JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 31 (2025), pp. 3–25
doi:10.5311/JOSIS.2025.31.375

RESEARCH ARTICLE

Preserving privacy of spatial distances using randomized geometric surface calculations

Jonas Klingwort

Department of Research & Development, Statistics Netherlands (CBS), the Netherlands

Sarah Redlich

Statistical Office of North Rhine-Westphalia (IT.NRW), Germany

Received: May 14, 2024; returned: July 11, 2024; revised: December 4, 2024; accepted: March 29, 2025.

Abstract: This paper introduces and evaluates a novel method for privacy-preserving distance computations. The method is based on randomized geometric surface calculations and replaces coordinates with contextual variables representing information about the coordinates or the distances between coordinates. The method is presented with an accompanying step-by-step workflow. Its applicability is demonstrated with real-world spatial data sets from Germany and the Netherlands that contain information about hospital and school locations. Open data was used to enable reproducibility. The method's utility is evaluated in detail using correlations, the relative root mean squared error (RRMSE), a Monte Carlo simulation, and the Wasserstein distance. The results show that the method yields high correlations, provides reasonably accurate results as an RRMSE of about 20 % is achieved, converges fast, and preserves the spatial distribution of the true coordinates.

Keywords: geographical data, geo-referenced data, geomasking, distance matrix, privacy, disclosure risk, confidentiality

1 Introduction

Spatial and geo-referenced data containing information about the distance between the place of residence and points of interest, such as schools or hospitals, is relevant for several research questions in medical or social science research [3, 26]. For example, the distance to schools is relevant for questions concerning education access and equity, the physical activity and health of students [7, 24] as well as real estate values [8, 13]. The distance to hospitals is relevant regarding healthcare access and quality, health outcomes, and emergency response times [1, 5, 27, 29, 32, 43].

These examples of different research fields have in common that detailed spatial data about the distance between a place of residence and a point of interest must be known and available to the researcher. However, access to datasets containing geo-referenced information is restricted to ensure privacy. Furthermore, even when available, geo-referenced information is often separated from information potentially identifying an individual [4, 36].

To overcome privacy concerns, geomasking methods are used to preserve spatial privacy while maintaining spatial information [2]. Geomasking methods are evaluated by their capability to maintain a high utility of the spatial information and their associated risk of re-identification [9, 10]. It has been shown that most geomasking methods either result in too much information loss or do not protect privacy adequately [33]. The most promising methods for maintaining a high utility and having an associated low risk of re-identification are methods that release contextual data instead of the actual coordinates [33]. Current methods in this category [22, 34] are computationally complex and/or trim large distances. Our proposed geomasking method preserves utility and privacy and does not have the abovementioned drawbacks.

The method proposed in this paper is based on randomized geometric surface calculations, more precisely on the surface of a two-dimensional triangle, to encode distance matrices. A surface area-based value replaces the geographical distance between coordinates. Thus, ensuring that the distance cannot be used to identify the original location but the proxy (contextual information) allows the use of standard spatial statistical methods such as spatial clustering methods.

The paper is organized as follows. First, we give an overview of the current research background. Second, we formally introduce the proposed method and provide a step-by-step workflow. This section is supported by Appendix A (relationship between distance and expected surface area) and Appendix B (empirical simulation of the distance–area relationship). To test the method's utility, we next describe the data that will be considered in the application and introduce metrics to evaluate the method. The results are presented in Section 6, including several performance evaluation results, a hardening method against attacking approaches, and an evaluation of the re-identification risk. Lastly, we discuss the method and end with a conclusion.

2 Background

Geoprivacy has been an important topic in various fields. For a short introduction, see, e.g., [14]. In the past, many geographical masking methods have been proposed to ensure geoprivacy while still allowing researchers to use spatial information. According to [14], geographic masking methods are commonly subsumed into aggregation (examples are

given by [2, 20, 42]), coordinate modification (examples are given by [2, 15, 35, 38]), and releasing contextual data. In the latter, a set of contextual variables are attached to the individual microdata while removing the precise locations from the dataset [14]. Examples of how releasing contextual data works in practice are give by [19, 22, 34]. However, aggregation methods result in major information loss due to the fact that distributed points in an area are assigned the same value. Thus, for example, accessibility analyses are meaningless. Therefore, only perturbation methods or methods providing contextual information are potentially acceptable for the above-stated examples. Recently, it has been shown that even perturbation methods do not prevent identifying the unmasked, original location [33]. However, releasing contextual data seems promising, but the number of existing methods is limited.

One of the two most current methods is Lipschitz embedding for anonymizing geographical distance matrices proposed by [22]. This method proposes to release only the distance matrix with perturbed distances by using two parameters, dimension and size, which affect the variance of approximated distances. The second method, proposed by [34], is based on intersecting sets of randomly labeled grid points. This method also uses two parameters: the size of the radius and the number of sampled grid points, which affect the precision of approximated distances. A real-world implementation has been demonstrated by [19].

In contrast to the methods by [22,34], the method proposed in this paper has the advantage that only one parameter is required in the anonymization process. That is the number of sampled grid points. Thus, finding the optimal parameter needs fewer resources. In contrast to [34], large distances are not trimmed, which is needed if the relation between the points of the given dataset is of interest. Furthermore, the proposed method is far less computationally complex, thus allowing it to be applied to large datasets with less computational resources.

Another method, also based on geometric surface calculations, is proposed by [25] and called 'triangular displacement'. Based on a series of questions about the sensitivity of the data, a minimum and maximum displacement distance is set. Based on random numbers between the minimum and maximum displacement distances and the Pythagorean equation, two displacement distances are determined by adding or subtracting from the coordinate parts [25]. Thus, this method is a coordinate modification method, while we propose replacing the coordinate with a distance proxy.

3 Method

As outlined in Section 2, the proposed method aims to replace original distances with proxy information. In the following, we will explain and demonstrate the method to obtain the proxy. For this purpose, we consider matrix $\boldsymbol{D}$ with dimensions $m \times p$, and element d_{ij} representing a geographic distance between two coordinates in a two-dimensional space. Further, we introduce matrix $\tilde{\boldsymbol{D}}$ with dimensions $m \times p$ and element $\tilde{d}_{ij}$ representing the proxy. For the proxy, we propose to use the average surface area $\bar{A}$, and define $\bar{A} = \tilde{d}_{ij}$. This surface area is based upon a two-dimensional triangle that is constituted by a pair of existing coordinates and a third random pair of coordinates.

The distance between the two coordinates is a measure of length, for example, the Euclidean distance. The surface area is a measure of area typically associated with two-

dimensional objects, such as triangles. The relationship between distance and surface area depends on the specific geometric shape. In the context of a triangle, the lengths of its sides will affect the surface area. However, this relationship involves more than just a single distance; it requires consideration of all three side lengths and possibly additional information such as angles. With the proposed method, one side of the triangle is given, and the random pair of coordinates determines the other two.

The proposed method is based on the idea that smaller distances will also yield smaller surface areas associated with these distances, and larger distances will yield larger surface areas associated with these distances. That is, if $d_1 < d_2 < ... < d_n$, then $\bar{A}_1 < \bar{A}_2 < ... < \bar{A}_n$, and vice versa if $d_1 > d_2 > ... > d_n$ then $\bar{A}_1 > \bar{A}_2 > ... > \bar{A}_n$. This statement assumes a linear relationship; if the first distance is smaller than the nth distance, then the average surface area associated with the first distance will also be smaller than the average surface area associated with the nth distance. Or vice versa, if the first distance is larger than the nth distance, then the average surface area associated with the first distance will also be larger than the average surface area associated with the nth distance.

The surface areas $A_1, ..., A_n$ are not solely determined by the distances $d_1, ..., d_n$ alone, as each area also depends on the height of the corresponding triangle, which is a function of the randomly drawn point R_n. However, since the random points are sampled uniformly from a bounding box designed to avoid geometric bias, the expected value of the height becomes approximately independent of the baseline distance. Thus, the expected surface area $\bar{A}$ increases proportionally with the base distance d. Appendix A provides a mathematical justification. Appendix B empirically validates the mathematical assumption. In the following, we describe the workflow step-by-step. The workflow consists of five steps. The first three steps of the method are visualized in Figure 1. The entire example is programmed in the statistical programming language R [30].

Step 1 – Generate bounding box (optional)

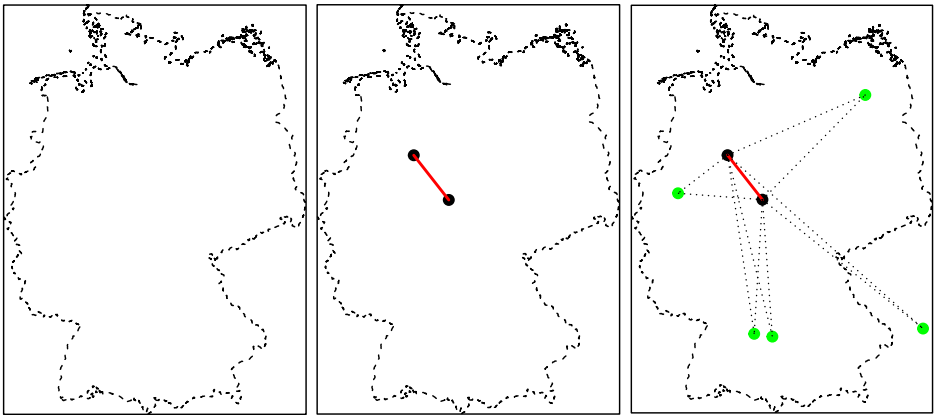
First, enlarge the area considered for computation by creating a bounding box around the geographical boundaries (see Figure 1a). As a result, the shape of the original geographical area has no potential effects. The method proposed by [34] showed that the quality of the distance approximations increased when the potential effects of the shape were diminished.

Step 2 – Calculate true distances

Consider a two-dimensional space $\mathbb{R}^2$ with two coordinates $X(x_1, x_2)$ and $Y(y_1, y_2)$ and d_{XY} being the geographical distance between X and Y. This is shown in Figure 1b. $\mathbb{R}^2$ is defined by the bounding box around the geographical borders of Germany. X and Y are the two geo-locations within $\mathbb{R}^2$ (black dots). In Figure 1b, d_{XY} is depicted by the red line connecting the two black dots (X and Y).

Step 3 – Draw random coordinates and obtain triangles

Draw a sample of uniformly distributed random coordinates $R_1, ..., R_n \in \mathbb{R}^2$ of size n. The random coordinates are required to obtain the triangles $\triangle XYR_n$. For this example, we draw a random sample with $n = 5$, thus creating five triangles. Figure 1c shows X, Y,



(a) Step 1 (b) Step 2 (c) Step 3

Figure 1: Step 1 shows the enlarged area considered for computation by creating a bounding box.

d_{XY}, $R_1, ..., R_5$, and triangles $\triangle XYR_1, ..., \triangle XYR_5$. For each d_{XY}, a different sample of n random coordinates is drawn. If an R_n coincidentally results in a surface area of 0, for example, when R_n lies on d_{XY}, a new R_n should be drawn. After completing steps 1–3, the next step will be calculating the surface areas required to obtain the distance proxy.

Step 4 – Calculate surface areas

With X and Y as base, calculate the surface area $a_{XYR_n} = \frac{1}{2} d_{XY} h_{XYR_n}$. The geographical distance d_{XY} is used as the base side of the triangle and calculated as

$$d_{XY} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{1}$$

h_{XYR_n} is the height of the base side, which depends on the location of R_n, and is calculated as

$$h_{XYR_n} = \frac{||(x_2 - x_1)(y_1 - y_{R_n}) - (x_1 - x_{R_n})(y_2 - y_1)||}{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}. \tag{2}$$

The results for the example are shown in Table 1. The results in Table 1 are not in meters or any interpretable distance unit because the coordinates used are in decimal degrees (longitude/latitude). Degrees are not linear units like meters or feet—they are angular. Before calculating distances and areas, they were not converted into a linear system. For the example and the method itself, this is not a problem because there is no entitlement for the derived proxy information to be in an easily interpretable unit, like meters.

From Step 4 and the Equations 1 and 2, it is evident that the area of the triangle a_{XYR} depends directly on the distance d_{XY}. Thus, if d_{XY} increases, the triangle $\triangle XYR_n$, and respectively the surface area of this triangle increases, because it is directly proportional to the baseline d_{XY}. From this, it follows that as the Euclidean distance between two points in a two-dimensional space (which forms the baseline of a triangle) increases, the area of the triangle also increases.

R_n	d_{XY}	h_{XYR_n}	a_{XYR_n}
R_1	1.361	1.731	1.178
R_2	1.361	5.516	3.754
R_3	1.361	2.713	1.846
R_4	1.361	3.688	2.510
R_5	1.361	2.657	1.808

Table 1: Base side of triangle (d_{XY}), height of the base side (h_{XY}), and surface area (a_{XYR_n}).

Step 5 – Calculate the proxy

Calculate the arithmetic mean $\bar{A} = \frac{1}{n}\sum_{i=1}^{n} a_{XYR_n} = \tilde{d}_{ij} = 2.219$ using the information from Table 1 and store in $\tilde{\boldsymbol{D}}$. Finally, repeat Steps 2 to 5 for each distance to be masked. The result is the complete distance matrix $\tilde{\boldsymbol{D}}$ with dimensions $m \times p$ and element $\tilde{d}_{ij}$ representing the proxy.

4 Data

Data from Germany (GER) and the Netherlands (NL) will be used to demonstrate the proposed method. The two countries differ considerably in size: the land mass of the Netherlands is about 12 % of the size of Germany (Germany 357,588 km^2, the Netherlands 41,850 km^2). Hence, the distribution of the distances between the points of interest differs considerably: there are smaller distances in the Netherlands than in Germany. For details, see Table 2. Thus, we can analyze how the method performs concerning differences in the underlying geographical data.

Country	Min.	25^{th} Quartile	50^{th} Quartile	Mean	75^{th} Quartile	Max.
Germany	0.11	199.62	315.03	320.39	431.35	875.67
Netherlands	0.01	53.90	88.03	95.35	131.24	323.88

Table 2: Distribution of geographical distances (Haversine distance in km) for Germany and the Netherlands.

We use open data from OpenStreetMap and publicly available register data to guarantee reproducibility. For Germany, the German hospital register is used [6]. The addresses are available online and have been geo-coded with `R` [30] using `tmaptools` [39]. The dataset contains $2,322$ locations of hospitals in Germany. As a proxy for patient residential addresses, the locations of 261 'general stores' from OpenStreetMap are used and provided by [12]. For the Netherlands, only OpenStreetMap data provided by [12] is used. Here, $3,006$ schools are selected. The 292 'kiosks' locations are used as a proxy for student residential addresses. We consider using general stores and kiosks as a proxy for residential addresses a feasible approach, given that these small shops are usually located within or close to residential areas. For details on the OpenStreetMap data, see [31]. The shapefiles of the geographical boundaries used in this study are publicly available and obtained from [11]. Figure 2 shows the geographic locations (bounding box not shown).

Data from both countries will be analyzed separately. In both analyses, the distances between all geo-locations are considered. Those are for Germany, $2,322 \times 261 = 606,042$, and for the Netherlands, $3,006 \times 292 = 877,752$ geographical distances.

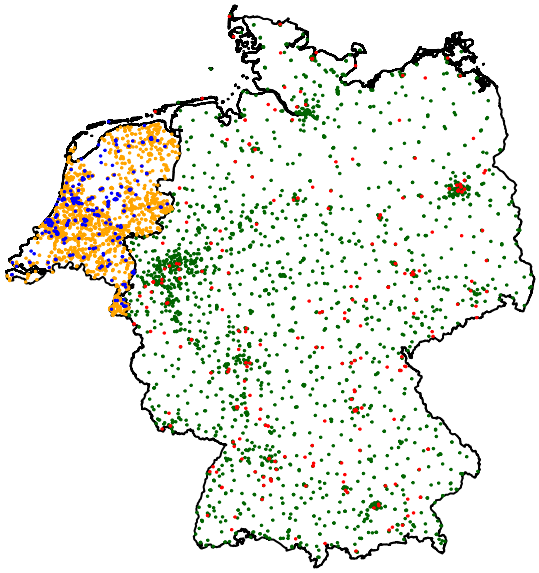


Figure 2: Geographical boundaries of Germany and the Netherlands. The spatial distribution of hospitals (green dots) and general stores (red dots) in Germany and schools (orange dots) and kiosks (blue dots) in the Netherlands are shown.

5 Evaluating the method

The evaluation of the utility of the proposed method consists of various steps. When using this method in practice, these evaluation methods can also be used to find the number of random points R_n needed, given the required amount of accuracy in the proxy. First, Pearson's correlation coefficient is used to evaluate the relationship between the true distance and the proxy. As true (geographical) distances, the Haversine and the Vincenty distances are used. The latter is computationally more intensive but more accurate as it assumes an elliptic shape of the globe compared to the Haversine distance assuming a sphere [16, 40]. Often, these two measures yield almost identical results [23, 28]. The results using only one sampled point are compared to those using up to 300 sampled points to compare for the influence of the number of sampled points (thus the number of triangles used to compute the average surface). We use the relative root mean squared error (RRMSE) to assess the proxy's performance. The RRMSE is obtained as

$$\text{MSE} = \frac{1}{mp} \sum_{ij} \left(\tilde{d}^*_{ij} - d^*_{ij} \right)^2$$

$$\text{RRMSE} = 100 \frac{\sqrt{\text{MSE}}}{\frac{1}{mp} \sum_{ij} d^*_{ij}}$$

The superscript '$*$' denotes that min-max normalized values are used. Given the different units for the geographical and proxy values, the RRMSE is based upon min-max normalized values (feature scaling), which linearly transforms both distances measures d_{ij} and $\tilde{d}_{ij}$ into the interval [0, 1]. By this, the units are made dimensionless and comparable.

$$d^*_{ij} = \frac{d_{ij} - \min(d_{ij})}{\max(d_{ij}) - \min(d_{ij})}, \tilde{d}^*_{ij} = \frac{\tilde{d}_{ij} - \min(\tilde{d}_{ij})}{\max(\tilde{d}_{ij}) - \min(\tilde{d}_{ij})}$$

In the second step, it is evaluated if the results are consistent for smaller and larger distances. Again, Pearson's correlation coefficient is used. However, in this step, the geographical distances are divided into two groups (smaller than the mean distance and larger than the mean distance), and the geographical distances are compared by group to the proxy. For the means, see Table 2.

Third, to study the method's variance and convergence, a Monte Carlo simulation is conducted. We generated $1,000$ random inputs for different numbers of sample points chosen (thus, the number of triangles). We calculate the proxy's sample variance and evaluate the number of iterations needed for convergence. Ideally, the results should show a small variance, and convergence should be achieved quickly.

Finally, it is evaluated if the distribution of the proxy remains similar to the distribution of the true distance. The Wasserstein Distance (also known as Earth Mover Distance) is used to do so. The Wasserstein distance reveals the minimum 'cost' needed to get from one distribution to another. Thus, similar distributions have a lower Wasserstein distance than dissimilar distributions. Here, the min-max normalized values are also used.

6 Results

6.1 Correlations and performance

Figure 3 shows for Germany and the Netherlands separately the Pearson correlation coefficient between the true distances and the proxy for the different number of sampled points used. For the true distance, the Haversine and the Vincenty distances are compared. There is no noticeable difference in the correlation coefficient when using the Haversine and Vincenty distance. Therefore, the Haversine distance will be used in subsequent analyses given its less computational effort.

Comparing the results by the number of random points (R_n) used, one random point achieved a moderate and positive correlation (r=0.49 Germany, r=0.58 Netherlands). The correlation increases continuously with an increasing number of random points. A considerable increase can be observed up to and including ten random points (r=0.83 Germany,

r=0.90 Netherlands). After ten random points, the correlation increases only minorly, as seen from the flattening curve. Considering the Haversine distance, the largest correlation in this experimental setting for the German and Dutch data is achieved with 300 random points (r=0.93). Considering the Vincenty distance, the largest correlation in this experimental setting is achieved with 200 points for Germany (r=0.93) and 300 points for the Dutch data (r=0.98). It can be expected that the correlation will also increase with an increasing number of random points. However, the results show that a strong correlation can be achieved with a small number of random points. Comparable results were found for the Spearman correlation coefficient (results not shown). Thus, the method preserves the order reasonably well.

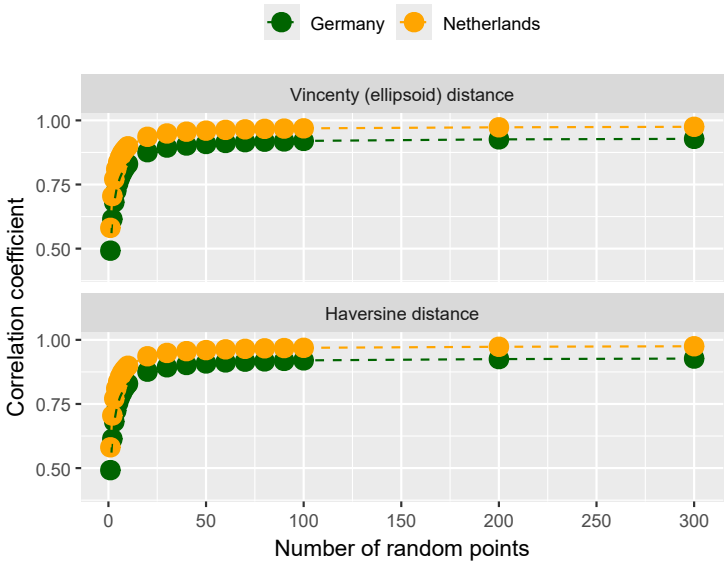


Figure 3: Correlation coefficient between geographical distance measures (Haversine distance in the top panel, Vincenty (ellipsoid) in the bottom panel) and proxy split by country (Germany in green, the Netherlands in orange). The x-axis shows the number of randomly sampled points used to obtain the proxy and the y-axis shows the correlation coefficient.

Moreover, it can be seen that the correlations obtained using the data from the Netherlands are larger than those obtained from Germany, indicating that small distances might be better preserved than larger distances. Thus, we grouped distances into smaller and larger distances and calculated group-wise correlations. Smaller distances were defined as being below or equal to the mean distance (the used means are shown in Table 2). Larger distances were defined as being larger than the mean distance. The results are shown in Figure 4. The lowest correlations are obtained for the large distances in Germany. The small distances in Germany achieve about the same correlations as the large distances in the Netherlands. The highest correlations are found for the small distances in the Netherlands. Thus, while the method yields good correlations, the method preserves smaller distances better than larger distances.

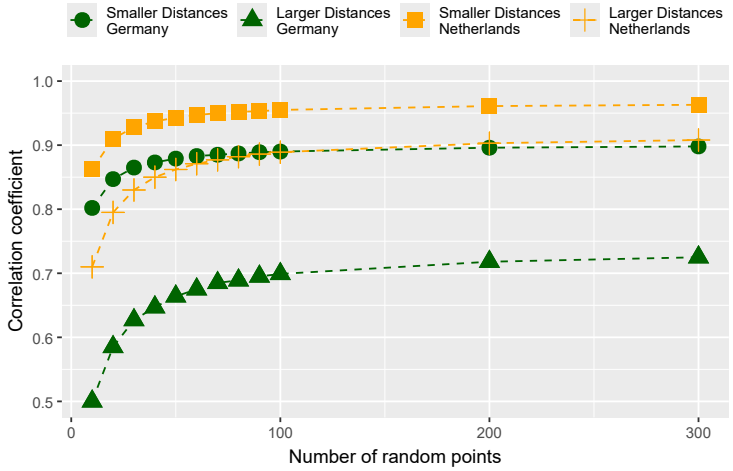


Figure 4: Correlation coefficient between geographical distance measures (Haversine distance) and the proxy, averaged over number of random points and split by country (Germany in green, the Netherlands in orange).

To measure the performance of the proxy, we used the RRMSE (shown in Figure 5). With only one random point, an RRMSE of about 100 % for both datasets is achieved. However, the RRMSE decreased considerably in both datasets when the number of random points increased. The RRMSE becomes stable at around 100 random points. With more than 100 random points, no further substantial decrease in the RRMSE was observed. With 300 random points, the RRMSE is about 18 % for Germany and the Netherlands (see Section 7 for a discussion on this result).

6.2 Monte Carlo simulation

For the Monte Carlo simulations, the Haversine distance was used, given that comparable results as with the Vincenty (ellipsoid) distance are achieved with less computational effort. Table 3 shows the point estimates based on the original data and the MC-based mean, standard error, and confidence interval. Only the results with one random point for the Netherlands and 300 random points for Germany are shown. The MC mean is close to the point estimates (rounded values are shown). The standard errors are near zero, and the confidence intervals are narrow.

Country	Random points	Point estimate	MC mean	MC standard error	MC confidence interval
Germany	300	0.93	0.93	0	[0.93, 0.93]
Netherlands	1	0.58	0.58	0	[0.58, 0.58]

Table 3: Results of Monte Carlo simulation. The number of random points, the point estimate, the MC mean, the MC standard error, and the MC confidence interval are shown.

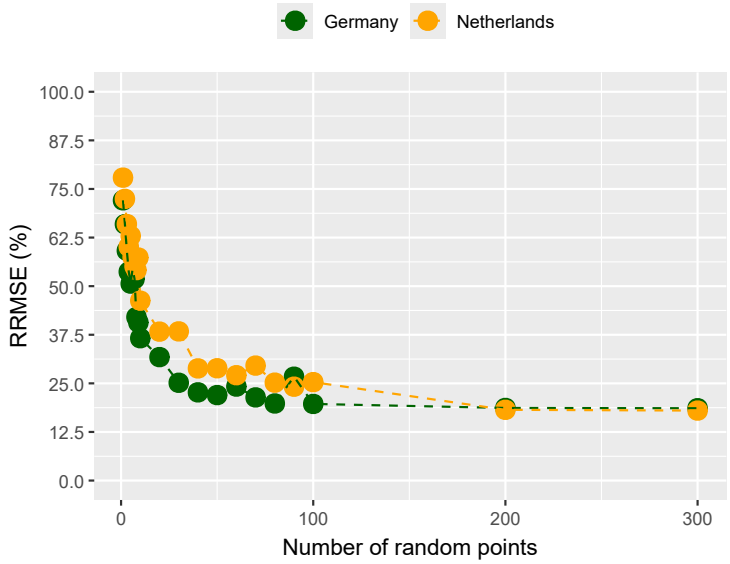
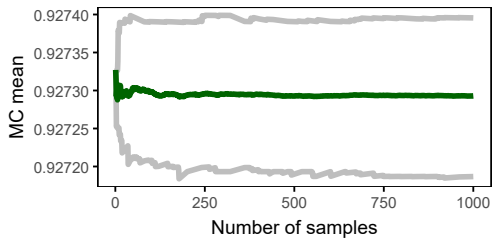
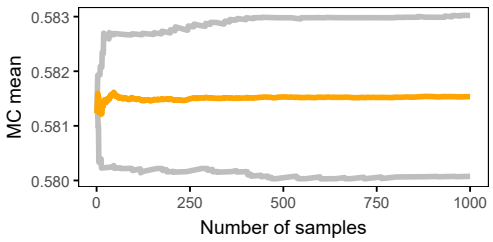


Figure 5: Relative root mean squared error split by country (Germany in green, the Netherlands in orange).

The convergence of the estimates was also analyzed. Figure 6 shows that the point estimates already became stable around 50 iterations. That holds for both the results based on one and 300 random points. Hence, the MC mean converges fast to the true mean. The upper and lower limits of the MC confidence interval limits take up to 500 iterations to converge. These findings are evident both for datasets. Thus, the method yields small variances and converges fast.



(a) German dataset and 300 random points. The green line shows the point estimate, and the gray lines show the lower and upper confidence interval limits.



(b) Dutch dataset and one random point. The orange line shows the point estimate, and the gray lines show the lower and upper confidence interval limits.

Figure 6: Convergence plots based on 1,000 Monte Carlo simulations. The left panel shows the German dataset and the right panel shows the dataset of the Netherlands. The X-axis shows the number of samples, and the y-axis shows the MC mean.

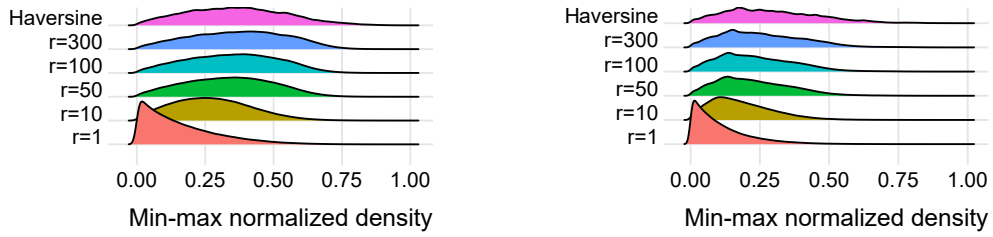
6.3 Wasserstein distance

Finally, we report the Wasserstein distance for both datasets for the proxy based on one, 100, and 300 random points. The results are shown in Table 4.

Country	Random points	Wasserstein
Germany	1	0.206
Germany	100	0.014
Germany	300	0.014
Netherlands	1	0.184
Netherlands	100	0.057
Netherlands	300	0.035

Table 4: Wasserstein distances by country and number of random points.

As seen for both countries, the minimum 'cost' needed to get from one distribution to another is much smaller, with 300 points, than with one random point as seen by the Wasserstein distance close to zero. Hence, the more random points, the closer the proxy distribution is to the original distance distribution. This is shown in Figure 7.



(a) German dataset.

(b) Dutch dataset.

Figure 7: Distribution plots of min-max normalized true distances (Haversine), and min-max normalized proxy values for different numbers of random points. The left panel shows the German dataset and the right panel shows the dataset of the Netherlands. The X-axis shows the density; the y-axis shows different sets of random points and the original distance.

6.4 Hardening against attacking approaches

Since no coordinates are released, it remains unknown in which area the locations are, and since no true distances are released, it cannot be approximated straightforwardly. However, especially in rural areas spatial points are prone to being re-identified easily. To reduce the risk a population-density-based error term is introduced for low-population density areas.

We tested this approach using the German dataset and population information on NUTS-3 level, which is a classification of administrative and municipal districts. For each municipal district (M, with $m = 1, ..., 417$), the population totals (U_m) and population densities per km^2 (D_m) are obtained from official statistics [37]. We tested the following three error terms to account for the low population density areas when obtaining the proxy:

$$\bar{A}_1 = \begin{cases} \bar{A} \cdot \frac{1}{U_m} & \text{if } U_m \in [P_{10}, P_{20}], \\ \bar{A} & \text{otherwise.} \end{cases}$$

$$\bar{A}_2 = \begin{cases} \bar{A} \cdot \frac{1}{\log(D_m)} & \text{if } U_m \in [P_{10}, P_{20}], \\ \bar{A} & \text{otherwise.} \end{cases}$$

$$\bar{A}_3 = \begin{cases} \bar{A} \cdot \frac{1}{\log(U_m)} & \text{if } U_m \in [P_{10}, P_{20}], \\ \bar{A} & \text{otherwise.} \end{cases}$$

When U_m is within the 10^{th} or 20^{th} percentile of its distribution, the proxy is multiplied by an error term. Otherwise, the proxy without an error term is used. Thus, ensuring that only low population densities are penalized. Figure 8 shows the results.

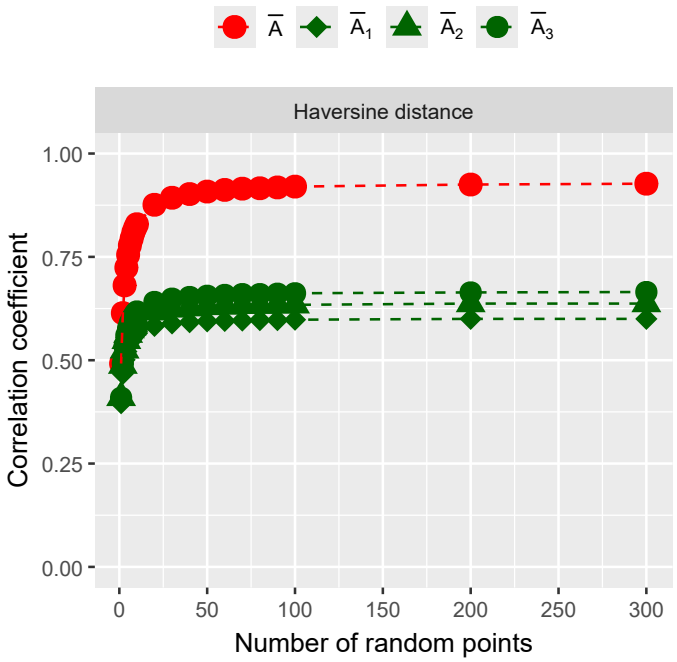


Figure 8: Correlation coefficient between Haversine distance and different versions of the proxy. The x-axis shows the number of randomly sampled points used to obtain the proxy and the y-axis shows the correlation coefficient. The color and shapes indicate the version of the proxy information.

The correlation coefficient between the original data and the proxy without the error term is shown in red as a benchmark. Adding an error term affects areas with lower population density, reducing the correlation coefficient. Areas with higher population densities are not penalized and, thus, do not contribute to the decreasing correlation. Differences between the error terms are small. With 300 random points, the correlation coefficient for

the proxy without error term ($\bar{A}$) is 0.93 (see also Table 3). For $\bar{A}_1$, $\bar{A}_2$, and $\bar{A}_3$, the coefficients are 0.6, 0.64, and 0.67 (with 300 random points). These results demonstrate that all approaches distort the relationship between the original distances and the proxy in areas with lower population densities. While the decrease in correlation is considerable, the resulting coefficients still indicate a moderate-strength relationship.

6.5 Re-identification risk

The risk of re-identifying a geomasking method is usually assessed using spatial k-anonymity, defined as the number of potential locations in a defined region for a given masked location [41]. Such a method is not applicable to a geomasking method that does not release coordinates but contextual information. However, an individualized attack method is usually used for those methods, see, e.g., [21].

We conducted a re-identification risk assessment to evaluate the proposed method's privacy-preserving capability. The approach assumes that an attacker attempts to reconstruct the original geographic distances from the proxy values (i.e., average triangle surface areas) using machine learning techniques. It is also assumed that a potential attacker accessed a fraction of the original dataset. The leaked data includes the true Euclidean distances and their corresponding proxy values. We considered this dataset as training data. A predictive model (random forest using default settings) was trained using this dataset. The trained model was then used to estimate the remaining true distances (test set). The accuracy of this prediction was assessed using the Mean Absolute Error (MAE) between the true and the predicted distance. This metric informs about the potential privacy leakage: lower predictive accuracy indicates higher privacy protection. The simulation was performed using the following input parameters:

- **`Random points`**: A numeric vector specifying the number of random points that were used to generate the proxy in the leaked data. Values used: `1` and `300`.
- **`Leakage fraction`**: A numeric vector representing the proportion of the data leaked to be used for training the model. Values used: 1 % and 10 %.
- **`Hardened`**: A logical vector indicating whether the hardening method discussed in Section 6.4 was used. Values used: `TRUE` and `FALSE`.

The results of this attacking scenario are shown in Figure 9. Each panel compares the predicted geographic distances (via a machine learning model) against the true observed distances. Since we used a large dataset and points can overlap in the figure, the density of points is shown in a gradient from yellow (few observations) to blue (many observations). Each subplot varies along three key dimensions: Number of random points used to compute the proxy (1 or 300), fraction of training data leaked (1 % or 10 %), and whether the hardening method was applied (TRUE or FALSE). Each plot also reports the Mean Absolute Error (MAE), a measure of prediction accuracy (lower MAE = higher risk, i.e., better re-identification).

If more random points are used, the relationship between the proxy and the original distance will be much closer, which enables more accurate re-identification (lower MAE), thus compromising privacy. However, an MAE of 51 km is still large, and the re-identification risk is still small (in scenarios without hardening). The hardening method increases the MAE, indicating a reduced risk of re-identification. The hardening method is especially

effective when a few random points are used to create the proxy. Leakage impact is minimal with 1 point or hardening. Increasing the training data leak from 1 % to 10 % does not affect MAE when either only one random point is used or hardening is applied. Our results demonstrate that despite the underlying monotonicity assumption, the proxy transformation introduces sufficient geometric ambiguity (see the MAE values) to prevent accurate distance reconstruction, thereby contributing to the privacy-preserving capability and data confidentiality.

Moreover, even if the underlying true distance is predicted, information about the data's general location is still needed to transform the distance matrix to coordinates.

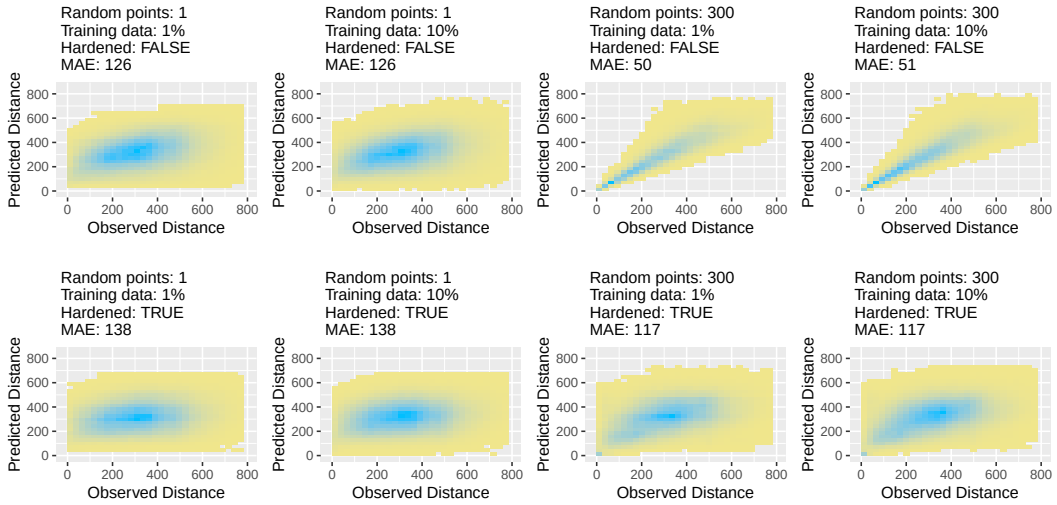


Figure 9: Re-identification risk assessment based on adversarial inference split by different scenarios of the geomasking method and by different attacking scenarios.

7 Discussion and conclusion

Geomasking methods aim to maintain the use of spatial information while reducing the risk of re-identification. To demonstrate the usefulness of the masked spatial information, we evaluated (1) the relationship between the true distance and the proxy, (2) the consistency of the method for smaller and larger distances, (3) the variance and convergence of the method and (4) the comparison of the distribution of the true distance and the proxy distances. The results showed that the obtained proxy strongly correlates with the true distance and that the underlying distribution of the original distances is preserved. Thus, the proxy information can be used in clustering and spatial autocorrelation methods, for example.

The results achieved by the method are of good quality. An RRMSE of about 18 % was achieved for the German dataset. Relating the 18 % error to the four minimum and maximum values provided in Table 2 results in errors between 0.0018 km and 157.2 km. To consider this error in a practical example, we take the ten nearest hospitals per residential

address and average these distances over all residential addresses. The mean distance to the ten nearest hospitals would be about 3 km for the Netherlands and about 13 km for Germany. Assuming an average speed of an ambulance of 60 km/h, the time needed for these distances would be 3 minutes for the Netherlands or about 13 minutes for Germany. Considering the error of 18 %, it would be about 0.5 km more or less for the Netherlands and about 2.3 km more or less for Germany. For the Netherlands, this would reduce the required time to about 2.5 minutes or increase it to about 3.5 minutes, and for Germany, it would reduce the required time to about 10.7 minutes or increase it to 15.3 minutes. In non-emergency cases, differences of less than 30 minutes are typical without consequence on patient outcome [17,18,29].

As no coordinates are released and no true distances are released, it remains unknown which area the locations are in, and it cannot be approximated straightforwardly. However, given that the algorithm is known, it is possible to simulate the relationship between the original distances and the proxies if the coordinates of one source are known, e.g., the points of interest. By comparing the given proxies with the simulated proxies, the original distances could be derived. This could lead to potential re-identifications, especially in regions with low population density. The fact that the bounding box is unknown makes this more difficult. As a solution, a population-density-based error term could be introduced when calculating the proxy to overcome this problem. By this, the proxies in areas with lower population densities get a penalty term. A test of this hardening method showed that this will slightly distort the relationship between the proxy and the original distance only for rural areas.

Our evaluation of the re-identification risk using machine learning techniques to reconstruct the original distance from the proxy values showed that sufficient geometric ambiguity is introduced, especially with the hardening method, so an accurate distance reconstruction is not achieved. Other attacking approaches, such as graph-theoretical approaches, can, in principle, be used as well to attack every geo-masking method when external information about the data is available [21,22]. A comprehensive risk analysis of the proposed method is an ongoing project of the authors.

To conclude, we proposed and discussed a new method to preserve spatial privacy in large distance matrices with medical and social microdata examples. Of course, the method applies to other spatial datasets as well. The method is straightforward to implement and requires neither excessive computational effort nor exceeding memory requirements. We consider this method a valuable enhancement of the methodological toolbox when working with masking methods for geolocation data.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of their affiliations. The authors would like to thank the three anonymous referees for careful reading and providing useful comments on a former draft of this manuscript.

Data availability statement

Data from OpenStreetMap and publicly available register data were used in this study to guarantee reproducibility. The German hospital register was used; data is available here:

[6]. OpenStreetMap data is available here: [12]. Shapefiles of the geographical boundaries are available here: [11]. Materials to reproduce the results of this article are available on GitHub [https://github.com/jkwort/ppdc.git].

References

[1] AMBROGGI, M., BIASINI, C., DEL GIOVANE, C., FORNARI, F., AND CAVANNA, L. Distance as a barrier to cancer diagnosis and treatment: Review of the literature. *Oncologist 20*, 12 (2015), 1378–1385. doi:10.1634/theoncologist.2015-0110.

[2] ARMSTRONG, M. P., RUSHTON, G., AND ZIMMERMAN, D. L. Geographically masking health data to preserve confidentiality. *Statistics in Medicine 18*, 5 (1999), 497–525.

[3] BANERJEE, S. Spatial data analysis. *Annual Review of Public Health 37*, 1 (2016), 47–60. doi:10.1146/annurev-publhealth-032315-021711.

[4] CHRISTEN, P., RANBADUGE, T., AND SCHNELL, R. *Linking Sensitive Data*. Springer International Publishing, Cham, Switzerland, 2020.

[5] CURRIE, J., AND REAGAN, P. B. Distance to hospital and children's use of preventive care: Is being closer better, and for whom? *Economic Inquiry 41*, 3 (2003), 378–391. doi:10.1093/ei/cbg015.

[6] DEUTSCHE KRANKENHAUS TRUSTCENTER UND INFORMATIONSVERARBEITUNG GMBH. Deutsches Krankenhausverzeichnis, 2022. https://www.deutsches-krankenhaus-verzeichnis.de/app/suche/landkarte.

[7] D'HAESE, S., VAN DYCK, D., DE BOURDEAUDHUIJ, I., DEFORCHE, B., AND CARDON, G. The association between objective walkability, neighborhood socio-economic status, and physical activity in belgian children. *International Journal of Behavioral Nutrition and Physical Activity 11*, 1 (2014), 1–8. doi:10.1186/s12966-014-0104-1.

[8] DOWNES, T. A., AND ZABEL, J. E. The impact of school characteristics on house prices: Chicago 1987–1991. *Journal of Urban Economics 52*, 1 (2002), 1–25. doi:10.1016/S0094-1190(02)00010-4.

[9] DUNCAN, G. T., AND FIENBERG, S. E. Obtaining information while preserving privacy: A Markov perturbation method for tabular data. In *Statistical Data Protection: Proceeding of the Conference*. Lisbon, 25 to 27 March 1998. Luxembourg: European Commission, Statistical Office of the European Communities, 1999, pp. 351—362.

[10] DUNCAN, G. T., KELLER-MCNULTY, S. A., AND STOKES, S. L. Disclosure risk vs. data utility: The R-U confidentiality map, 2001. Tech. rep. 121. National Institute of Statistical Science.

[11] GADM. GADM maps and data, 2023. https://www.gadm.org/index.html.

[12] GEOFABRIK GMBH. OpenStreetMap data extracts, 2023. https://download.geofabrik.de/europe/netherlands.html.

[13] GIBBONS, S., AND MACHIN, S. Valuing school quality, better transport, and lower crime: Evidence from house prices on JSTOR. *Oxford Review of Economic Policy 24*, 1 (2008), 99–119. doi:10.1093/oxrep/grn008.

[14] GUTMANN, M., WITKOWSKI, K., COLYER, C., O'ROURKE, J. M., AND MCNALLY, J. Providing spatial data for secondary analysis: Issues and current practices relating to confidentiality. *Population Research and Policy Review 27*, 6 (2008), 639–665. doi:10.1007/s11113-008-9095-4.

[15] HAMPTON, K. H., FITCH, M. K., ALLSHOUSE, W. B., DOHERTY, I. A., GESINK, D. C., LEONE, P. A., SERRE, M. L., AND MILLER, W. C. Mapping health data: Improved privacy protection with donut method geomasking. *American Journal of Epidemiology 172*, 9 (2010), 1062. doi:10.1093/aje/kwq248.

[16] HIJMANS, R. J. *geosphere: Spherical Trigonometry*, 2021. R package version 1.5-14.

[17] JANG, W. M., LEE, J., EUN, S. J., YIM, J., KIM, Y., AND KWAK, M. Y. Travel time to emergency care not by geographic time, but by optimal time: A nationwide cross-sectional study for establishing optimal hospital access time to emergency medical care in South Korea. *PLOS ONE 16*, 5 (2021), e0251116. doi:10.1371/journal.pone.0251116.

[18] KELLY, C., HULME, C., FARRAGHER, T., AND CLARKE, G. Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? a systematic review. *BMJ Open 6*, 11 (2016), e013059. doi:10.1136/bmjopen-2016-013059.

[19] KLINGWORT, J., SCHNELL, R., AND SIXT, M. Geo-masking of coordinates of BiLO respondents for future privacy-preserving distance calculations [in german: Geo-Masking von Koordinaten der BiLO Befragten für zukünftige datenschutzgerechte Distanzberechnungen]. Tech. Rep. 87, Leibniz-Institut für Bildungsverläufe, Bamberg, 2020.

[20] KOUNADI, O., AND LEITNER, M. Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets. *Computers, Environment and Urban Systems 57* (2016), 59–67. doi:10.1016/j.compenvurbsys.2016.01.004.

[21] KROLL, M. A graph theoretic linkage attack on microdata in a metric space. *ArXiv e-prints* (2014). doi:10.48550/arXiv.1402.3198.

[22] KROLL, M., AND SCHNELL, R. Anonymisation of geographical distance matrices via Lipschitz embedding. *International Journal of Health Geographics 15*, 1 (2016), 1–14. doi:10.1186/s12942-015-0031-7.

[23] LAWHEAD, J. *Learning Geospatial Analysis With Python: An Effective Guide to Geographic Information System and Remote Sensing Analysis Using Python 3*, 2 ed. Packt Publishing, Birmingham, 2015.

[24] MENDOZA, J. A., WATSON, K., NGUYEN, N., CERIN, E., BARANOWSKI, T., AND NICKLAS, T. A. Active commuting to school and association with physical activity and adiposity among US youth. *Journal of physical activity & health 8*, 4 (2011), 488. doi:10.1123/jpah.8.4.488.

[25] MURAD, A., HILTON, B., HORAN, T., AND TANGENBERG, J. Protecting patient geo-privacy via a triangular displacement geo-masking method. In *GeoPrivacy '14: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Privacy in Geographic Information Collection and Analysis*. New York, NY, USA, 2014, pp. 1–9. doi:10.1145/2675682.2676399.

[26] NGUYEN, H. L., TSOLAK, D., KARMANN, A., KNAUFF, S., AND KÜHNE, S. Efficient and reliable geocoding of German Twitter data to enable spatial data linkage to official statistics and other data sources. *Frontiers in Sociology 7* (2022). doi:10.3389/fsoc.2022.910111.

[27] NICHOLL, J., WEST, J., GOODACRE, S., AND TURNER, J. The relationship between distance to hospital and patient mortality in emergencies: An observational study. *Emergency Medicine Journal : EMJ 24*, 9 (2007), 665–668. doi:10.1136/emj.2007.047654.

[28] PANIGRAHI, N. *Computing in Geographic Information Systems*. CRC Press, Boca Raton, 2014.

[29] PHIBBS, C. S., AND LUFT, H. S. Correlation of travel time on roads versus straight line distance. *Medical Care Research and Review 52*, 4 (1995), 532–542. doi:10.1177/107755879505200406.

[30] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[31] RAMM, F. OpenStreetMap data in layered GIS format: Free shapefiles - 2022-04-29, 2022. http://download.geofabrik.de/osm-data-in-gis-formats-free.pdf.

[32] RAVELLI, A., JAGER, K. J., DE GROOT, M. H., ERWICH, J., DRIEL, G. C. R.-V., TROMP, M., ESKES, M., ABU-HANNA, A., AND MOL, B. Travel time from home to hospital and adverse perinatal outcomes in women at term in the netherlands. *BJOG: An International Journal of Obstetrics & Gynaecology 118*, 4 (2011), 457–465. doi:10.1111/j.1471-0528.2010.02816.x.

[33] REDLICH, S. *Quantitative Analysis of Geomasking Methods*. Doctoral dissertation, University Duisburg-Essen, 2022. doi:10.17185/duepublico/76045.

[34] SCHNELL, R., KLINGWORT, J., AND FARROW, J. M. Locational privacy-preserving distance computations with intersecting sets of randomly labeled grid points. *International Journal of Health Geographics 20*, 1 (2021), 14–16. doi:10.1186/s12942-021-00268-y.

[35] SEIDL, D. E., PAULUS, G., JANKOWSKI, P., AND REGENFELDER, M. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography 63* (2015), 253–263. doi:10.1016/j.apgeog.2015.07.001.

[36] SHARMA, S. *Data Privacy and GDPR Handbook*. Wiley, Hoboken, NJ, USA, 2019.

[37] STATISTISCHES BUNDESAMT. District-free cities and rural districts by area, population and population density on 31.12.2022 [german title: Kreisfreie Städte und Landkreise nach Fläche, Bevölkerung und Bevölkerungsdichte am 31.12.2022], 2023. https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/Administrativ/04-kreise.html.

[38] STINCHCOMB, D. *Procedures for Geomasking to Protect Patient Confidentiality*. Presented at the ESRI International Health GIS Conference held in Washington, DC, 2004.

[39] TENNEKES, M. *tmaptools: Thematic Map Tools*, 2021. R package version 3.1-1.

[40] VINCENTY, T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey Review 23*, 176 (1975), 88–93. doi:10.1179/sre.1975.23.176.88.

[41] WANG, J., AND KWAN, M.-P. Daily activity locations k-anonymity for the evaluation of disclosure risk of individual GPS datasets. *International Journal of Health Geographics 19*, 7 (2020), 1–14. doi:10.1186/s12942-020-00201-9.

[42] WOLF, M. K. Microaggregation and disclosure avoidance for economic establishment data. In *1988 Proceeding of the Business and Economic Statistics Section*. Papers Presented at the Annual Meeting of the American Statistical Association, New Orleans, Louisiana, August 22-25, 1988, pp. 355—360.

[43] YANTZI, N., ROSENBERG, M. W., BURKE, S. O., AND HARRISON, M. B. The impacts of distance to hospital on families with a child with a chronic condition. *Social Science & Medicine 52*, 12 (2001), 1777–1791. doi:10.1016/S0277-9536(00)00297-5.

Appendix A: Relationship between distance and expected surface area

We provide a mathematical justification for the assumption that the expected surface area $\bar{A}_{XY}$ of the triangle $\triangle XYR_n$, constructed from a fixed pair of points $X, Y \in \mathbb{R}^2$ and a uniformly random point $R_n \in \mathbb{R}^2$, increases monotonically with the Euclidean distance d_{XY} between X and Y.

Let:

- $d = d_{XY}$ be the fixed distance (the base of the triangle),
- $R_n = (x_R, y_R)$ be a uniformly drawn random point from a rectangular region (i.e., the bounding box),
- h_{XYR_n} be the perpendicular height from R_n to the line segment XY,
- $a_{XYR_n} = \frac{1}{2} d \cdot h_{XYR_n}$ be the area of the triangle $\triangle XYR_n$.

We define the expected area over n samples as:

$$\mathbb{E}[\bar{A}_{XY}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}d \cdot h_i\right] = \frac{1}{2}d \cdot \mathbb{E}[h],$$

where $h_i = h_{XYR_i}$, and $\mathbb{E}[h]$ is the expected height of the triangle given random sampling of R_n.

Key Assumption

We assume that the random points R_n are drawn uniformly over a region that does not systematically favor positions closer or farther from the segment XY, regardless of the length of d. The bounding box constructed in Step 1 is designed to ensure this.

Result

Under this assumption, the expected value $\mathbb{E}[h]$ is approximately independent of d, and hence:

$$\mathbb{E}[\bar{A}_{XY}] \propto d.$$

This means the expected surface area grows linearly with the base length. Therefore, for two pairs (X_1, Y_1) and (X_2, Y_2), if $d_{X_1Y_1} > d_{X_2Y_2}$, then $\mathbb{E}[\bar{A}_{X_1Y_1}] > \mathbb{E}[\bar{A}_{X_2Y_2}]$.

Conclusion

The expected proxy value $\bar{A}_{XY}$ preserves the rank order of original distances d_{XY}. While the individual triangle areas depend on both d and the stochastic height h, their expected value scales linearly with d under uniform sampling, justifying the use of $\bar{A}_{XY}$ as a distance proxy.

Appendix B: Empirical simulation of the distance–area relationship

To empirically validate the theoretical assumption that the expected triangle surface area increases linearly with distance, we conducted the following simulation in R. For multiple baseline distances d, we applied the proposed geomasking method and visualized the baseline distance and their yielding mean surface area using 1, 10, 50, 100, and 300 points. The R code for this simulation is as follows:

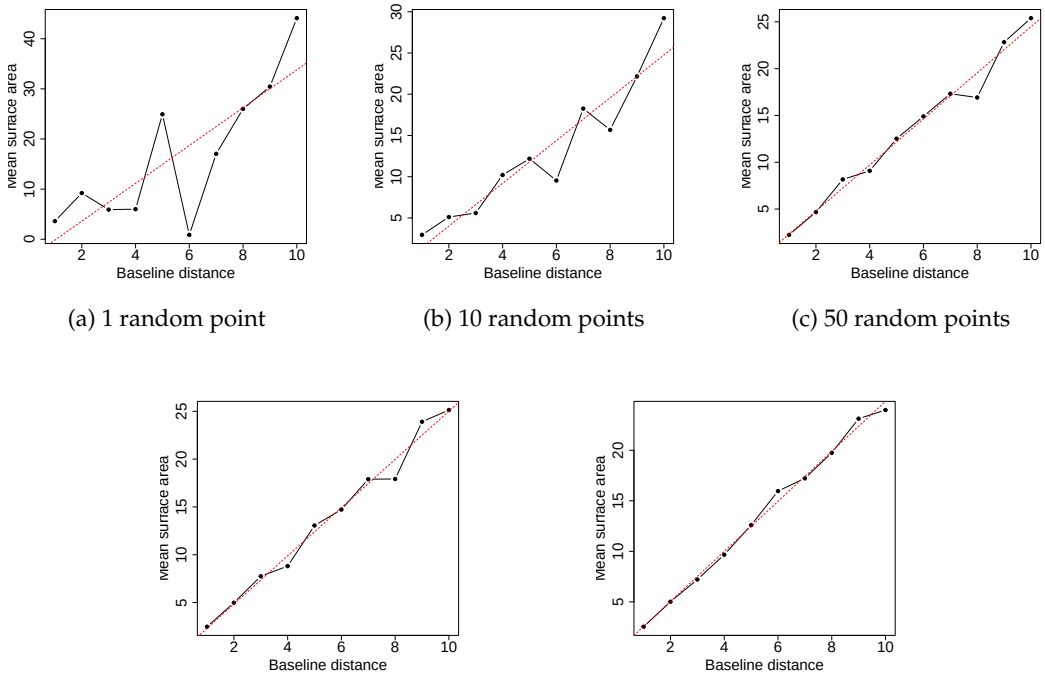
```r
# Set seed for reproducibility
set.seed(21082024)

# Function to compute area of triangle given two fixed points and one random point
triangle_area <- function(x1, y1, x2, y2, xr, yr) {
  d <- sqrt((x2 - x1)^2 + (y2 - y1)^2)
  h <- abs((x2 - x1)*(y1 - yr) - (x1 - xr)*(y2 - y1)) / d
  area <- 0.5 * d * h
  return(area)
}

# Simulation parameters
n_random <- 10     # number of random points per baseline
distances <- seq(1, 10, by = 1)  # varying baseline distances
mean_areas <- numeric(length(distances))

# Simulation loop
for (i in seq_along(distances)) {
  d <- distances[i]
  x1 <- 0; y1 <- 0
  x2 <- d; y2 <- 0  # horizontal segment of length d
  rand_x <- runif(n_random, -5, 15)
  rand_y <- runif(n_random, -10, 10)

  areas <- mapply(
    triangle_area,
    x1, y1, x2, y2,
    rand_x, rand_y
  )

  mean_areas[i] <- mean(areas)
}

# Plotting the results
plot(distances, mean_areas, type = "b", pch = 19,
     xlab = "Baseline distance",
     ylab = "Mean surface area",
     cex.axis = 2,
     cex.lab = 2)
abline(lm(mean_areas ~ distances), col = "red", lty = 2)
```

Listing 1: Simulation of proxy surface areas for distance masking

Results of empirical simulation of the distance–area relationship

Each panel is based on a different number of random points. The x-axis shows the baseline distance. The y-axis shows the average surface area. The red line is a fitted linear regression model where the average area is the dependent variable and the baseline distance is the independent variable.

As can be seen, even with a few random points, the mean average surface area increases with increasing baseline distances. Thus, if $d_1 < d_2 < ... < d_n$, then $\bar{A}_1 < \bar{A}_2 < ... < \bar{A}_n$.



(a) 1 random point (b) 10 random points (c) 50 random points

(d) 100 random points (e) 300 random points

Figure 10: Results of empirical simulation of the distance–area relationship for selected number of random points.