r}$ indicates a higher level of confidentiality (i.e. geoprivacy protection).

For the five widely used spatial analysis methods being evaluated, the default parameter settings are used when implementing these methods in ArcGIS. However, for the KDE and the PDE, we use 50 m as the cell size and 500 m as the bandwidth. For each spatial analysis method, we examine to what extent the analytical accuracy of the original data is changed by applying each geomasking method with a specific radius setting. A high level of error introduced by a geomasking method indicates reduced analytical

accuracy. Specifically, for the ANN, we calculate the relative error of the index generated from using the geomasked data. For the MCP and SDE, we compute the ratio of the area of the shape (e.g. polygon or ellipse) that is not preserved after applying a geomasking method to the area of the original shape. For the KDE and PDE, we calculate the Root Mean Square Error (RMSE) to represent the level of error introduced by a geomasking method.

To sum up, for the eight geomasking methods implemented with ten different radii, we first calculate the average spatial k-anonymity ($K_{m,r}$) that indicates the level of data confidentiality. A higher value of $K_{m,r}$ indicates a higher level of confidentiality. Second, we calculate the analytical accuracy of five spatial analysis methods for geomasking methods with various radii. These two steps are repeated for the 9 different simulation scenarios with the residential and sensitive locations in various spatial patterns. In addition, the entire evaluation process is independently repeated ten times to ensure the reliability of the results.

4. Results

4.1. Evaluation based on the average spatial k-anonymity

The average spatial k-anonymity of the sensitive locations geomasked by various methods with different radius settings are compared in the nine simulated scenarios (Figure 3). Not surprisingly, spatial k-anonymity increases (higher confidentiality level) as the masking radius increases. However, the growth rates are not the same for different methods. Interestingly, the point aggregation method breaks away from others when longer radii are used ($r \geq 300m$ in the experiment). Point aggregation (PA) achieves lower confidentiality compared to other methods in all simulated scenarios. The differences in spatial k-anonymity between aggregation and other geomasking methods are the largest for the scenario rgS, followed by ctS and rdS.

In addition, the spatial k-anonymity of most geomasking methods (except point aggregation) is not notably affected by the pattern of the sensitive locations while affected by the residential location pattern when shorter masking radii are used. In the scenarios of rgR, the location swapping methods (LSC and LSA) do not perform as well as other methods regarding spatial k-anonymity. However, they stand out and achieve higher spatial k-anonymity in the scenarios of ctR than other methods when longer radii are used.

4.2. Evaluation of the average nearest neighbor (ANN) index

The differences in the results of ANN before and after applying geomasking, considered as the error introduced, were evaluated in the nine simulated scenarios with different geomasking methods with various masking radii. Figure 4 shows results with respect to the patterns of residential and sensitive locations. The y-axis of these graphs indicates the error (percent of difference) introduced by the geomasking methods. Thus, higher y values indicate larger errors (or lower analytical accuracy). As the figure shows, error increases as radius increases for most geomasking methods. However, the error introduced by point aggregation increases dramatically when longer masking radii are used.

It is worth noting that in the scenarios of rdS and rgS, point aggregation performs better than other methods when shorter radii are used, while the contrary is the case when longer radii are employed. For reasons discussed later (Section 5), the performance of geomasking methods (except point aggregation) increases with longer radii being used in the scenario of rgS. Differently, point aggregation has a larger error than the other 7 geomasking methods when applied to ctS regardless of the size of the radius. The errors introduced by point aggregation grow linearly as the masking radius increases, while errors introduced by other masking methods remain stable. Regarding the effects of the residential location patterns, however, there is no notable difference.

4.3. Evaluation of the minimum convex polygon (MCP)

The differences in the results of the MCP before and after applying geomasking to the sensitive datasets are shown in Figure 5. Not surprisingly, the error increases as the geomasking radius increases for most geomasking methods. However, the results show less consistent trends in the errors when compared to other analysis methods investigated in this study. Observing the effects of the sensitive location patterns on the error of the MCP, in the rdS scenarios, most geomasking methods have similar trends with different radii, while the point aggregation method has much larger errors when longer radii are used. In contrast, in the ctS scenarios, point aggregation has relatively smaller errors regardless of masking radii. However, the trend is less consistent as the errors of point aggregation are unstable as the radius increases. As to the effects of residential location patterns on the errors, similar trends in errors are observed in rdR and rgR. It is worth noting that random direction has relatively higher errors in rgR and

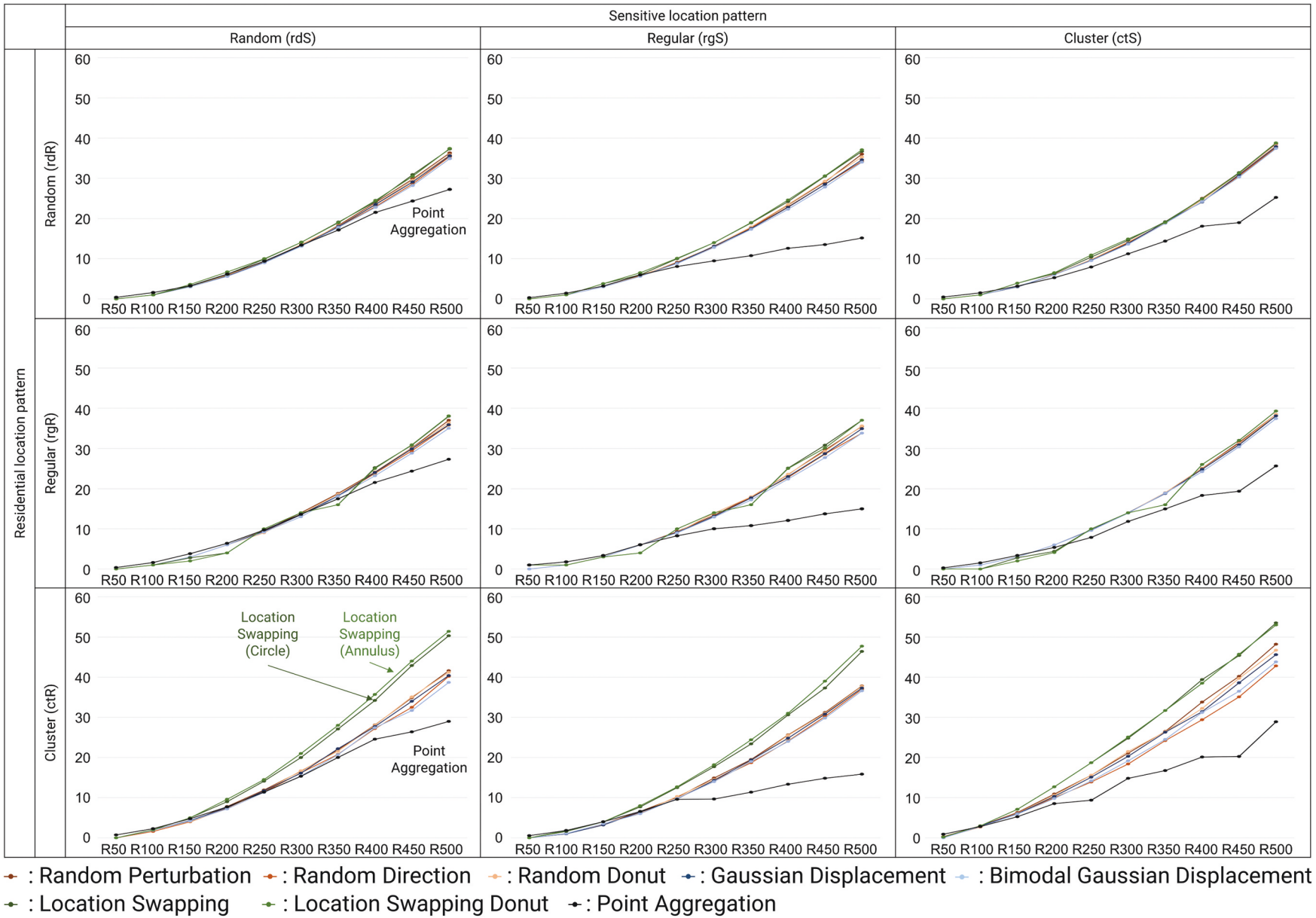

Figure 3. Results of the spatial k-anonymity regarding patterns of residential locations and sensitive locations. [X-axis of each graph indicates the radius (50–500 m); Y-axis of each graph indicates the mean spatial k-anonymity obtained from 10 independent tests. Higher values indicate higher confidentiality].

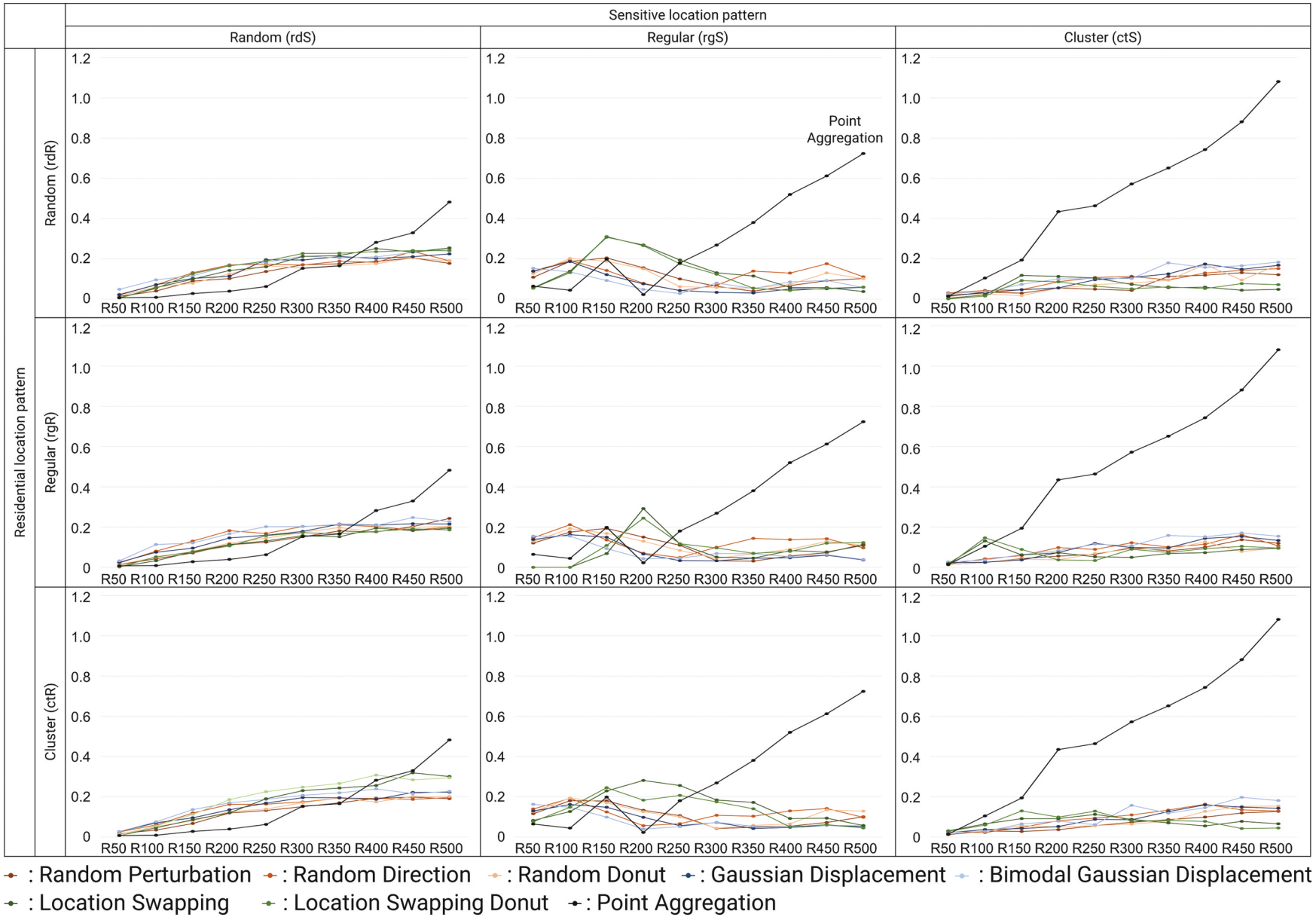

Figure 4. Results of the errors of the average nearest neighbor index regarding patterns of residential locations and sensitive locations. [X-axis of each graph indicates the size of the radius (50–500 m); Y-axis of each graph indicates the mean error of the average nearest neighbor index obtained from 10 independent tests. Higher values indicate higher errors (i.e. lower analytical accuracy)].

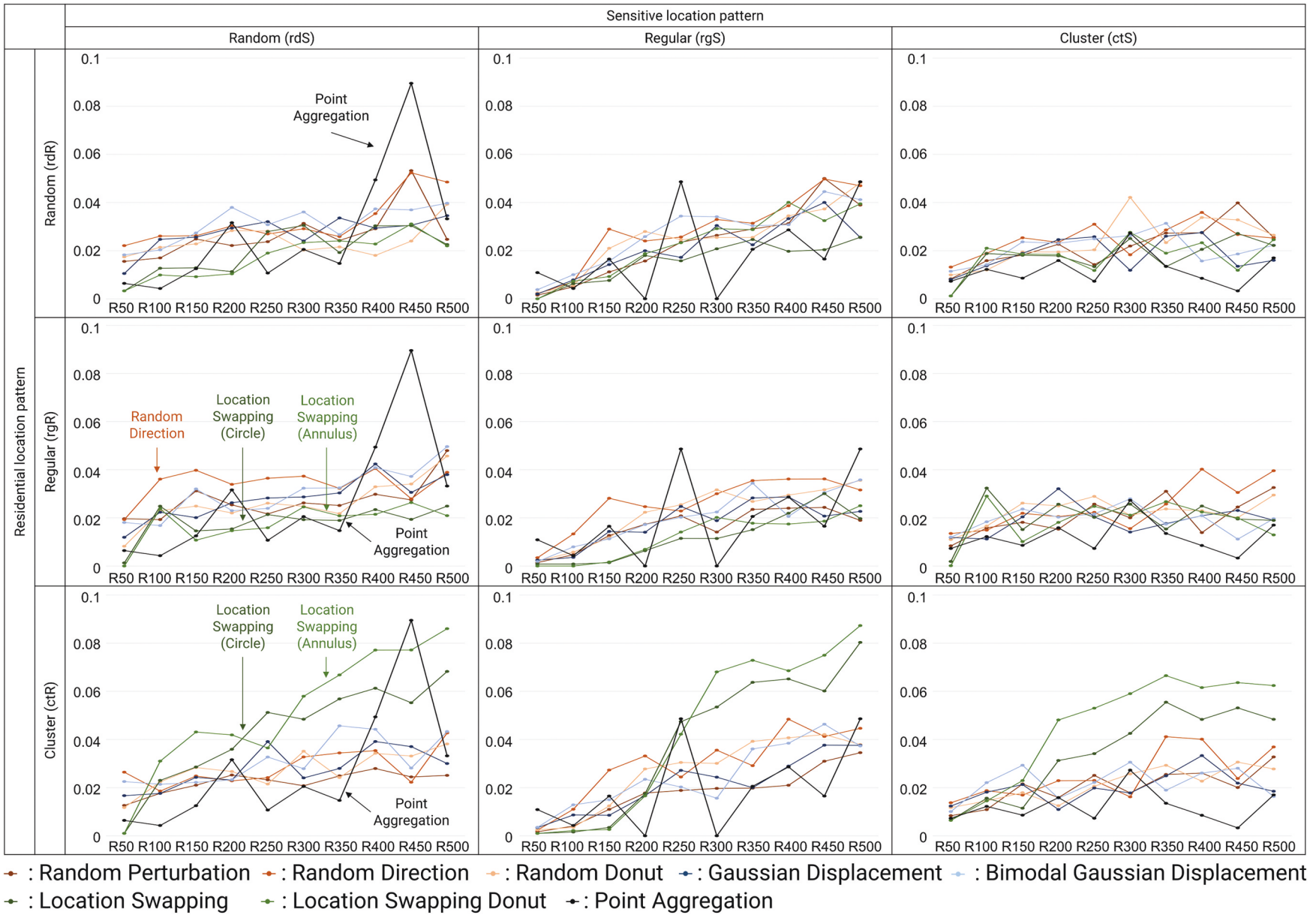

Figure 5. Results of the errors of the minimum convex polygon (MCP) regarding patterns of residential locations and sensitive locations. [X-axis of each graph indicates the size of the radius (50–500 m); Y-axis of each graph indicates the mean error of minimum convex polygon obtained from 10 independent tests. Higher values indicate higher errors (i.e. lower analytical accuracy)].

location swapping methods show relatively higher errors in the ctR scenarios compared to other geomasking methods.

4.4. Evaluation of the standard deviation ellipse (SDE)

Figure 6 illustrates the errors of the SDE for different patterns of residential and sensitive locations when various geomasking methods with different radii are applied. Observing the effects of the patterns of sensitive locations on the errors of the SDE, location swapping methods show higher errors in rgS when compared to rdS. Also, the differences in errors between the point aggregation and other geomasking methods are especially higher in ctS than in rdS and rgS regardless of the size of the radius. As to the effects of the patterns of residential locations on analytical errors, location swapping methods show higher errors in ctR when compared to rdR when longer radii are used. It is worth noting that, similar to the case of the MCP, the results show less consistent trends in the errors.

4.5. Evaluation of kernel density estimation (KDE)

Figure 7 shows the differences (evaluated by RMSE) in the result of the KDE before and after geomasking with respect to different patterns of residential and sensitive locations. Consistent with other analytical methods, error increases as the radius used in a geomasking method increases. For reasons discussed later (Section 5), the point aggregation shows lower errors than other geomasking methods regardless of the spatial patterns of the datasets.

Observing the effects of the patterns of sensitive locations, the ctS scenario gives the highest errors, followed by rgS and rdS. However, there is no notable difference in errors between geomasking methods applied to rdS, except for point aggregation, which gives the lowest errors. In rgS or ctS, the random direction geomasking method has relatively higher errors than other geomasking methods when longer radii are used. As to the effects of residential locations, there is no notable difference.

4.6. Evaluation of point density estimation (PDE)

Same as the KDE, we used the RMSE to assess the differences in the PDE results before and after applying geomasking (Figure 8). Error increases as radius increases. The point aggregation has higher errors than other geomasking methods when applied to ctS. In rdS and rgS, the errors of the point aggregations are smaller than the errors of other geomasking methods when shorter radii are used. However, errors of the point aggregations are larger than those of other geomasking methods when longer radii are applied. Regarding the effects of the patterns of residential locations on the errors of PDE, similar trends are observed for all three patterns.

As discussed in the Methods section, the entire evaluation process is independently repeated ten times to ensure the reliability of the results. The spatial k-anonymity and analytical errors were calculated as the average of results from the ten rounds of testing when different geomasking methods are applied to the patterns of residential and sensitive locations. Therefore, we investigated the distribution of spatial k-anonymity value and analytical errors since the average values could potentially hide extreme values that may be of concern. The evaluation of the boxplot results of spatial k-anonymity and analytical errors in two different radii, 500 m and 250 m, for different geomasking methods indicates a low variation, so the average values are sufficient to represent the general performance of geomasking methods.

5. Discussion and conclusion

In this study, the effectiveness of different geomasking methods was evaluated considering both data confidentiality and analytical accuracy. Spatial k-anonymity was used to assess data confidentiality, while the error introduced to the results of spatial analysis after applying geomasking was used to assess analytical accuracy. Consistent with previous studies (e.g. Armstrong et al., 1999; Kwan et al., 2004), our results suggest that there is a trade-off between geoprivacy protection and analytical accuracy when applying geomasking methods. Also, the findings provide general guidance on how to properly select geomasking methods based on the spatial patterns of sensitive datasets.

Regarding geoprivacy protection, the results suggest that point aggregation performs poorly when compared to other geomasking methods in almost all the simulated scenarios, and that location swapping methods perform better than other geomasking methods in ctR scenarios but perform worse in most rgR scenarios. This may be explained by the fact that the k value becomes higher when more residential locations are clustered near the geomasked points (Zhang et al., 2017), which leads to higher spatial k-anonymity of the location swapping methods. Based on the results of spatial k-anonymity, we suggest that it is better not to use point aggregation regardless of the spatial patterns of

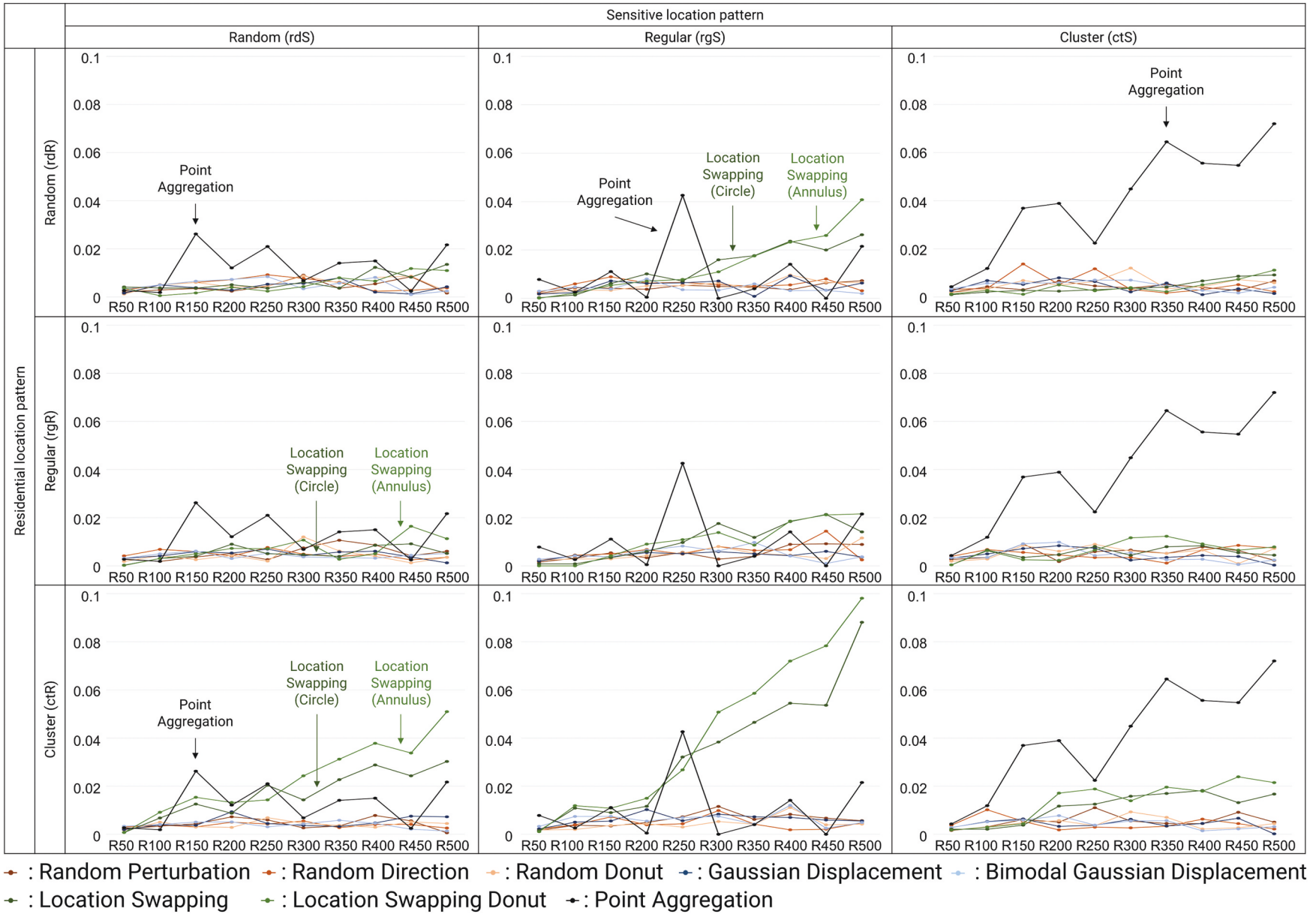

Figure 6. Results of the errors of the standard deviation ellipse (SDE) regarding patterns of residential locations and sensitive locations. [X-axis of each graph indicates the size of the radius (50–500 m); Y-axis of each graph indicates the mean error of standard deviation ellipse obtained from 10 independent tests. Higher values indicate higher errors (i.e. lower analytical accuracy)].

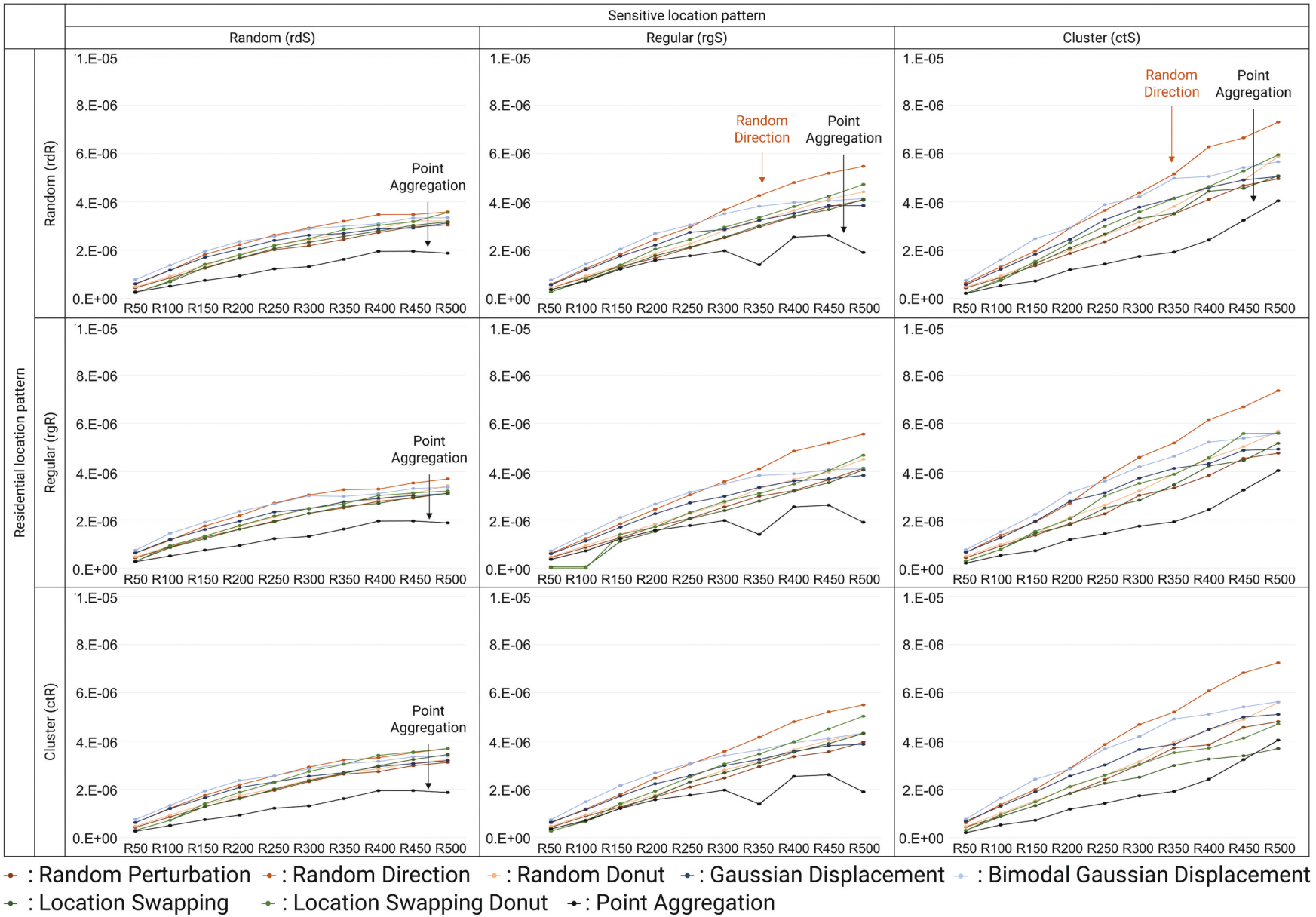

Figure 7. Results of the errors (RMSE) of kernel density estimation regarding patterns of residential locations and sensitive locations. [X-axis of each graph indicates the size of the radius (50–500 m); Y-axis of each graph indicates the average RMSE of kernel density estimation obtained from 10 independent tests. Higher values indicate higher errors (i.e. lower analytical accuracy)].

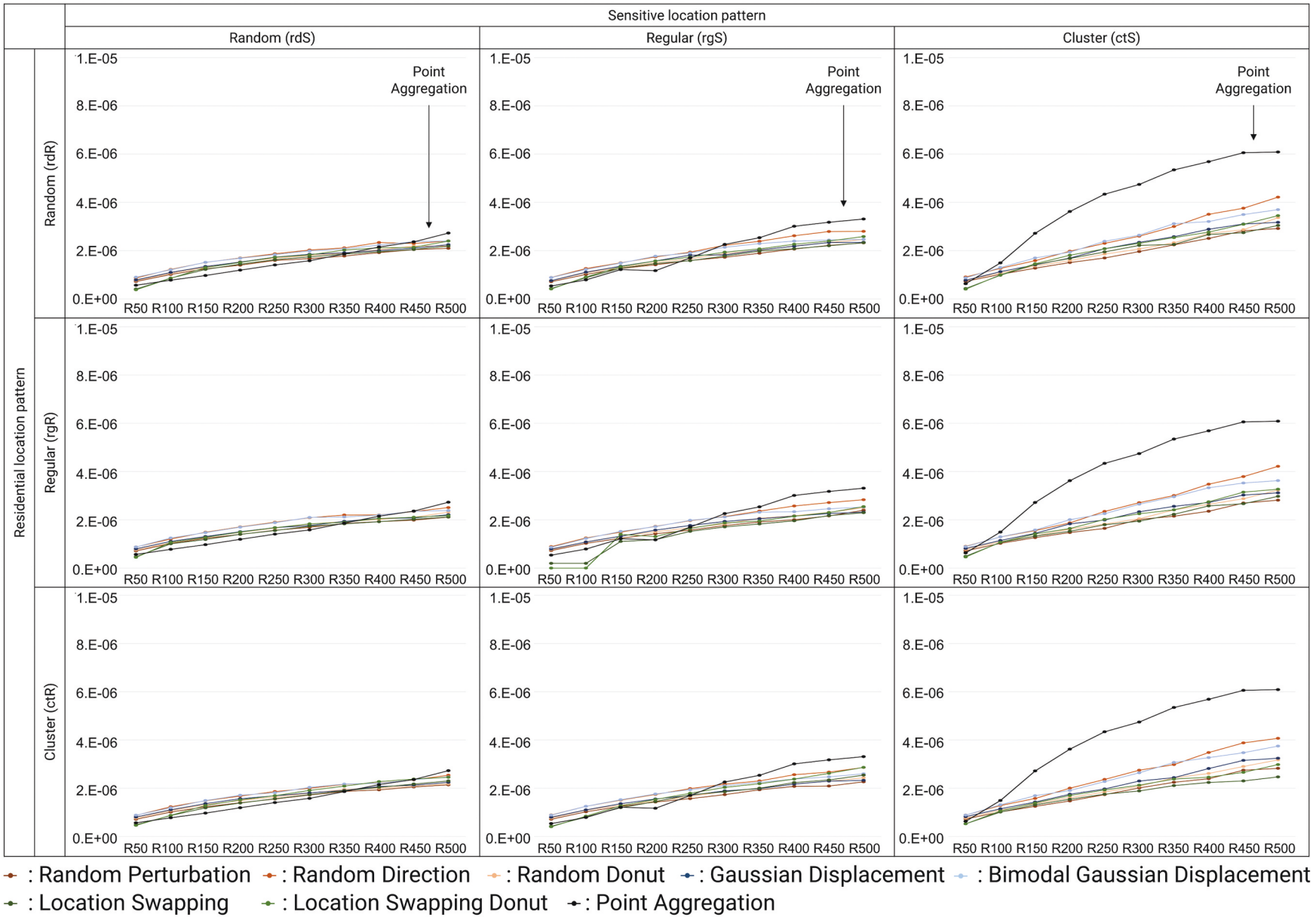

Figure 8. Results of the errors (RMSE) of point density estimation regarding patterns of residential locations and sensitive locations. [X-axis of each graph indicates the size of the radius (50–500 m); Y-axis of each graph indicates the average RMSE of point density estimation obtained from 10 independent tests. Higher values indicate higher errors (i.e. lower analytical accuracy)].

the residential and sensitive locations, and that location swapping geomasking methods seem preferable for ctR scenarios.

The results of the ANN suggest that point aggregation has higher errors than other geomasking methods when longer radii are used and has lower errors when shorter radii are used, especially for rdS and rgS scenarios. These findings can be explained by the fact that the ANN is evaluated largely based on the distances between two nearest points. Thus, if the overall distribution of the distances between two nearest points is preserved after geomasking, the error introduced by geomasking would be lower. Specifically, when longer radii are used, point aggregation leads to higher errors (than other geomasking methods) because it distorts the overall distribution of the distances between two nearest points to a larger extent. This is especially true for the case of ctS because the clustered patterns may be highly distorted by geomasking, which leads to higher errors. For example, the histograms in Figure 9 show the frequencies of different distances to the nearest points. As seen in this figure, the histogram of geomasked data by point aggregation is notably different from that of the original points, while the one by location swapping is similar to that of the original points. Generally, when data analysis is conducted with the ANN, we suggest not using point aggregation with long radii in all scenarios, but point aggregation with short radii may be used in rdS.

For MCP analysis, the results suggest that the errors introduced by geomasking vary considerably. This can be explained by the fact that the MCP is highly sensitive to specific locations of points (e.g. outliers). Recall that the MCP is defined by the smallest polygon that contains all points. If the few original points comprising the edges of the MCP are relocated far from the original points (due to geomasking), the overall shape of the MCP can change significantly. In this light, the MCP is particularly sensitive to geomasking, which leads to higher variations in errors. In addition, it was observed that higher errors were introduced by the location swapping methods with long radii for ctR. It may be because the location swapping methods consider surrounding residential locations to relocate the geomasked points: if a sensitive location is far from the clusters of the residential locations, these methods may lead to higher errors given that the MCP is sensitive to the specific locations of points. Thus, based on the findings

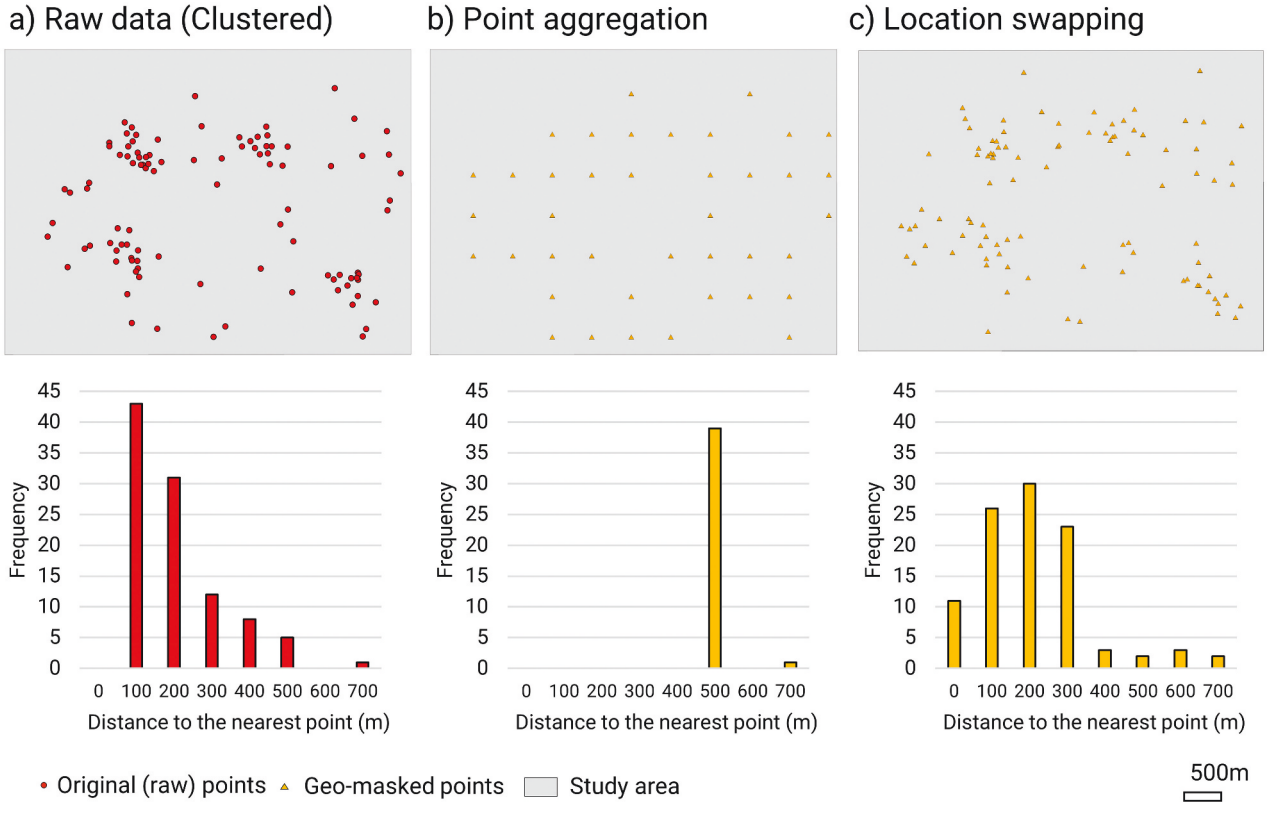

Figure 9. The histogram of the distance to the nearest points in the original data and geomasked data: a) original points (clustered pattern) and the histogram of frequency by distance to the nearest point, b) geomasked points by point aggregation method with 500 m radius and the histogram of frequency by distance to the nearest point, c) geomasked points by location swapping method with 500 m radius.

and for data analysis using the MCP, we recommend avoiding using location swapping methods regardless of the length of the radius, especially for ctR scenarios.

Notably, the variation in errors introduced by geomasking methods in SDE analysis is smaller than that in the MCP. This is because, unlike the MCP where the exact locations of points are directly used for defining the polygon, the SDE captures the general dispersion and orientation of points. Therefore, errors of the SDE tend to be less sensitive than those of the MCP. In general, when data analysis is conducted with the SDE, we suggest not using the location swapping methods for rgS when long radii are used, not using the point aggregation for ctS regardless of the length of the radius, and not using the location swapping methods for ctR when long radii are used.

Interestingly, the results of the kernel density analysis show that point aggregation has lower errors than other geomasking methods. This may be explained by the fact that KDE calculates the overall density of observations at each raster cell by using a smoothly curved surface (i.e. the kernel) that fits points within the bandwidth (Environmental Systems Research Institute, 2022; Yin, 2020). The point aggregation geomasking method aggregates points to the nearest centroids of grid cells, and those aggregated centroids may somewhat capture the local density of points. Actually, the aggregating points can be considered as an approximation of KDE. For instance, Figure 10 shows that the kernel density surface generated from geomasked data by point aggregation has high-density regions (i.e. reddish color) that are similar to those generated from the original points, while the random direction method significantly changed the original high-density regions. Thus, the errors of KDE when using point aggregation can be smaller than those of other geomasking methods. In summary, when data analysis is conducted using KDE, we recommend that point aggregation may be used regardless of the patterns of the residential and sensitive locations, and not using random direction geomasking method for rgS or ctS when long radii are used.

Different from kernel density analysis, the results of PDE suggest that the point aggregation method has higher errors than other geomasking methods, especially for ctS. This may be explained by the fact that point density is largely influenced by the number of points within a certain bandwidth (illustrated in Figure 11). Unlike KDE, which estimates a smoothly curved surface within the bandwidth, point density is estimated by counting points located within the bandwidth. If the number of points within the bandwidth changes considerably (which largely occurs in the point aggregation method), point density would also change considerably. It implies that the locations of points play an important role in PDE. Therefore, when data analysis is conducted with PDE, we recommend not using point aggregation for ctS regardless of the radius or for rgS when long radii are used.

Summarizing the results of all the tests in this study, we propose Tables 2–4 as the guidelines of the geomasking method with respect to the spatial patterns of sensitive and residential locations. Researchers can refer to this table to select the suitable geomasking methods for their analysis: first, identify the spatial patterns of the sensitive locations (e.g. home locations of AIDS patients) and residential locations (e.g. existing residential locations in the study area) being studied; second, refer to the corresponding cell of the table for both suggested (denoted by "+") and not suggested (denoted by "-") geomasking methods based on the spatial patterns of the point datasets and spatial analysis method. Additionally, "O" in the table indicates there is no notable difference observed or no specific suggestion concluded from the test and thus researchers should consider testing different geomasking methods based on their data because geomasking methods may be sensitive to the unique spatial pattern.

Taking MCP analysis as an example, according to the guidelines in Table 4, if the sensitive locations have a regular pattern and the residential locations have a clustered pattern, the location swapping methods (LSC and LSA) are not suggested to be used with longer geomasking radii. Instead, point aggregation methods are recommended in this scenario.

Further, when using the proposed guidelines to select the proper geomasking methods for their data, we strongly recommend that researchers apply the guidelines in a flexible manner. For example, when a guideline suggests PA, it does not imply that researchers must select the point aggregation method. Instead, it suggests that point aggregation may be one of the suitable choices for the case. Also, when a guideline does not suggest LSC, for instance, it indicates that researchers should consider using other geomasking methods than the location swapping within a circle method, by testing the performance of other geomasking methods according to the framework, and then choosing the ones with the best performance for their data. Since the guidelines are not exhaustive, we recommend researchers interpret the guidelines in a flexible manner and pay more attention to the unique spatial patterns of their data when selecting geomasking methods. The best practice may be testing the performance of several

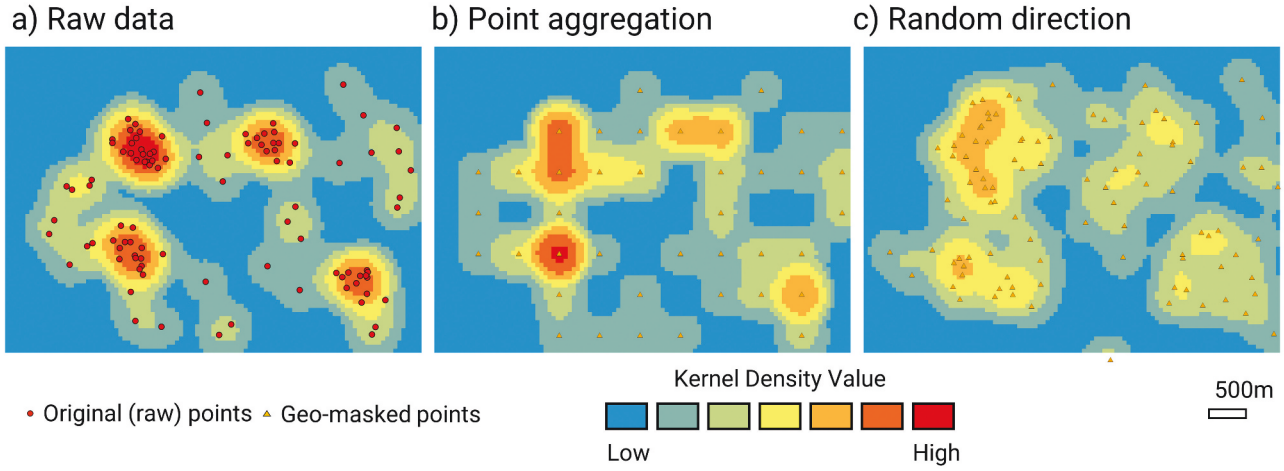

Figure 10. The kernel density estimation (KDE) generated from original data and geomasked data: a) KDE generated from original points; b) KDE generated from geomasked points by point aggregation with 500 m radius; c) KDE generated from geomasked points by random direction with 500 m radius.

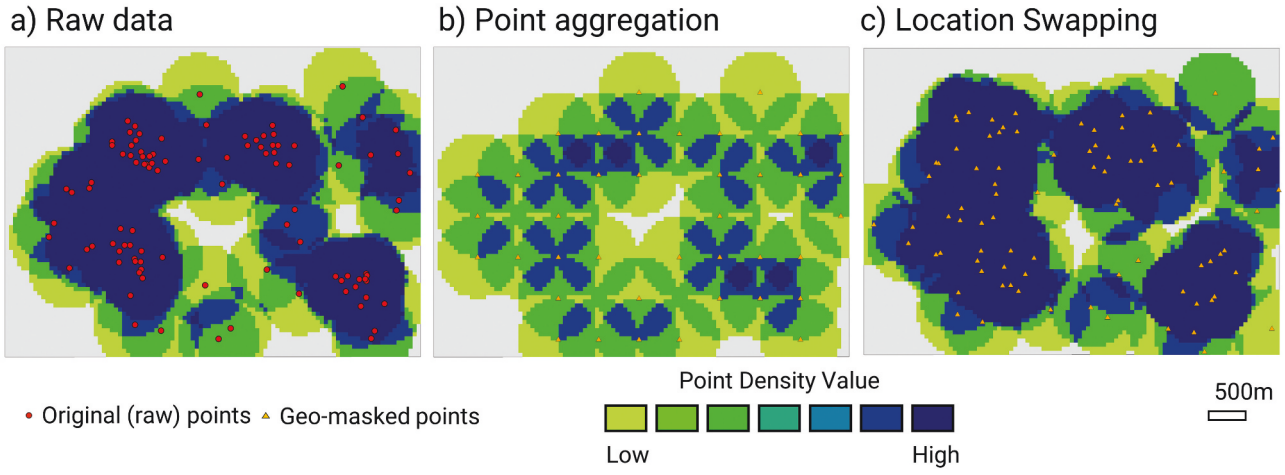

Figure 11. The point density estimation (PDE) generated from original data and geomasked data: a) PDE generated from original points; b) PDE generated from geomasked points by point aggregation with 500 m radius; c) PDE generated from geomasked points by location swapping with 500 m radius.

geomasking methods for a given dataset by following the evaluation framework proposed in this study and choosing the geomasking method that performed best.

Although the tests we conducted in this study cover many spatial analysis methods that are widely used in the field of geography and other relevant fields, many issues still need further exploration and to be addressed in future studies. First of all, the assessment in this study is not exhaustive, so there are many other spatial analysis methods (e.g. Moran's *I*, spatial regression models) that were not evaluated. In addition, these guidelines focus only on individual-level geospatial data with confidential locations. However, researchers may want to explore proper geomasking methods for more complex data, such as GPS trajectories. We did not address GPS trajectories data because geomasking methods for GPS trajectories data are limited (Seidl et al., 2016; Wang & Kwan, 2020). Furthermore, this study focuses on geoprivacy concerns raised by the spatial component of sensitive datasets. However, geospatial data often contains personal socio demographic information, such as age, gender, income level, and health status, which may also cause serious privacy concerns if not handled properly. Also, the masking performance may be affected when population density is considered. For example, random perturbation and donut masking are easily performed while considering population density and doing so would significantly affect the trade-off between k-anonymity and analytical accuracy. It calls for further investigation in future studies. Another interesting

Table 2. Geomasking method guidelines for residential locations in random pattern with various spatial patterns of sensitive locations.

		Sensitive location pattern					
		Random (rdS)		Regular (rgS)		Cluster (ctS)	
		Shorter Masking Radii	*Longer Masking Radii*	*Shorter Masking Radii*	*Longer Masking Radii*	*Shorter Masking Radii*	*Longer Masking Radii*
Residential location in random pattern (rdR)	SKA	O	-PA	O	-PA	O	-PA
	ANN	+PA	-PA	O	-PA	-PA	-PA
	MCP	O	-PA	O	O	+PA	+PA
	SDE	-PA	O	-PA	-LSC -LSA	-PA	-PA
	KDE	+PA	+PA	O	+PA -RD	+PA	+PA -RD
	PDE	O	O	O	-PA	-PA	-PA

Notes: SKA: Spatial k-anonymity; ANN: Average Nearest Neighbor Index; MCP: Minimum Convex Polygon; SDE: Standard Deviation Ellipse; KDE: Kernel Density Estimation; PDE: Point Density Estimation. PA: Point Aggregation; LSC: Location Swapping within a circle; LSA: location swapping within an annulus; RD: Random Direction; +: the specific method is suggested; -: the specific method is not suggested; O: no notable difference observed among the tested masking methods.

Table 3. Geomasking method guidelines for residential locations in regular pattern with various spatial patterns of sensitive locations.

		Sensitive location pattern					
		Random (rdS)		Regular (rgS)		Cluster (ctS)	
		Shorter Masking Radii	*Longer Masking Radii*	*Shorter Masking Radii*	*Longer Masking Radii*	*Shorter Masking Radii*	*Longer Masking Radii*
Residential location in regular pattern (rgR)	SKA	-LSC -LSA	-PA	-LSC -LSA	-PA	-LSC -LSA	-PA
	ANN	+PA	-PA	O	-PA	-PA	-PA
	MCP	-RD	-PA	-RD	-RD	O	-RD
	SDE	-PA	O	-PA	-LSC -LSA	-PA	-PA
	KDE	+PA	+PA	O	+PA -RD	+PA	+PA -RD
	PDE	O	O	+LSC +LSA	-PA	-PA	-PA

Notes: SKA: Spatial k-anonymity; ANN: Average Nearest Neighbor Index; MCP: Minimum Convex Polygon; SDE: Standard Deviation Ellipse; KDE: Kernel Density Estimation; PDE: Point Density Estimation. PA: Point Aggregation; LSC: Location Swapping within a circle; LSA: location swapping within an annulus; RD: Random Direction; +: the specific method is suggested; -: the specific method is not suggested; O: no notable difference observed among the tested masking methods.

Table 4. Geomasking method guidelines for residential locations in clustered pattern with various spatial patterns of sensitive locations.

		Sensitive location pattern					
		Random (rdS)		Regular (rgS)		Cluster (ctS)	
		Shorter Masking Radii	*Longer Masking Radii*	*Shorter Masking Radii*	*Longer Masking Radii*	*Shorter Masking Radii*	*Longer Masking Radii*
Residential location in clustered pattern (ctR)	SKA	+LSC +LSA	+LSC +LSA -PA	+LSC +LSA	+LSC +LSA -PA	+LSC +LSA	+LSC +LSA -PA
	ANN	+PA	-PA	O	-PA	-PA	-PA
	MCP	+PA	-LSC -LSA	O	+PA -LSC -LSA	O	+PA -LSC -LSA
	SDE	-PA -LSC -LSA	-LSC -LSA	O	-LSC -LSA	-PA	-PA -LSC -LSA
	KDE	+PA	+PA	O	+PA -RD	+PA	+PA -RD
	PDE	O	O	O	-PA	-PA	-PA

Notes: SKA: Spatial k-anonymity; ANN: Average Nearest Neighbor Index; MCP: Minimum Convex Polygon; SDE: Standard Deviation Ellipse; KDE: Kernel Density Estimation; PDE: Point Density Estimation. PA: Point Aggregation; LSC: Location Swapping within a circle; LSA: location swapping within an annulus; RD: Random Direction; +: the specific method is suggested; -: the specific method is not suggested; O: no notable difference observed among the tested masking methods.

future research direction is to examine the feasibility of integrating geomasking into online geocoders or data analysis platforms, which return masked geocoded locations to protect sensitive geospatial data.

In this exploratory assessment, a limited number of the widely used basic geomasking methods and geoprivacy measurement in the field of spatial data privacy research were studied. We acknowledge there are other geomaksing methods (e.g. the Voronoi masking, adaptive areal elimination or masking, street masking, and MGRS masking) and measures of disclosure risk (e.g. l-diversity and t-closeness), and further study based on other methods or measures are needed in the future. Readers can refer to existing open-source codes and tools of different geomaksing and geoprivacy measurement algorithms, such as adaptive geographical masking (Kounadi, 2020), adaptive areal anonymization ArcGIS toolbox (Charleux & Schofield, 2020), MaskMy.XYZ (Swanlund, Schuurman, et al., 2020a), Privy (Ajayakumar et al., 2019).

Additionally, false identification (Kim et al., 2021; Seidl et al., 2018) is an emerging concern of applying geomasking methods, which indicate the linking of the masked data points to incorrect persons or households (Polzin & Kounadi, 2021). The false identification transferred the potential negative effects of being identified from the true persons or households to individuals who were not part of the research (National Research Council, 2007). There are newly developed geomasking methods that target addressing this issue, such as the adaptive Voronoi masking (Polzin & Kounadi, 2021). In future studies, false identification should be involved in the assessment when comparing the newer geomasking methods. Further, the guideline obtained from the test results can be scale-dependent, so researchers should carefully consider the study area context (especially spatial scale) when selecting geomasking methods based on our guideline. Lastly, the evaluation was not implemented with real-world datasets due to privacy concerns. That is the major reason for using simulated data. However, the performance of geomasking may be affected by the specific characteristic (e.g. the shape of the study area is irregular) of real-world applications. Real-world applications in various locations, on the premise of ensuring geoprivacy, need to be tested in future studies. Also, applying geomasking techniques on large real-world datasets may be prohibitively computationally intensive, and CyberGIS can be a promising direction to explore for addressing this issue in future studies (Delmelle et al., 2022).

Rather than the performance of geomasking methods, applying geomasking needs to consider the confidential degree of different datasets and the legal framework of various countries. Geospatial data comes with different levels of confidentiality, thus requiring various levels of geomasking to be protected. For instance, the dataset containing residential locations of sexual assault victims may require a higher level of geomasking when compared to the one with locations of street vandalism. Further, the degree of geomasking may also depend on the legal framework of the country within which the study area is located. In the United States, there are no other formal laws about the protection of personal location privacy except the Privacy Act of 1974. Other countries (e.g. the European Union) may have stricter personal privacy laws compared to the US, such as the European Union's Data Protection Directive and General Data Protection Regulation (GDPR), Australian Information Privacy Principles under the Privacy Act of 1988, Japan's Personal Information Protection Law, and Singapore's E-commerce Code for the Protection of Personal Information and Communications of Consumers of Internet Commerce.

Privacy protection of personal geospatial data is a systematic project, and more studies are needed in this field to explore and discuss the issue. This research is an exploratory study to investigate the performance of some selected geomasking methods under different urban pattern scenarios, explore practical ways to evaluate them and provide preliminary guidelines for potential users. This research may shed new light on the geoprivacy protection research and help the construction of guidelines for preserving personal location privacy. The findings of this research facilitate researchers to understand the effectiveness of geomasking methods and provide practical guidelines on how to properly apply geomasking methods related to the spatial structure of their data. Ultimately, this study may promote the sharing of geospatial data, which will encourage collaboration among disciplines and promote research reproducibility while protecting personal geoprivacy, thereby benefiting not only the academic community but also humans individually.

Acknowledgments

The authors would like to thank the editors and anonymous reviewers for their helpful comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the U.S. National Science Foundation (Grant No. BCS-2025783) and the University of Toronto (UTM Start-up Funding).

ORCID

Jue Wang http://orcid.org/0000-0002-6305-4298
Junghwan Kim http://orcid.org/0000-0002-7275-769X
Mei-Po Kwan http://orcid.org/0000-0001-8602-9258

Data availability statement

The data that support the findings of this study are available with the identifier at https://doi.org/10.17605/OSF.IO/HXD5W and subject to Creative Commons license (CC-By Attribution 4.0 International).

References

Ajayakumar, J., Curtis, A. J., & Curtis, J. (2019). 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. https://doi.org/10.1186/s12942-019-0194-8

Allshouse, W. B., Fitch, M. K., Hampton, K. H., Gesink, D. C., Doherty, I. A., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. *Geocarto International*, *25*(6), 443–452. https://doi.org/10.1080/10106049.2010.496496

Armstrong, M. P., & Ruggles, A. J. (2005). Geographic information technologies and personal privacy. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *40*(4), 63–73. https://doi.org/10.3138/ru65-81r3-0w75-8v21

Armstrong, M. P., Rushton, G., & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, *18*(5), 497–525. https://doi.org/10.1002/(sici)1097-0258(19990315)18:5<497::aid-sim45>3.0.co;2-%23

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R*. Springer. https://doi.org/10.1007/978-1-4614-7618-4

Boulos, M. N. K., Peng, G., & Vopham, T. (2019). An overview of GeoAI applications in health and healthcare. *International Journal of Health Geographics*, *18*(1), 1–9. https://doi.org/10.1186/s12942-019-0171-2

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*(1), 56. https://doi.org/10.1186/1476-072X-5-56

Carr, J., Vallor, S., Freundschuh, S., Gannon, W. L., & Zandbergen, P. (2014). Hitting the moving target: Challenges of creating a dynamic curriculum addressing the ethical dimensions of geospatial data. *Journal of Geography in Higher Education*, *38*(4), 444–454. https://doi.org/10.1080/03098265.2014.936313

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*(2), 160–165. https://doi.org/10.1197/jamia.M1920

Chaix, B., Kestens, Y., Duncan, D. T., Brondeel, R., Méline, J., El Aarbaoui, T., Pannier, B., & Merlo, J. (2016). A GPS-based methodology to analyze environment-health associations at the trip level: Case-crossover analyses of built environments and walking. *American Journal of Epidemiology*, *184*(8), 579–589. https://doi.org/10.1093/aje/kww071

Charleux, L., & Schofield, K. (2020). True spatial k-anonymity: Adaptive areal elimination vs. adaptive areal masking. *Cartography and Geographic Information Science*, *47*(6), 537–549. https://doi.org/10.1080/15230406.2020.1794975

Clarke, K. C. (2016). A multiscale masking method for point geographic data. *International Journal of Geographical Information Science*, *30*(2), 300–315. https://doi.org/10.1080/13658816.2015.1085540

Clifton, K. J., & Gehrke, S. R. (2013). Application of geographic perturbation methods to residential locations in the Oregon household activity survey. *Transportation Research Record: Journal of the Transportation Research Board*, *2354*(1), 40–50. https://doi.org/10.3141/2354-05

Curtis, A., Mills, J. W., Agustin, L., & Cockburn, M. (2011). Confidentiality risks in fine scale aggregations of health data. *Computers, Environment and Urban Systems*, *35*(1), 57–64. https://doi.org/10.1016/j.compenvurbsys.2010.08.002

Curtis, A., Mills, J., & Leitner, M. (2006). Spatial confidentiality and GIS: Re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics*, *5*(1), 44. https://doi.org/10.1186/1476-072X-5-44

Delmelle, E. M., Desjardins, M. R., Jung, P., Owusu, C., Lan, Y., Hohl, A., & Dony, C. (2022). Uncertainty in geospatial health: Challenges and opportunities ahead. *Annals of Epidemiology*, *65*, 15–30. https://doi.org/10.1016/j.annepidem.2021.10.002

Duncan, D. T., Castro, M. C., & Blossom, J. C. (2012). Response to Geocoding-protected health information using online services may compromise patient privacy-Comments on "Evaluation of the positional difference between two common geocoding methods" by Duncan et al. *Geospatial Health*, *6*(2), 158–159. https://doi.org/10.4081/gh.2012.133

Duncan, G. T., & Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, *6*(3), 219–232. https://doi.org/10.1214/ss/1177011681

Environmental Systems Research Institute (ESRI). (2022). *How Kernel Density works*. https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-analyst/how-kernel-density-works.htm

Fuller, D., Shareck, M., & Stanley, K. (2017). Ethical implications of location and accelerometer measurement in health research studies with mobile sensing devices. *Social Science & Medicine*, *191*, 84–88. https://doi.org/10.1016/j.socscimed.2017.08.043

Ghinita, G., Zhao, K., Papadias, D., & Kalnis, P. (2010). A reciprocal framework for spatial K-anonymity. *Information Systems*, *35*(3), 299–314. https://doi.org/10.1016/j.is.2009.10.001

Gutmann, M. P., Witkowski, K., Colyer, C., O'Rourke, J. M., & McNally, J. (2008). Providing spatial data for secondary analysis: Issues and current practices relating to confidentiality. *Population Research and Policy Review*, *27*(6), 639–665. https://doi.org/10.1007/s11113-008-9095-4

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L., & Miller, W. C. (2010). Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology*, *172*(9), 1062–1069. https://doi.org/10.1093/aje/kwq248

Kim, J., & Kwan, M. P. (2021). Travel time errors caused by geomasking might be different between transportation modes and types of urban area. *Transactions in GIS*, *25*(4), 1910–1926. https://doi.org/10.1111/tgis.12751

Kim, J., Kwan, M.-P., Levenstein, M. C., & Richardson, D. B. (2021). How do people perceive the disclosure risk of maps? Examining the perceived disclosure risk of maps and its implications for geoprivacy protection. *Cartography and Geographic Information Science*, *48*(1), 2–20. https://doi.org/10.1080/15230406.2020.1794976

Kounadi, O. (2020). *Adaptive Geographical Masking*. https://github.com/okounadi/Geoprivacy

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*, *9*(4), 34–45. https://doi.org/10.1177/1556264614544103

Kounadi, O., & Leitner, M. (2016). Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems*, *57*, 1:1–1:17. Dagstuhl Publishing, Germany. https://doi.org/10.1016/j.compenvurbsys.2016.01.004

Kounadi, O., & Resch, B. (2018). 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. https://doi.org/10.1177/1556264618759877

Kounadi, O., Resch, B., & Petutschnig, A. (2018). Privacy threats and protection recommendations for the use of geosocial network data in research. *Social Sciences*, *7*(10), 191. https://doi.org/10.3390/socsci7100191

Kwan, M.-P. (2012). How GIS can help address the uncertain geographic context problem in social science research. *Annals of GIS*, *18*(4), 245–255. https://doi.org/10.1080/19475683.2012.727867

Kwan, M.-P., Casas, I., & Schmitz, B. C. (2004). Protection of geoprivacy and accuracy of spatial information. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *39*(2), 15–28. https://doi.org/10.3138/X204-4223-57MK-8273

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, Istanbul, Turkey. IEEE. https://doi.org/10.1109/ICDE.2007.367856

Lu, Y., Kawamura, K., & Zellner, M. L. (2008). Exploring the influence of urban form on work travel behavior with agent-based modeling. *Transportation Research Record*, *2082*(1), 132–140. https://doi.org/10.3141/2082-16

Lu, Y., Yorke, C., & Zhan, F. B. (2012). Considering Risk Locations When Defining Perturbation Zones for Geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *47*(3), 168–178. https://doi.org/10.3138/carto.47.3.1112

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3. https://doi.org/10.1145/1217299.1217302

Marshall, W. E., & Garrick, N. W. (2010). Street network types and road safety: A study of 24 California cities. *Urban Design International*, *15*(3), 133–147. https://doi.org/10.1057/udi.2009.31

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229. https://doi.org/10.1126/science.1250475

National Academies of Sciences. (2019). *Reproducibility and replicability in science*. National Academies Press. https://doi.org/10.17226/25303

National Research Council. (2007). *Putting people on the map: Protecting confidentiality with linked social-spatial data*. National Academies Press. https://doi.org/10.17226/11865

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press. http://www.sup.org/books/title/?id=8862

Owusu, C., Delmelle, E., Tang, W., Silverman, G., & Dye, S. (2020). A multistage, geocoding approach for the development of a database of private wells in Gaston County, North Carolina. *Journal of Environmental Health*, *83*(4), 8–16. https://link.gale.com/apps/doc/A639890265/AONE?u=nhais_hsck&sid=googleScholar&xid=9613e13d

Owusu, C., Lan, Y., Zheng, M., Tang, W., & Delmelle, E. (2017). Geocoding fundamentals and associated challenges. In H. Karimi & B. Karimi (Eds.), *Geospatial data science techniques and applications* (pp. 41–62). CRC Press. https://doi.org/10.1201/b22052

Polzin, F. S., & Kounadi, O. (2021). Adaptive Voronoi Masking. A method to protect confidential discrete spatial data. In K. Janowicz & J. A. Verstegen (Eds.), *11th International Conference on Geographic Information Science*. https://doi.org/10.4230/LIPIcs.GIScience.2021.II.1

Richardson, D. B., Kwan, M.-P., Alter, G., & McKendry, J. E. (2015). Replication of scientific research: Addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research. *Annals of GIS*, *21*(2), 101–110. https://doi.org/10.1080/19475683.2015.1027792

Seidl, D. E., Jankowski, P., & Clarke, K. C. (2018). Privacy and false identification risk in geomasking techniques. *Geographical Analysis*, *50*(3), 280–297. https://doi.org/10.1111/gean.12144

Seidl, D. E., Jankowski, P., & Tsou, M.-H. (2016). Privacy and spatial pattern preservation in masked GPS trajectory data. *International Journal of Geographical Information Science*, *30*(4), 785–800. https://doi.org/10.1080/13658816.2015.1101767

Seidl, D. E., Paulus, G., Jankowski, P., & Regenfelder, M. (2015). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography*, *63*, 253–263. https://doi.org/10.1016/j.apgeog.2015.07.001

Sherman, J. E., & Fetters, T. L. (2007). Confidentiality concerns with mapping survey data in reproductive health research. *Studies in Family Planning*, *38*(4), 309–321. https://doi.org/10.1111/j.1728-4465.2007.00143.x

Stinchcomb, D. (2004). *Procedures for geomasking to protect patient confidentiality*. October 17-24, 2004, ESRI International Health GIS Conference held in Washington, DC. https://proceedings.esri.com/library/userconf/health04/papers/pap3012.pdf

Swanlund, D., Schuurman, N., & Brussoni, M. (2020a). MaskMy.XYZ: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS*, *24*(2), 390–401. https://doi.org/10.1111/tgis.12606

Swanlund, D., Schuurman, N., Zandbergen, P., & Brussoni, M. (2020b). Street masking: A network-based geographic mask for easily protecting geoprivacy. *International Journal of Health Geographics*, *19*(1), 1–11. https://doi.org/10.1186/s12942-020-00219-z

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Tellman, N., Litt, E. R., Knapp, C., Eagan, A., Cheng, J., & Lewis, J., Jr. (2010). The effects of the Health Insurance Portability and Accountability Act privacy rule on influenza research using geographical information systems. *Geospatial Health*, *5*(1), 3–9. https://doi.org/10.4081/gh.2010.182

United States Census Bureau. (2021). *Urban Areas Facts*. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/ua-facts.html

VanWey, L. K., Rindfuss, R. R., Gutmann, M. P., Entwisle, B., & Balk, D. L. (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences*, *102*(43), 15337–15342. https://doi.org/10.1073/pnas.0507804102

Wang, J., & Kwan, M.-P. (2018). An analytical framework for integrating the spatiotemporal dynamics of environmental context and individual mobility in exposure assessment: A study on the relationship between food environment exposures and body weight. *International Journal of Environmental Research and Public Health*, *15*(9), 2022. https://doi.org/10.3390/ijerph15092022

Wang, J., & Kwan, M.-P. (2020). Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets. *International Journal of Health Geographics*, *19*(1), 7. https://doi.org/10.1186/s12942-020-00201-9

Yin, P. (2020). Kernels and density estimation. In J. P. Wilson (Ed.), *The geographic information science & technology body of knowledge*. Association of American Geographers. http://gistbok.ucgis.org/bok-topics/kernels-and-density-estimation

Yoo, E., Rudra, C., Glasgow, M., & Mu, L. (2015). Geospatial estimation of individual exposure to air pollutants: Moving from static monitoring to activity-based dynamic exposure assessment. *Annals of the Association of American Geographers*, *105*(5), 915–926. https://doi.org/10.1080/00045608.2015.1054253

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine*, *2014*, 1–14. https://doi.org/10.1155/2014/567049

Zhang, S., Freundschuh, S. M., Lenzer, K., & Zandbergen, P. A. (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science*, *44*(1), 22–34. https://doi.org/10.1080/15230406.2015.1095655