Kernel density estimation under masking of geolocations with applications to DHS data

Lorena Gril
Md Jamal Hossain
Nikos Tzavidis
Ulrich Rendtel

School of Business & Economics
Discussion Paper

Economics

2026/3

Kernel density estimation under masking of geolocations with applications to DHS data

Lorena Gril[1], Md Jamal Hossain[2], Nikos Tzavidis[2], Ulrich Rendtel[1]

Abstract

The availability of geocoordinates offers valuable insights into spatial patterns of economic, demographic and health outcomes. However, disclosing the exact geolocation of statistical units to secondary analysts contravenes the responsible use of data. To protect privacy, anonymisation methods are used. A commonly applied anonymisation method is the one used by Demographic and Health Surveys (DHS). The DHS anonymisation scheme works by first aggregating data at small spatial units followed by random (donut) displacement of the geocoordinates. It is reasonable for secondary analysts to be concerned about the impact of anonymisation on the analyses. In this paper, the DHS anonymisation scheme is used as a basis for studying how anonymisation impacts on kernel density estimation. We propose methodology to account for the impact of the anonymisation process on density estimation. The proposed methodology is based on deriving the distribution of the true coordinates given the observed (anonymised) coordinates. Density estimation is then implemented by using the theoretical distribution and an iterative algorithm that accounts for both aggregation and displacement. The aim is to approximate the original population density using generated pseudo-coordinates under the assumption that the anonymisation process is known. The proposed method is illustrated by using DHS data from the Rajshahi Division in Bangladesh to estimate the density of households below the poverty line. The results show that accounting for measurement error due to anonymisation leads to a more accurate picture of the spatial distribution of poverty.

Key words: Aggregation, Confidentiality, Measurement error, Random (donut) displacement.

1 Introduction

Georeferenced datasets which include information about individuals' or households' locations enable detailed spatial analyses of socioeconomic patterns. Access to precise spatial information enhances statistical analyses and provides valuable insights into the distribution of economic, demographic, and health outcomes. However, making the geolocation of the units of analysis available carries the risk of identification. To balance this risk against the utility of the data, anonymisation methods can be used to ensure that geocoded data can be safely published and used.

A globally recognised source of high-quality, nationally representative data with geographical information collected in over 90 countries is provided by the Demographic and Health Surveys (DHS). DHS collect information on, among other topics, poverty, fertility, maternal and child health, HIV/AIDS, malaria, and nutrition, which are critical for monitoring health, demographic and economic trends and forming public health and humanitarian policies. Smoothing techniques such as kernel density estimation are frequently used alongside geolocated information to provide a continuous representation of variables of interest and assist with the identification of spatial clusters. Generally speaking, two approaches are used for anonymising geolocated data: (a) modifying the geographic locations through so-called geomasking methods, or (b) introducing measurement error in the variables of interest at the given geographic coordinates. DHS ensures confidentiality by using geomasking techniques i.e., by modifying

[1]Freie Universität Berlin, Department of Economics - Chair of Applied Statistics, Garystr. 21, 14195 Berlin, Germany
[2]University of Southampton, University Road, Southampton SO17 1BJ, United Kingdom

the geographic coordinates while keeping the attribute (variable) values unchanged. Introducing error in geographic coordinates affects the use of smoothing techniques and spatial analysis of the the target variables. In this paper, we develop methodology to account for measurement error in kernel density estimation when analysing geolocated data that are perturbed by using geomasking techniques.

DHS inform data users about the type of geomasking mechanism applied to the data. The procedure, henceforth referred to as the DHS anonymisation scheme, employs the sequential application of two well-established anonymisation strategies: aggregation and random displacement. First, multiple households within a given area are aggregated to the centroid of a so-called Enumeration Area (EA). Next, the EA centroids are displaced by a randomly chosen angle and distance depending on whether the coordinate belongs to an urban or rural areas, as described by Burgert et al. (2013). Coordinates on which both aggregation and displacement is applied are referred to as anonymised. Random displacement by choosing a random angle and distance was first proposed by Stinchcomb (2004a) and Hampton et al. (2010) and applied by Lu et al. (2012) to anonymised street addresses, by Allshouse et al. (2010) in applications with health data, by Clifton and Gehrke (2013) in travel data applications, and by Kounadi and Leitner (2016) in applications with crime data. Several studies have made extensive use of DHS data to investigate health and socio-economic patterns in low- and middle-income countries, cf. Balk et al. (2004), Feldacker et al. (2010), Perez-Heydrich et al. (2016), and Lohela et al. (2012). Although the DHS anonymisation scheme preserves respondent confidentiality, it introduces measurement error in statistical analysis that makes the use of geolocated data challenging. Despite this, many studies have used the anonymised coordinates without accounting for the error, treating the released geolocations as true, cf. Balk et al. (2004), Pande et al. (2008), Feldacker et al. (2010), and Lohela et al. (2012). Other studies have sought to mitigate the effects of anonymisation. Based on the work of Perez-Heydrich et al. (2016) and Gething et al. (2015), DHS recommend averaging the covariate values within the maximum displacement area around the anonymised coordinate, assuming that all points within this area are equally likely to represent the true location. However, these approaches do not explicitly account for the underlying anonymisation mechanism in estimation and inference.

Warren et al. (2016) proposed a simulation-based approach that generates likely true coordinates by re-displacing the anonymised coordinates according to the DHS displacement mechanism. However, this simulation-based approach is not based on the correct distribution of the true coordinates given the observed coordinates under the anonymisation scheme. Altay et al. (2024) further improve the estimation of model parameters and prediction. While this approach accounts for the uncertainty introduced by random displacement, it does not provide an explicit mathematical formulation for the displacement distribution under the anonymisation mechanism. Additionally, it does not account for the nested anonymisation process used by the DHS, i.e., aggregation of the coordinates to the EA centroid and displacement of the centroids. It is worth noting that most studies that use anonymised DHS data have focused on parameter estimation, whereas comparatively little attention has been paid to the effect of anonymisation on density estimation. A related line of research has considered anonymisation arising from the aggregation or rounding of coordinates rather than random displacement. In this context, Groß et al. (2017) developed a multivariate non-parametric kernel density estimation approach based on a Monte-Carlo Markov Chain algorithm that treats aggregation as a measurement error process. The proposed methodology generates pseudo-coordinates that attempt to approximate the true but unobserved spatial distribution more precisely than the anonymised data.

We extend the work by Warren et al. (2016) and adapt the aggregation framework of Groß et al. (2017) by deriving the theoretical distribution of the true coordinates given the anonymised coordinates under the DHS anonymisation mechanism that includes both aggregation and random displacement. We jointly model both aggregation and random displacement mechanisms within a measurement error framework for kernel density estimation (KDE). Estimation is performed by using a stochastic expectation–maximisation (SEM) algorithm, cf. Hossain (2023).

The remainder of the paper is organised as follows. The DHS anonymisation process is presented in detail in Section 2. Section 3 presents the theoretical foundations of the paper which includes deriving the distribution under the DHS scheme and formularizing the measurement error model. Therefore, aggregation and random (donut) displacement are analysed separately and then the measurement error model is extended to include both anonymisation methods. In Section 4 the proposed methods are first evaluated in a simulation study and further compared to a naive approach that ignores the effects of anonymisation. In Section 5.1 the proposed methodology is illustrated using DHS data from the Rajshahi Division in Bangladesh to estimate the density of households below the poverty line. The impact of both random displacement and random displacement combined with aggregation is assessed. The results show that accounting for the measurement error in geolocations due to anonymisation leads to a more accurate picture for the spatial distribution of poverty in that region of Bangladesh.

2 Anonymisation strategies and the DHS scheme

Disclosing the geolocations of the units of analyses (individuals or households) carries the risk of identification. This can be mitigated by using geomasking methods such as aggregation to predefined spatial units and/or displacement of geographical coordinates. Demographic and Health Surveys (DHS) use a combination of aggregation and displacement as the preferred geomasking method. This is known as the DHS anonymisation scheme and comprises the sequential application of aggregation followed by displacement. In this section, we describe aggregation and random displacement before formally defining the DHS anonymisation scheme. Therefore, let $X_i = (X_{i1}, X_{i2})$ denote the true two-dimensional coordinates and $W_i = (W_{i1}, W_{i2})$ the observed coordinates after applying a geomasking method, where $W_i = X_i +$ error term.

2.1 Aggregation

A common strategy for anonymising geocoordinates is to aggregate these into spatial units.

In offical statistics, an administrative area system with in total A areas is used for aggregation. Here, geocoordinates within a given area are spatially aggregated. These administrative areas are usual of irregular size and complex structure. Formally, if the point to be anonymised X_i lies in the polygon P_a of area a, then it is placed at the centre of P_a, i.e., $W_i = M_a = X_i + \nu_i$. Hence, the set of centroids and the number of geocoordinates aggregated to it are observed. It should be noted that the anonymised location depends on both the true location and the area system. When displaying aggregated data, for example as a map or table, the type of anonymisation, i.e., the underlying area system, must be known.

The geometry of the aggregation areas can also be very simple. The simplest way of aggregation is to round the true geolocations into a rectangle, i.e., separately along each coordinate. Again, we denote $M_a,\ a = 1, \ldots,\ A$ as the centroids of the rectangle obtained when rounding along each axis, and r_1 the length of the rectangle along the horizontal axis and r_2 along the vertical axis. The anonymised point for X_i is $W_i = W_{i,a} = M_a$, if $X_i \in (M_{a1} - 0.5 \cdot r_1, M_{a1} + 0.5 \cdot r_1) \times (M_{a2} - 0.5 \cdot r_2, M_{a2} + 0.5 \cdot r_2)$ for $a = 1,\ \ldots,\ A$, i.e., the M_a corresponds to the middle of the rectangle. Hence, M_a can be interpreted as the centroid of the rectangle. The values of r_1 and r_2 are called rounding values and are assumed in this paper to be equal. Here, the geometry of the aggregation is a rectangle, and when defining as a polygon the vertices of the rectangle are used.

2.2 Random donut displacement

Another strategy for masking geolocated data consists of randomly shifting the geocoordinates. There are various ways for doing this. In this paper, we focus on random donut displacement, which was thoroughly presented by Hampton et al. (2010) in their study on improved privacy protection rules for mapping health data. Although the concept had already been described by Stinchcomb (2004b) who described it as random perturbation, the innovation introduced by Hampton et al. (2010) consisted of establishing a minimum level of displacement. Random donut displacement is applied in various contexts, including the protection of address-level data in public health studies in North Carolina, USA, cf. Allshouse et al. (2010), and in epidemiological research to anonymise patient data, cf. Fox et al. (2024).

Random donut displacement shifts the true coordinate in a random direction by at least a minimum distance, while ensuring that the displacement does not exceed a maximum distance. A random direction is first determined, using a random angle between 0 and 2π, i.e., $2\pi \cdot v_1$ with $v_1 \sim \text{Uniform}(0,1)$. Secondly, using a minimum and a maximum displacement radius, δ_{min} and δ_{max}, the random distance is generated, i.e., $\delta_{\text{min}} + v_2(\delta_{\text{max}} - \delta_{\text{min}})$ with $v_2 \sim \text{Uniform}(0,1)$. The additive offset of the point along the coordinates is determined by $e_{i1} = \sin(2\pi v_1) \cdot \big(\delta_{\text{min}} + v_2(\delta_{\text{max}} - \delta_{\text{min}})\big)$ and $e_{i2} = \cos(2\pi v_1) \cdot \big(\delta_{\text{min}} + v_2(\delta_{\text{max}} - \delta_{\text{min}})\big)$. The displaced location W_i is generated by adding the offset to the true coordinates, i.e., $W_i = X_i + e_i$. For further analysis of the joint distribution of e_{i1} and e_{i2}, we refer to Appendix A.

The use of methods that stochastically overlap the original coordinates, such as random donut displacement, carries the risk that the coordinates can be shifted outside the area of interest. Examples include uninhabited or uninhabitable areas such as forests, wetlands, or outside the country administrative unit. Therefore, let D be the donut area around some point $\widetilde{X} = (\widetilde{X}_1, \widetilde{X}_2)$ to be displaced, i.e., $D = \{P' \in R^2 \mid \delta_{min} < ||\widetilde{X} - P'||_2 < \delta_{max}\}$. The set Z defines areas, i.e., polygons, in which the point should not be displaced. Hence, the displaced point $\widetilde{W} \in D \setminus Z$.
From a computational point of view, the displacement of a geocoordinate accounting for areas that should not include any points can be achieved either by narrowing down the area in which the displaced point can be placed i.e., rejecting random angles and directions not fulfilling the requirements, or by drawing multiple points, restricting these to the feasible area. Nevertheless, the exclusion of uninhabited or uninhabitable areas is computationally expensive. In the application in this paper, urban and rural households are distinguished, which means that the parameters of the error term depend on the true location. For further details, see Hossain (2023) discussing random displacement.

2.3 DHS anonymisation scheme

The DHS anonymisation scheme is designed to ensure respondent confidentiality and comprises both aggregation and random displacement. The country of interest is divided into A disjunct Enumeration Areas (EAs). Each EA a contains n_a households which are placed at EA centroid. Hence, the true coordinates X_i of unit i are first aggregated to the midpoint $M_{a,i}$ of the corresponding EA. The aggregated coordinates are then further anonymised by using random displacement, i.e., shifting $M_{a,i}$ to $W_{a,i}$ for $a = 1, \ldots, A$. All geocoordinates aggregated to a centroid M_a are jointly displaced to W_a. In the DHS anonymisation scheme, coordinates are displaced by up to two kilometres in urban areas, while in rural areas, the maximum displacement radius is up to five kilometres, and for 1% of selected randomly EAs in rural areas, it is up to ten kilometres, cf. Burgert et al. (2013), Burgert-Brucker et al. (2016), and DHS Program (2025).

When original geocoordinates are *only* randomly displaced using the minimum and maximum distances specified by the DHS, the data is referred as *displaced* and the procedure as *DHS displacement scheme*. When, however,

original geocoordinates are first aggregated at the Enumeration Area (EA) level and subsequently displaced – as prescribed by the DHS protocol – the resulting data are considered *anonymised*, and the procedure is described as *DHS anonymisation scheme*.

Figure 1 shows an exemplary illustration of the DHS anonymisation scheme starting with true coordinates X, shown in green. These coordinates are then aggregated to their corresponding polygons, depicted in purple, i.e., X_i, X_j and X_k located in area a are aggregated to the corresponding polygon centroid $M_a = \cdots = M_{a,i} = M_{a,j} = M_{a,k}$. The centroid M_a is further displaced to W_a by drawing a random angle and distance, where the distance is restricted by δ_{min} and δ_{max}. The blue shaded donut area around the centroid M_a depicts the potential area the centroid can be displaced to. Then all geocoordinates aggregated to M_a are displaced to W_a, which is observable.

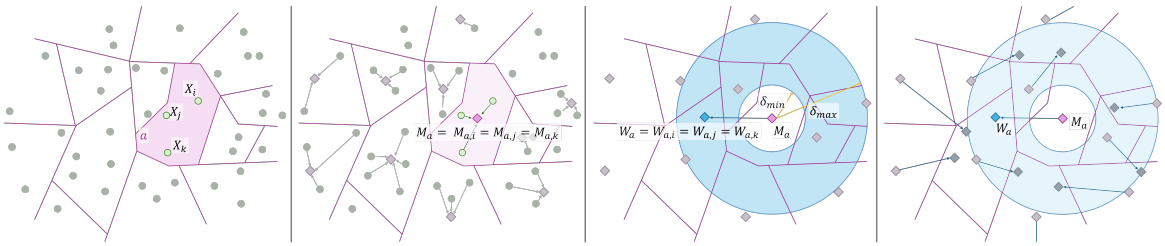

Figure 1: Geographical coordinates modified by the DHS anonymisation scheme

Although the DHS is highly transparent about the procedures used to anonymise the data, the Enumeration Areas (EAs) themselves remain unknown to data users. As a consequence, the underlying analysis requires the artificial construction of EAs. Therefore, we assume that aggregation by rounding or using complex geometries provide a reasonable approximation to the unknown aggregation geography. The use of different aggregation mechanisms allows us to assess the sensitivity of the results to this assumption. While no original EAs are provided by DHS, we retain the term EA for consistency in terminology. Depending on the construction method, these artificial EAs can take different forms.

In the simplest case, the data are rounded, resulting in EAs with rectangular shapes. In both the simulation study and the empirical application, this form of aggregation is referred to as *grid-based aggregation*. The rectangles, in which the geocoordinates are located and to whose centroids the points are perturbed, are referred to as *aggregation cells or cells*. This special case is considered in the following chapters due to the simplicity of its practical implementation.

In contrast to rounding, the generation of complex geometries is by no means easy. In general, the DHS tries to produce EAs with similar population sizes. Hence, we synthetically generate polygons in a systematic manner. Therefore, the polygons were generated using a *targeted cluster size*, which specifies the approximate number of geocoordinates per polygon. Initially, the geocoordinates are spatially clustered into groups of approximately equal size using the predefined targeted cluster size. For each cluster, a Voronoi tessellation is generated based on the mean coordinates in vertical and horizontal direction of all clusters. The resulting Voronoi tessellation is constrained by the convex hull of the overall point cloud. When constructing the Voronoi tessellation based on the mean coordinates, some points may fall within another polygons. Therefore, the number of points within each polygon varies slightly.
Note that the term *cluster* or related terminology are only used when describing the generation of the synthetic EAs. The process of generating these polygons is referred to as *polygon-based aggregation*. The EAs obtained by Voronoi tessellation are referred to as *polygons* or *aggregations area*, to ensure clear distinction to the grid-based case.

3 Kernel density estimation in the presence of geomasking

Our focus is on the smooth estimation of anonymised data. A common tool to obtain smooth spatial density estimates for geocoordinates is kernel density estimation (KDE). Therefore, we start by defining KDE based on exact geoocoodinates before analysing the problem when these coodinates are anonymised.

3.1 Kernel density estimation

Multivariate kernel density estimation (KDE) is a non-parametric approach to estimate the probability density of continuous bivariate variables. For the case presented in this paper, we consider $X = \{X_1, X_2, \ldots, X_n\}$ to be a sample of exact geocoodinates of size n from a bivariate random variable, with $X_i = (X_{i1}, X_{i2})$ denoting the longitude and latitude coordinates and the unknown density $f(\cdot)$. Let $\{x_g \mid g = 1, \ldots, G\}$ be a fine grid on which the density is evaluated. The kernel density estimator at a grid point x_g is given by

$$\hat{f}_H(x_g) = \frac{1}{n|H|} \sum_{i=1}^{n} K\Big(H^{-1}(x_g - X_i)\Big), \tag{1}$$

where $K(\cdot)$ is a multivariate kernel function, H denotes a symmetric positive definite bandwidth matrix and $|\cdot|$ denotes its determinant. In this setting, the kernel function $K(\cdot)$ is the multivariate Gaussian kernel, due to its smoothness behaviour. The choice of bandwidth H is very important for the performance of a KDE. Different approaches for its selection have been discussed in the literature; see Izenman (1991) and Silverman (1986). Here, the plug-in approach of Wand and Jones (1994) is used, which is computationally efficient as it relies on analytic approximations and avoids intensive resampling procedures. It directly targets the minimization of the Mean Integrated Squared Error (MISE), offering strong theoretical performance, see Wand and Jones (1994).

Equation (1) highlights that having knowledge of the true geolocations X is crucial for accurate density estimation. However, anonymisation through aggregation, displacement, or a combination of both can distort the true spatial patterns. In these cases, the true coordinates X are not observed, and only the anonymised coordinates W are available. If the true coordinates are distorted, the distortion is also transferred to the density estimate, see Groß et al. (2017). This motivates the development of methods that account for the measurement error introduced by anonymisation, assuming that the type of anonymisation applied is known. These methods aim to approximate the true density of the geocoordinates better than when using anonymised data alone, hence ignoring the process of anonymisation.

3.2 Geocoordinate anonymisation as a measurement error problem

Let $X_i,\ i = 1, \ldots, n$ be the true geocoordinates. However, the true coordinates are not observable but $W_i = X_i + e_i$, whereas e_i refers to the i.i.d. known error distribution F. Note that observed coordinates $W = (W_1, \ldots, W_n)$ depends on the anonymisation strategy. The coordinates W and the error distribution are given, i.e., knowledge about the anonymisation strategy must be given. We aim to estimate the density $f(X)$ from the sample $X_i,\ i = 1, \ldots, n$ using the observed values W. Under the assumption that the anonymisation strategy of the true coordinates is known, we can formulate a measurement error model $f(W|X)$ for W. From Bayes theorem it follows that

$$f(X|W) \propto f(W|X)f(X). \tag{2}$$

The first term is the conditional density $f(W|X)$, describing the anonymisation mechanism, that is, the probability distribution governing how the true coordinates X are transformed into the observed anonymised coordinates

W. The formulation of the anonymisation mechanism differs depending on the anonymisation applied on the true geocoordinates.

The second term in the Bayes theorem (2) describes the density of the true coordinates X. However, as X is unknown, consequently, $f(X)$ is unknown. This motivates the introduction of a simulation framework to estimate $f(X)$. Hence, $f(X)$ needs to be replaced by an estimate obtained by the multivariate kernel density estimation. For a KDE, samples approximating X are needed for an meaningful estimation of $f(X)$.

Therefore, pseudo-samples of X are drawn from $f(X|W)$ to allow the estimation $f(X)$. In detail, a pseudo-coordinate for X_i is drawn in a restricted area $\mathcal{G}^i$ around the observed point taking the anonymisation process into account. The definition of such an area ensures that the true coordinate is contained within it. In a numerical sense, let $\mathcal{G} = \{x_g \mid g = 1, \ \ldots, \ G\}$ denote a fine grid over the area of interest representing all possible pseudo-samples. For each observation X_i, this set is restricted to a subset $\mathcal{G}^i \subset \mathcal{G}$, which defines the admissible pseudo-coordinates for that observation. Note that the definition $\mathcal{G}^i$ depends on the anonymisation strategy applied on the true coordinates.

Hence, linking these two terms, the pseudo-samples are drawn in an iterative manner by taking the anonymisation mechanism and the density estimate from the previous iteration as sampling weights into account. Numerically, the anonymisation mechanism is described by the restricted area from which a pseudo-sample can be drawn and the conditional distribution, i.e., describing the probabilities within the area. These probabilities are further multiplied with the density estimate from the previous iteration to complete the sampling weights, i.e., fulfil Bayes theorem (2).

This leads us to come back to the first term $f(W|X)$ of our analysis and to formulate the conditional density $f(W|X)$ for the considered anonymisation mechanism. For a numerical implementation we need to define the restricted area $\mathcal{G}^i$ in each case.
To be able to formulate a measurement error model (MEM) for the DHS anonymisation scheme, involving both aggregation and displacement, we will proceed systematically. First we analyse both anonymisation strategies individually and then come to the subsequent use of them.

3.2.1 MEM in the presence of aggregation (MEM-Agg)

Starting with aggregation, it can be observed that the conditional density $f(W|X)$ collapses to the distribution of the area to which the true coordinate X is assigned. Under the assumption that the aggregation process of X_i is known, the measurement error model is $f(W|X) = \prod_{i=1}^{n} f(W_{a,i}|X_i)$ with

$$f(W_i = W_{i,a}|X_i) = \mathbb{1}_{\{X_i \in P_a\}} \ . \tag{3}$$

Equation (3) formalizes the aggregation mechanism: given the true coordinate X_i, the aggregated observation W_i refers to the centroid of the aggregation unit, in which the true coordinate X_i lied, and hence, is deterministically assigned to the polygon P_a containing X_i. Thus, the aggregation unit is known with certainty, while the exact location of X_i within P_a remains unidentified and is assumed to be uniformly distributed over the polygon.
Hence, the known anonymisation mechanism is reflected by known aggregation units. Here, we assume to have knowledge about the geometries of the aggregation areas $P_a, \ a = 1, \ \ldots, \ A$.

For numerical implementation, a pseudo-sample mimicing X_i is drawn from the aggregation area. Hence, the set of possible pseudo-samples $\mathcal{G}$ is restricted to

$$\mathcal{G}^i = \{x_g \in \mathcal{G} : x_g \in P_a\}, \tag{4}$$

the feasible set for X_i.

These pseudo-samples mimic X and enable an estimate of the density $f(X)$ by the multivariate kernel density estimation. In an iterative manner pseudo-samples are drawn according to the Bayes theorem, i.e., proportional to equation (3) and the estimate of $f(X)$. From a numerical perspective, enforcing the feasible set restriction $\mathcal{G}^i$ for X_i is sufficient to implement $f(W|X)$ in the Bayes Theorem, as the aggregation provides no additional information on the within-polygon location.

Note that on each point of the feasible set the KDE needs to be evaluated. In the implementation, the evaluation grid of the KDE (see equation 1) refers to the set of possible pseudo-samples to avoid double computation.

For further details we refer to Groß et al. (2017).

3.2.2 MEM in the presence of random donut displacement (MEM-RDD)

When considering random donut displacement, given the displacement radii δ_{min} and δ_{max}, the true location lies within the donut-shaped area around the displaced point. The conditional distribution within the area reads as

$$f(W_i|X_i) = \frac{1}{2\pi(\delta_{max} - \delta_{min})||W_i - X_i||_2} \; \mathbb{1}_{\{\delta_{min} < ||W_i - X_i||_2 < \delta_{max}\}}, \tag{5}$$

where $||W_i - X_i||_2$ represents the Euclidean distance between X_i and W_i and $\mathbb{1}$ the indicator function. Under the assumption that the random donut displacement process of X_i is known, the measurement error model is $f(W|X) = \prod_{i=1}^{n} f(W_i|X_i)$. Thus, the donut-shaped area around the displaced point contains with certainty the exact location of X_i, however, its location remains unknown. In contrast to aggregation its location is not assumed to be uniformly distributed over the area, but locations closed to the anonymised point have a higher probability, cf. Figure 12.
See Appendix A for details in deviation. To achieve equation (5) above, equation (12) in Appendix A needs to be adapted.

For numerical implementation, a pseudo-sample for X_i need to be drawn to update the estimate of $f(X)$ by the multivariate kernel density estimation. In detail, the true coordinates are repeatedly drawn from the donut-shaped area around the displaced point taking the displacement procedure and the current density estimate as sampling weights into account. Based on the displaced point W_i the set of potential sample coordinates is defined as

$$\mathcal{G}^i = \left\{x_g \in \mathcal{G} \mid \delta_{min} < ||W_i - x_g||_2 < \delta_{max}\right\}. \tag{6}$$

Hence, following the conditional distribution of X_i given W_i, for $x_g \in \mathcal{G}^i$ we obtain

$$f(X_i = x_g|W_i) \propto \frac{1}{2\pi(\delta_{max} - \delta_{min})||W_i - x_g||_2 \mathbb{1}_{\{x_g \in \mathcal{G}^i\}}} f(X_i = x_g) \; . \tag{7}$$

Note that for $x_g \notin \mathcal{G}^i$ it is zero.

3.2.3 MEM in the presence of aggregation and displacement (DHS anonymsiation scheme) (MEM-ARDD)

The DHS anonymisation scheme protects respondent confidentiality by first aggregating true household coordinates to the centroid of an Enumeration Area (EA) and then applying random displacement within a specified radius. In short, true geocoordinates X are first aggregated to centroids M and these centroids are further displaced; i.e., the observed coordinates are denoted by W. To be able to describe the anonymisation mechanism, we need to assume knowledge of all aggregation geometries P_a and their respective centroids M_a, $a = 1, \ldots, A$, as well

as the displacement radii δ_{min} and δ_{max}. Due to the consecutive application of aggregation and random donut displacement on the true geocoordinates X, the conditional density $f(W|X)$ needs to be extended. Given the anonymised and observable point $W_{a,i}$ and the centroid M_a, we obtain

$$f(W_{a,i}, M_a|X_i) \propto f(W_{a,i}|M_a, X_i)f(M_a|X_i) \propto f(W_{a,i}|M_a)f(M_a|X_i), \tag{8}$$

as $W_{a,i}$ is independent of X_i, the displacement is conditional on M_a.
The first term of equation (8) represents the displacement process, i.e., the centroid M_a is perturbed to $W_{a,i}$ using random donut displacement under δ_{min} and δ_{max}. Hence, $f(W_{a,i}|M_a)$ can be described by equation (5) (where X needs to be replaced by M). Furthermore, the second term of equation (8) represents the aggregation process, i.e., the true geocoordinates X_i are aggregated to a midpoint M_a. Hence, $f(M_a|X_i)$ can be described by equation (3) (where W is replaced by the M). However, the centroid M_a corresponding to the anonymised and observable point $W_{a,i}$ is unknown. For meaningful anonymisation, multiple centroids lie within the donut-shaped area around the observed point. The observed point can usually not be associated with one centroid and hence one aggregation area. Therefore, will be denoted as W_i.
Building on the derivation (8), it would be intuitive to build a two-step approach, i.e., first choosing a new centroid under the consideration of random displacement mechanism and then drawing pseudo-samples within this new area. Following Bayes therorem, in a two-step approach the measurement error models for random (donut) displacement and for aggregation need to be nested. However, this two-step approach can be further enhanced. Especially, to improve computational efficiency, a one-step approach is proposed here.

In the one-step approach we drop the condition that all anonymised coordinates must origin from the same polygon. It omits the step of choosing the most likely centroid but rather draws pseudo-sample from a feasible set. Therefore, centroids are determined within the donut area and the pseudo-samples are drawn from the grid points within these polygons. This emphasises that sampling takes place directly at the coordinate level and that the intermediate step of drawing a centroid and thus a spatial unit is omitted. Unlike the two-step approach, aggregated points are not restricted to a single selected polygon but may be spread across multiple polygons, resulting in a smoother estimate. In contrast to the MEM-RDD, where pseudo-samples were drawn within the donut area around the displaced point, we are now drawing from the polygons whose centroids lie within the donut area of the anonymised point (= original geocoordinate which was aggregated and displaced). Hence, the feasible set may be located outside the (donut) displacement area around the observed point W or part of the displacement area may be excluded. This is particularly beneficial when working with large aggregation areas. Nevertheless, this comes at the cost of losing polygon-specific characteristics.

For numerical implementation, a pseudo-sample for X_i need to be drawn to update the estimate of $f(X)$ by the multivariate kernel density estimation. Based on the anonymised point W_i the set of potential sample coordinates is a subset of the fine grid on the area of interest $\mathcal{G}$. The feasible set is defined as

$$\mathcal{G}^i = \bigcup_{a:\ \delta_{\min} < \|W_i - M_a\|_2 < \delta_{\max}} \{x_g \in \mathcal{G} : x_g \in P_a\}. \tag{9}$$

(Following the example in Figure 1 illustrating the consecutive anonymisation by aggregation and then displacement, in the appendix the feasible set using the one-step approach is depicted in Figure 15.)
For $x_g \in \mathcal{G}^i$ from equation (9), the conditional distribution of X_i given W_i and M_a combines equation (7) describing

the displacement process of the centroids with equation (3) describing the aggregation process, and reads as

$$\begin{aligned} f(X_i = x_g | W_i, M_a) &\propto f(W_i|M_a) f(M_a|X_i = x_g) f(X_i = x_g) \\ &\propto \frac{1}{2\pi(\delta_{max} - \delta_{min})||W_i - M_a||_2 \mathbb{1}_{\{x_g \in P_a\}}} \mathbb{1}_{\{x_g \in \mathcal{G}^i\}} f(X_i = x_g). \end{aligned} \tag{10}$$

Additionally, $f(X_i = x_g)$ is estimated based on the previous estimation. However, it is important to note that the assumption of knowing the shapes of the aggregation areas when observing DHS anonymised coordinates is rather unusual.

In applications there could occur a lack of information with respect to the exact aggregation geometries, e.g., the DHS do not publish the EAs. To apply the algorithm considering aggregation and random donut displacement as measurement errors, these aggregation areas should be estimated. We propose to estimate them using a Voronoi tessellation on the anonymised data, i.e., on $W_{i,a}$ representing the aggregated and displaced data. A Voronoi tessellation partitions an area into cells around given points (i.e. the mean x- and y-coordinates), where each cell contains all points closer to its corresponding point than to any other, see Okabe and Boots (2000). The resulting polygons $\hat{P}_a$ as well as its centroids $\hat{M}_a$, $a = 1, \ldots, A$ can be used in MEM-ARDD instead of P_a and M_a, respectively.
Additionally, it may be of interest whether the estimation is needed or whether using the information of the displacement can reduce the information loss occurred by the anonymisation. When comparing results using MEM-ARDD having or not having information about the aggregation geometries, the estimation f_{ARDD} is equipped with an *known* or *unknown* in the subscript, respectively.

3.2.4 Pseudoalgorithm

From the observed coordinates $W = (W_1, \ldots, W_n)$ and the knowledge about the anonymisation mechanism, we aim to estimate the density $f(X)$ of the true locations. Using Bayes theorem $f(X|W) \propto f(W|X)f(X)$, we obtain a conceptual idea.

The first term $f(W|X)$ represents the anonymisation mechanism. Depening on the anonymisation applied on the true coordinate X, this term is derived. For aggregation, we assume knowlegment of the geometries P_a and their centroids M_a; for random donut displacement, we assume knowlegment of the displacement radii δ_{min} and δ_{max}; and for both – aggregation and displacement, we assume knowlegment of both the geometries P_a and their centroids M_a as well as the displacement radii δ_{min} and δ_{max}.

The second term in the Bayes theorem (2) describes the density of the unknown true coordinates X. A simulation framework to estimate $f(X)$ based on pseudo-samples is needed. Therefore, pseudo-samples of X are drawn from $f(X|W)$ to allow the estimation $f(X)$. In detail, a pseudo-coordinate for X_i are iteratively drawn respecting the Bayes Theorem.

The following pseudo-algorithm should give an overview.

Algorithm 1 Pseudo-code: MEM in the presence of aggregation, random donut displacement or the DHS anonymisation scheme

1: Given: $W_i,\ i = 1,\ \ldots,\ n$, knowledge about the anonymisation process, burn-in B and sampling phase S.
2: Set an evaluation grid $\mathcal{G}$ and compute the set of potential pseudo-samples $\mathcal{G}^i,\ i = 1,\ \ldots,\ n$.
3: Calculate naive kernel density estimation f_0 based on displaced data W_i with large bandwidth.
4: **for** t in 1:(B + S) **do**
5: Draw a pseudo-sample for X_i from the evaluation grid $\mathcal{G}^i$, which is proportional to f_{t-1} and under consideration of the anonymisation process $f(W|X)$, i.e., referring to eq. (3) for aggregation, eq. (7) for random donut displacement, and eq. (10) for the DHS anonymisation scheme.
6: Calculate a kernel density estimation f_t from pseudo-samples with plug-in bandwidth by Wand and Jones (1994).
7: **end for**
8: Estimated density $f_{RDD} = \frac{1}{\mathrm{S}} \sum_{t=1}^{\mathrm{S}} f_{t+B}$.

Further notes

The idea of the simulation concept originates from Groß et al. (2017), who introduced the approach for aggregated georeferenced data. For a detailed description of the algorithm, we refer to Gril et al. (2025), as the same notation is used in this work. Furthermore, Gril et al. (2025) demonstrated in their simulations on a measurement error model under aggregation clear advantages of using the plug-in bandwidth; hence, the same bandwidth selection method is applied here.

The consideration of random donut displacement within a measurement error framework, and subsequently in the context of the DHS anonymisation scheme, constitutes the novel contribution of this paper.

The deviation of the conditional distribution in the case of random donut displacement can be found in Appendix A. Furthermore, we refer to Hossain (2023) for the development in the case of random displacement (corresponding to T for the original coordinate and W for displaced ones).

When considering the MEM for consecutive aggregation and random donut displacement, it would be intuitive to build a two-step approach. In detail, a centroid M_a lying within the donut-shaped area around the displaced point is first selected using equation (5). Since each aggregation point that contains several points has been moved together according to the displacement rule, a new joint centroid is determined for all points. The aggregated points are then redistributed within the aggregation unit corresponding to the chosen centroid. Redistribution is performed proportional to the previously estimated density of X retaining specific EA characteristics.
In Appendix B.2, especially in Figure 13 - 15 the path to developing the one-step approach proposed is explained in detail. For a meaningful comparison of the approaches, we assume knowlegdment of the aggregation geometries and the displacement radii.

In fact, in Chapter 3 of Hossain (2023) proposed a two-step approach under the knowledge of the displacement radii and the rounding value in grid-based aggregation, i.e., first choosing a new centroid under the consideration of random displacement and then drawing pseudo-samples within an area around the chosen centroid. However, Hossain relaxes the assumption of knowing the exact location of the centroids.
In detail, centroids M_a are first sampled within the donut area induced by displaced points using equation (5). Since each aggregation point that contains several points has been moved together according to the displacement rule, a new joint centroid is determined for all points. The aggregated points are then redistributed within the grid cell proportional to the previously estimated density of X retaining specific EA characteristics. These grid cells are generated by using the rounding value.

For the MEM-ARDD having aggregation geometries is essential. However, the assumption of knowing the shapes of the aggregation areas when observing DHS anonymised coordinates is rather unusual. Note that in grid-based aggregation we could have some more information available, like the rounding value but not the centroids, and hence, it is not possible to reconstruct the polygons exactly. Having this additional information is not considered in this analysis, but only if complete information or no information about the aggregation process is available. Having knowledge about the rounding value is discussed in Chapter 3 of Hossain (2023).

4 Simulation study

In this section, we evaluate the performance of the proposed methodology. Using different scenarios for the intensity of the measurement error process, for both random donut displacement as well as aggregation and displacement, we assess if accounting for the error introduced by the geomasking process improves the accuracy of density estimates compared to standard kernel density estimation which ignores the measurement error. The simulated data are generated by using a mixture of three uncorrelated bivariate normal distributions as follows,

$$f(x) = \frac{1}{3}\phi(x|\mu_1, \Sigma_1) + \frac{1}{3}\phi(x|\mu_2, \Sigma_2) + \frac{1}{3}\phi(x|\mu_3, \Sigma_3) \quad \text{with} \tag{11}$$

$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \mu_2 = \begin{pmatrix} 5 \\ 3 \end{pmatrix},\ \mu_3 = \begin{pmatrix} -4 \\ 1 \end{pmatrix},\ \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

The simulation settings follow the design in Groß et al. (2017). Using a mixture of three Gaussian distributions with the proposed parameters leads to three agglomeration areas that are discernible, see Figure 17 Appendix B.2. However, displacement of sampling points from this density may cause the areas to overlap or blur, depending on the chosen parameters.

For the evaluation of the measurement error model when having displaced or anonymised (= aggregated and displaced) data, the data are generated using equation (11) for 500 independent repetitions. Each dataset S_0 is based on sample size of $n = 500$ random samples. The density of the sample data f_{S_0} is evaluated on a regular grid for both, the x- and y-coordinates, ranging from -10 to 10. Using a grid-width δ_x and grid-height δ_y of 0.1, evaluation points $x_g,\ g = 1, \ldots,\ G$ are constructed. Hence, the MEM-RDD and MEM-ARDD are evaluated on this evaluation grid.

The coordinates of S_0 are perturbed using either random donut displacement (RDD) or a combination of aggregation and random donut displacement (ARDD), obtaining a perturbed set of coordinates S_D. Based on S_D the naive kernel density estimation f_{S_D} is computed. The density estimator from the MEM-RDD is denoted by f_{RDD} and from the MEM-ARDD is denoted by f_{ARRD}, see Section 3.2.2 and 3.2.3 for details. Both the burn-in and sampling phases were set to 20 iterations each. For a further discussion of the burn-in and sampling phases we refer to Appendix B.

Two error measures are used to evaluate the performance of the models. First, the Root Mean Square Integrated Error (*RMISE*) is used to evaluate the global error and second, the *Jaccard Similarity* is used to evaluate the efficiency with which local hotspots are retrieved. The RMISE between f and $\hat{f}$, approximated by using the Riemann sum over the fine evaluation grid, is defined as follows,

$$\text{RMISE}(f, \hat{f}) = \sqrt{\mathbb{E}\left(\int \left[f(x) - \hat{f}(x) \right]^2 dx \right)} \approx \sqrt{\frac{1}{G} \sum_{g=1}^{G} [f(x_g) - \hat{f}(x_g)]^2 \delta_x \delta_y}\ .$$

Using the evaluation grid, denote by q% the region containing the top q% of the probability mass, by $f_{|q}$ the q% hotspot of f, and by $\hat{f}_{|q}$ the q% hotspot of $\hat{f}$. The Jaccard Similarity index is then define as follows,

$$\text{Jaccard Similarity}(f, \hat{f}; q) = \frac{|f_{|q} \cap \hat{f}_{|q}|}{|f_{|q} \cup \hat{f}_{|q}|} .$$

In the event of complete overlap between the hotspots of the estimated and true densities, the index is equal to 1. Conversely, in the absence of any overlap, the index is equal to 0.

4.1 Evaluation of density estimators under random donut displacement

We first consider the impact of random donut displacement on the original sample S_0 and the use of MEM-RDD, see Section 3.2.2. Each point is displaced uniformly in a random direction in the interval $[0, 2\pi)$, and a distance bounded between $\delta_{\min}$ and $\delta_{\max}$ from its original position. To ensure broad applicability, four different settings are developed and visualised in Figure 2. For each setting, a simulated set of original points S_0 (dark gray) and displaced points S_D (dark cyan) are presented. Additionally, two original points (black), the displacement radii and the corresponding displaced points (cyan) are highlighted. Note that overlaps between radii are possible. Motivated by the DHS displacement scheme, the minimum distance of the first two settings is set to zero. In Setting A, the maximum distance is set to ($\delta_{max} = 0.5$), while in Setting B it was set to ($\delta_{max} = 2$). In Settings C and D, the minimum distance is set to a non-zero value, which causes the area to which the original point is relocated to have the appearance of a donut. In Setting C, the displacement area is donut-shaped and wider, resulting in a large displacement area ($\delta_{min} = 0.5$ and $\delta_{max} = 1.5$). In Setting D, both distances are set to high values, the displacement area in the shape of a donut looks rather narrow ($\delta_{min} = 1.5$ and $\delta_{max} = 2$).

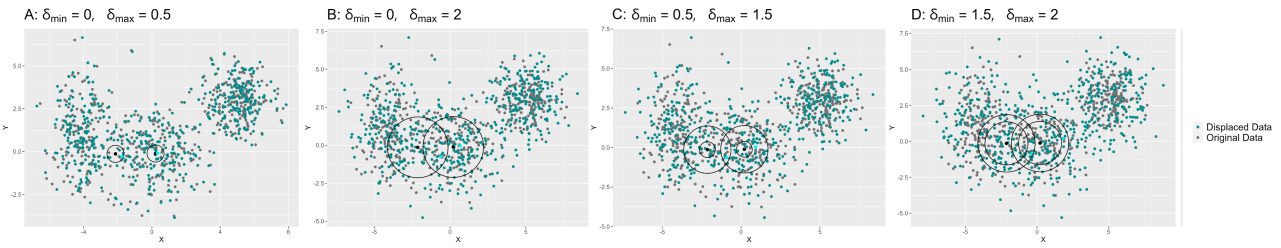

Figure 2: For each individual setting, a set of original points S_0 is drawn (dark gray) and displaced points S_D (dark cyan) are obtained by displacing each point in S_0 with corresponding minimum and maximum displacement distances. Each subplot highlights two original points (black), their displaced radii, and the displaced point (cyan).

In Figure 3 we present boxplots of the RMISE (first row) and the Jaccard similarity (second row) for different estimators under the four simulation different settings. The boxplots of the RMISE illustrate the information loss as a result of displacement. The RMISE of the naive approach f_{S_D} (shown in gray) is larger than that of the proposed density estimator f_{RDD} that accounts for random displacement as measurement error, and the difference increases with increasing displacement radii.

The Jaccard Similarity is used to evaluate the overlap between the hotspots identified using the density estimators (naive in gray colour and adjusted for displacement in green colour) and the true hotspots. Here we use the top 70%, 80%, and 90% intensity regions corresponding to the three boxplots under the four simulation settings. In Setting A, the displacement parameters are relatively small, resulting in minimal impact of the density estimation f_{S_D} based on displaced coordinates. Although the loss of information using the RMISE is clearly depicted as the displacement radii increases, differences in the Jaccard Similarity are less pronounced and more visible under Settings C and D. Overall, the results show that using the MEM-RDD reduces the RMISE and improves the ability to identify hotspots.

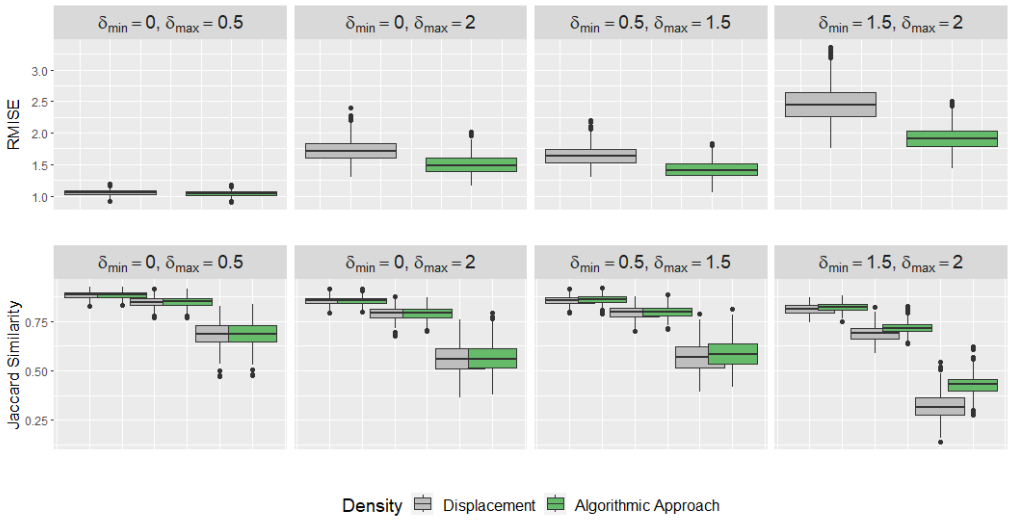

Figure 3: Performance measures of the proposed approach using the RMISE and the Jaccard Similarity by comparing the naive estimation f_{S_D} and the estimation obtained by the algorithm f_{RDD} to the data generation process for four fixed settings. For the Jaccard Similarity top 70% (left), 80% (middle), and 90% (right) hotspot regions are shown.

4.2 Evaluation of density estimators under aggregation and random donut displacement

In this section we assess the impact of the sequential application of the two anonymisation strategies, namely aggregation and displacement as is the case with DHS anonymisation scheme. Applying the MEM-ARDD means that both the geographies from aggregation process and it centroids, as well as the displacement radii are known.

As described in Section 2.3 to create aggregation geometries grid-based and polygon-based aggregation is used. These two aggregation strategies offer comprehensive insights into the aggregation process and allow us to assess the sensitivity of the results to this assumption. Nevertheless, several factors must be considered when selecting reasonable settings. First, the donut or disc-shaped displacement area around the anonymised point W_a should ideally contain several centroids M_a, $a = 1, \ldots, A$, as otherwise the displacement becomes meaningless. Note that the true aggregation centroid M_a corresponding to the anonymised centroid W_a lies by construction within this area. If no other centroids lie within the displacement area around the anonymised point W_a, the MEM-ARDD (see Section 3.2.3) reduces to redistributing the points within the true area.

Table 1 summarises the settings for grid-based and polygon-based aggregation and the corresponding displacement parameters applied to the aggregation centroids. For grid-based aggregation, each cell is defined as a rectangular area. The midpoint is therefore the point that is equally distant from all four vertices. For polygon-based aggregation, the midpoint depends on the shape of the polygon. The approach used to construct the polygons is described in detail in Section 2.3.

	Grid-based aggregation				**Polygon-based aggregation**			
Setting	**A1**	**A2**	**A3**	**A4**	**B1**	**B2**	**B3**	**B4**
δ_{min}	0	0	0.5	1.5	0	0	0.5	1.5
δ_{max}	0.5	2	1.5	2	0.5	2	1.5	2
Rounding value r	0.2	0.75	0.5	1	–	–	–	–
Number of polygons cl	–	–	–	–	410	140	250	100

Table 1: Anonymisation by aggregation and random donut displacement leads to different settings for two aggregation strategies: rounding on a regular grid and polygon-based aggregation.

4.2.1 Grid-based aggregation

In grid-based aggregation, the coordinates are rounded in x- and y-direction based on a specific rounding value r that defines the height and width of the aggregation cells, see Section 2.1 for details. The aggregation grid is regular, but the presence of agglomeration areas – see the mixture of three normal distributions as data generation process (11) – leads to an uneven distribution of sample points across cells, i.e., the resulting number of points aggregated to an aggregation cell vary in size. Moreover, since the scenario was repeated 500 times, i.e., with 500 independently drawn datasets, the number of aggregation cells varies across replications for fixed parameters. Table 2 summarises the number of aggregation cells as well as its sizes under different settings. For very small rounding values, as in Setting A1, many aggregation cells contain only a single original sample point, leading to a large number of cells. In contrast, larger rounding values, as seen in Settings A2 or A4, result in fewer aggregation cells, each of which consolidates multiple original data points, thereby forming fewer but larger cells.

	Number of aggregation cells						**Number of sample points per cell**					
Setting	**Min**	**1Q**	**Median**	**Mean**	**3Q**	**Max**	**Min**	**1Q**	**Median**	**Mean**	**3Q**	**Max**
A1	393	414	420	419.2	424	442	1	1	1	1.19	1	6
A2	123	136	140	139.7	143	154	1	1	3	3.57	5	26
A3	207	224	229	228.7	234	246	1	1	2	2.18	3	16
A4	78	91	93	93.52	96	108	1	2	4	5.34	8	40

Table 2: For the settings A1 - A4 reflecting grid-based aggregation the table shows the summary statistics for the number of aggregation cells and the number of sample points per cell.

For each individual setting, the anonymisation process of one set of original coordinates S_0 using the parameters for grid-based aggregation and random donut displacement from Table 1 (see Settings A1–A4) is demonstrated in Figure 4. Therefore, for each setting a set S_0 is generated. For visibility reasons, the points of S_0 are not shown directly. According to the rounding values, these original points are aggregated to the centroids M_a of the rectangles, which are shown as gray points. The size of the points shows how many points of S_0 have been aggregated to the corresponding centroid. These centroids are further displaced using RDD. Note that all points aggregated to a centroid are displaced jointly. The resulting anonymised points are shown in dark cyan and its size indicate again the number of original points anonymised to the receptive point. Furthermore, for two selected aggregation points per setting, their displacement radii are visualized as circles, and the corresponding randomly selected displacement point is shown in lighter cyan.
In Figure 4, it is evident from the size of the points that smaller rounding values, i.e. smaller rectangles, result in fewer points aggregated to their centroids (cf. Setting A1 and A3). When larger rectangles are used for aggregation, the rounding values must also be large to achieve meaningful anonymization. However, in this case, the underlying structure of the data may be completely lost (cf. Setting A4).

For the MEM-ARDD under the assumption of known aggregations geometries, the set of grid points from which pseudo-samples can be drawn is determined by the grid points that lie within the aggregation cell whose centroids lie within the displacement area of the anonymised point.
Hence, to ensure meaningful anonymisation for grid-based aggregation, the maximum displacement radius must exceed the rounding value and the displacement area must be large enough, i.e., the minimum distance should not be too large. Otherwise, the original aggregation point would be the only valid centroid within the defined radius, reducing the process to disaggregation.
Consequently, larger rounding values require proportionally larger maximum displacement radii. However, if both the rounding and displacement parameters are too large, the spatial structure of the data can become irreversibly distorted. For small rounding values, the effect of random displacement on kernel density estimation is negligible, see Groß et al. (2017). Therefore, in such cases, considering only the displacement may be sufficient.

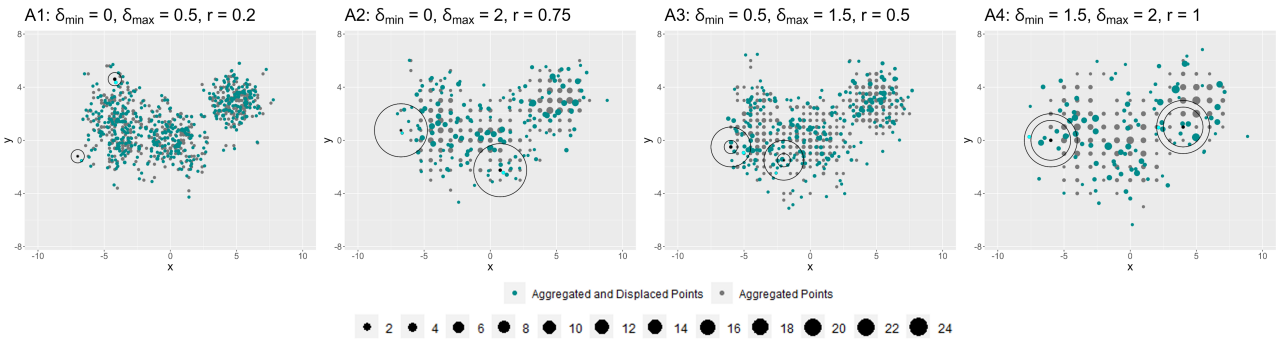

Figure 4: Set of original coordinates S_0 were first aggregated using varying rounding values. These aggregation points (gray) were further displaced (darkcyan). The parameters for the settings are stated in Table 1. For each setting, two aggregation points (black) are shown, along with their displacement radii and the corresponding randomly selected displacement point (cyan).

Figure 5 shows the relative RMISE and the Jaccard Similarity for the 80%, 85%, 90%, and 95% hotspots. These thresholds specify the hotspot area, i.e., only the upper 20%, 15%, 10% and 5% of the densities values are considered, with the given values corresponding to the respective quantiles. The performance of the proposed density estimator that accounts only for random displacement (MEM-RDD; Section 3.2.2) is presented in light turquoise. The estimator that accounts both for aggregation and displacement (MEM-ARDD; Section 3.2.3) is presented in dark turquoise. The performance of the density estimators that account for anonymisation improves as the effects of anonymisation increase. The RMISE of the proposed estimators reduces compared to the RMISE of the naive estimator and the Jaccard Similarity increases. Accounting for aggregation and displacement, instead of only displacement, appears to matter only in the last scenario that shows some improvement of the corresponding estimator over the estimator that accounts only for displacement.

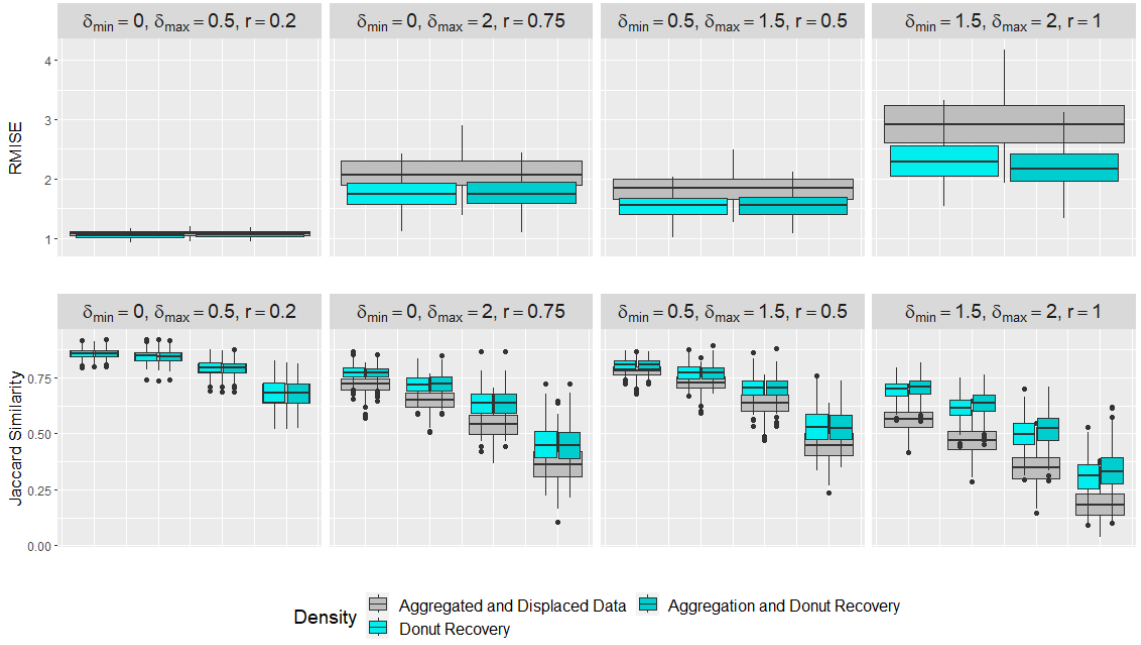

Figure 5: Performance measure of the proposed approaches using the RMISE and the Jaccard Similarity by comparing the naive estimation f_{S_D} and the estimation obtained by the algorithm f_{RDD} and f_{ARDD} to the data generation process for four fixed settings for grid-based aggregation.

4.2.2 Polygon-based aggregation

In polygon-based aggregation, the sample coordinates are aggregated to predefined polygons. In this simulation study, the polygons were synthetically generated in a systematic manner. The procedure used for polygon-based aggregation (see Section 2.3) results in rather equal populated aggregation areas, displayed in the summary statistics in Table 3. Compared to the grid-based approach (see Table 2), the aggregation areas (polygons) contain a more balanced number of data points.

Figure 6 displays an exemplary scenario for Settings B1 to B4, with gray points indicating the centroids of the polygons used for aggregating the sample points. As before, the original sample points themselves are not displayed directly. Due to the polygon generation procedure, the resulting polygons vary in size but contain a similar number of sample points.

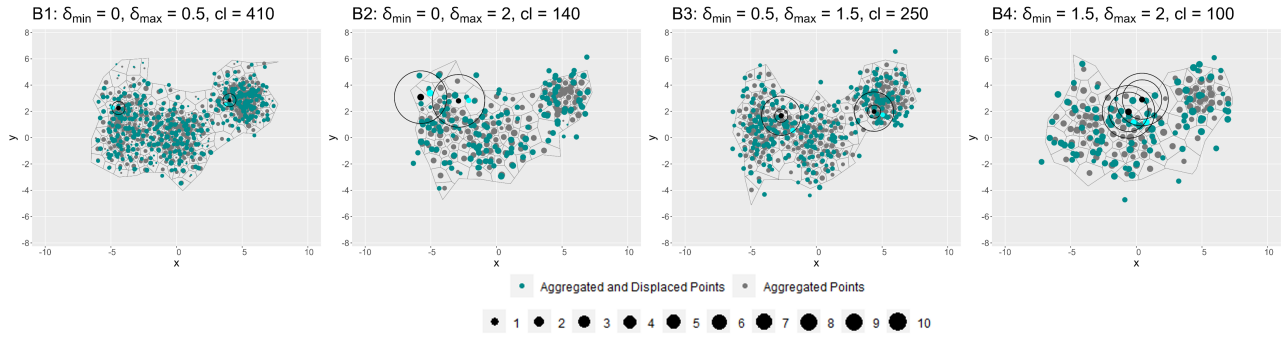

Figure 6: Sampled data were first aggregated based on polygons. The centroids of the polygons (gray) were then displaced using the random donut displacement (darkcyan), with parameters (e.g., minimum and maximum radius) set according to predefined tuning settings. For each setting, two aggregation points (black) were shown, along with their displacement radii and the corresponding randomly selected displacement point (cyan).

The number of polygons was fixed and Table 3 provides an overview of how many sample points were aggregated per polygon in all settings. The choice of displacement radii is crucial in this context. Compared to grid-based aggregation, it is more difficult to define a clear maximum displacement radius that guarantees anonymity. In particular, the varying polygon sizes introduce challenges: in regions with small polygons (i.e., small clusters), the displacement radii may be sufficient to ensure anonymisation, whereas in regions with larger polygons, the displacement distances might be too small to prevent reidentification. This issue is apparent in Settings B1 and B4, where large polygons near the edges are particularly prone to disclosure and may not be displaced sufficiently by a too small maximum radius (B1) or a too small displacement area (B4).

	Number of polygons	**Number of sample points per polygon**					
Setting		**Min**	**1Q**	**Median**	**Mean**	**3Q**	**Max**
B1	410	1	1	1	1.21	2	3
B2	140	1	3	4	3.571	4	8
B3	250	1	2	2	2.041	2	5
B4	100	2	4	5	5	6	9

Table 3: For the settings B1 - B4 reflecting polygon-based aggregation the table shows the number of polygons on the one hand and summary statistics for the number of data points per polygon on the other.

The performance when aggregating with irregularly sized and shaped polygons is similar to that using a regular aggregation grid and is shown in Figure 7. Accounting for the effect of anonymisation, reduces the RMISE and increases the Jaccard Similarity in Settings B2 to B4.

Aggregation process unknown

To apply the MEM-ARDD, we assume knowlegdment about the anonymisation process, i.e., both the displacement radii and the aggregation geometries. Typically, the aggregation of spatial data is readily apparent, as it is often visualized through choropleth maps. Such representations make it relatively easy to discern the aggregation geometries, i.e., aggregation cells or polygons. However, in our case, the aggregation centroids are further displaced, making it no longer straightforward to infer the underlying aggregation geometries. We propose to approximate the unknown geometries by using Voronoi tessellation on the displaced and aggregated point coordinates as centroids, and the convex hull of the point cloud as a boundary. Hence, the true geometries P_a are estimated and denoted by $\hat{P}_a$. For MEM-ARDD the assumption that the displacement radii are known is crucial and not relaxed. For a further distinction of the results of the MEM-ARDD, the resulting density under the knowledge of P_a is denoted $f_{ARDD,\text{known}}$, while when using $\hat{P}_a$ it is denoted $f_{ARDD,\text{unknown}}$.

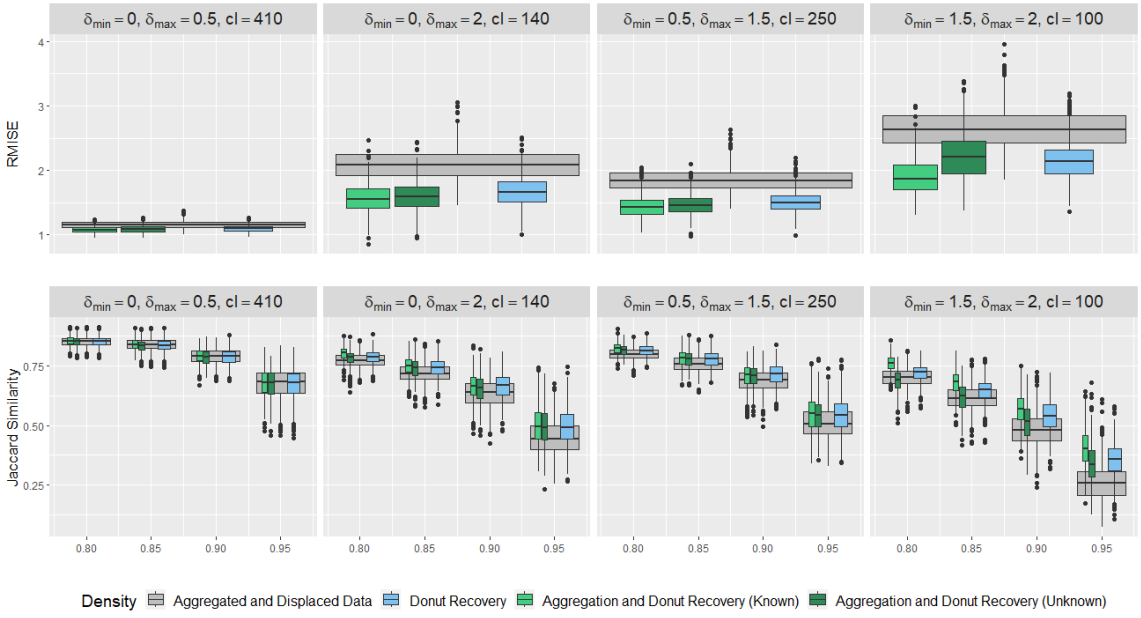

Figure 7: Performance measure of the MEM-RDD and MEM-ARDD using the RMISE and the Jaccard Similarity for four simulation settings and polygon-based aggregation. The naive estimator f_{S_D} (gray) is compared to the estimator of the MEM-RDD f_{RDD} (blue) and the estimators of the MEM-ARDD, applied using known geometries $f_{ARDD,\text{known}}$ (light green) and approximated geometries $f_{ARDD,\text{unknown}}$ (dark green). In all cases, the displacement radii are given.

As expected, the absence of knowledge about the geography used in the anonymisation process leads to a moderate decline in performance. Depending on the setting, accounting for aggregation, both with and without knowledge of the aggregation structure, can improve the performance compared to approaches that account only for random displacement.

In Setting B2 (Figure 7 of column 2), the performance of MEM-ARDD with both, having or not having knowledge of the aggregation, is better than using MEM-RDD. In contrast, for Setting B3 (panels of column 3), the performance of the estimator f_{RDD} and $f_{ARDD,\text{ unknown}}$, which is based on approximated polygons, is almost identical. Interestingly, in Setting B4 (panels of column 4), relying solely on MEM-RDD yields better results than incorporating the approximated aggregation structure. This is not consistent with the expectation that the consideration of aggregation is particularly effective when the aggregation areas are large. However, this could be caused by the naive estimation of the polygons which do not represent the true polygon P_a. Using the MEM-ARDD, the area, in which the potential pseudo-samples can be drawn, is restricted to the aggregation areas, which centroid lie within

the displacement area around the anonymised points. When approximating P_a, this area may be strongly deviating yielding poorer outcomes than when applying MEM-RDD.

In summary, the findings of the simulation study suggest that in the absence of knowledge about the aggregation process, accounting only for displacement may be a reasonable estimation approach.

5 Application

In this section, we evaluate the proposed methodology for identifying poverty hotspots using DHS data from the Rajshahi division (province) in Bangladesh. The Rajshahi division is located in northwestern Bangladesh, spans approximately 18,000 km^2, and is home to over 20 million people. Administratively, the division comprises 8 districts and 70 upazilas (Admin 3 units), making it a suitable and diverse setting for assessing the impact of spatial anonymisation.

Because the true household coordinates are not available in the DHS data, we generate synthetic household locations. The process of generating household locations is carried out independently of the DHS anonymisation process and is used solely to create a realistic underlying spatial distribution of households for evaluation purposes. For each upazila (Admin 3 unit), the total number of households is first obtained from the most recent population census. Household locations are then generated by sampling grid points from a 100 m resolution gridded nighttime light (NTL) intensity surface, with sampling probabilities proportional to the observed NTL intensity within each upazila. This ensures that census-derived household totals are preserved exactly at the upazila level, while allowing for spatial heterogeneity in household density that is consistent with observed patterns of human settlement.
Following the generation of household coordinates, households are classified as poor or non-poor using upazila-level direct estimates of poverty headcount ratios (HCR). The poverty status is assigned by drawing independent Bernoulli samples with probabilities equal to the corresponding upazila poverty estimate. Households are additionally labelled as rural or urban to enable the application of DHS-specific displacement radii in subsequent anonymisation steps. Assuming an average household size of approximately four individuals, this procedure results in a total of 4,624,884 synthetic household locations across the Rajshahi division. These locations are treated as the true (unobserved) coordinates in the subsequent evaluation of aggregation- and displacement-based anonymisation strategies.
As a remainder, household coordinates in the DHS are anonymised through a combination of aggregation and random displacement. First, household coordinates are aggregated to Enumeration Areas (EAs). Then, EA centroids are randomly displaced, i.e., up to 2 km in urban areas and up to 5 km in rural areas, with 1% of rural points shifted up to 10 km, to protect confidentiality, cf. Burgert et al. (2013) and DHS Program (2025). These procedures limit reidentification risks, cf. Burgert-Brucker et al. (2016). The proposed methodology in this paper is tested on the basis of the generated household dataset comparing the performance of different density estimators of poverty. Initially, displacement is applied to the household coordinates without aggregation. Then, both anonymisation strategies are applied. To combine aggregation and random donut displacement, information about the aggregation process is needed i.e., the structure of EAs in the case of DHS in Bangladesh. Since this information is not publicly available, we experimented with both grid-based and polygon-based aggregation methods of varying sizes. As in the simulation study, care was taken to ensure that the scenarios considered are reasonably comparable. Since the DHS anonymisation process assumes that there is no displacement outside a district, the procedure is applied separately for all eight districts in the Rajshahi division.

5.1 Random displacement

In this section, the true geocoordinates are subject to the DHS random displacement scheme without aggregation. As the DHS is transparent with the displacement radii applied on the data, the MEM-RDD can be used. Kernel

density estimates of (a) the original coordinates, (b) the displaced coordinates, and (c) the density estimated by using MEM-RDD are plotted in Figure 8. Note that the evaluation is based on a 250×250 metre grid. The visualisation reveals that while the anonymisation effect is noticeable in terms of density, it is relatively modest. The global error measure RMISE shows an improvement when comparing the KDE based on the original coordinates with the one based on the displaced coordinates ($4.913148 \cdot 10^{-4}$) as well as the one using MEM-RDD ($4.818337 \cdot 10^{-4}$).

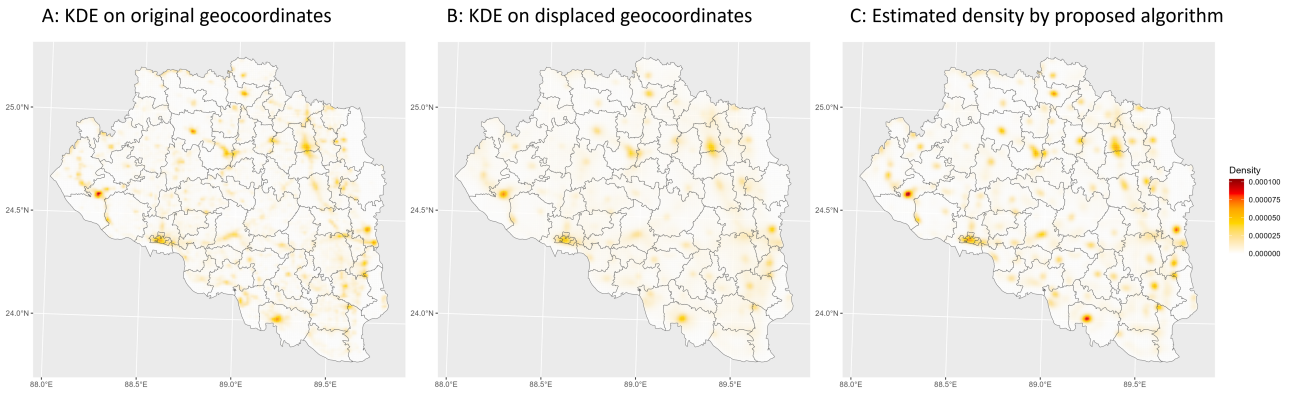

Figure 8: Kernel density estimates based on (A) the original coordinates, (B) the displaced coordinates, and (C) the algorithmically recovered density illustrate the effects of displacement.

In addition to analysing the densities and error measure based on them, we examine the effect of anonymisation on the poverty variable equipped with the geocoordinate. This variable indicates whether a household is classified as poor or not. Hence, for each upazila the proportion of poor households (poverty ratio) can be obtained.
When displacing the geocoordinates, the true locations can be perturbed beyond the boundary resulting in distorted poverty ratios. To assess these effects at a very fine spatial scale, the poverty ratios are evaluated at the upazila-level.

Figure 9 (a) presents the true proportion of poor households obtained by the original geocoordinates in each upazila. The proportion of poor households in the Rajshahi division ranges between 20% and 50%, with a clear boundary in west-east direction. Many neighbouring upazilas located away from the sharp boundary dividing the northern and southern regions exhibit similar poverty rates. However, since random displacement extends beyond upazila boundaries, the displacement of households results in distorted poverty estimates at the upazila level. Hence, households located on the boundaries of the upazila may be displaced to another one. Figure 9 (b) displays the difference between the poverty rates calculated using the original coordinates and those based on the displaced coordinates. Underestimation is shown in shades of blue, while overestimation is indicated in red. The closer a region is coloured paler colours, the smaller the differences. When using the displaced coordinates to compute poverty ratios, considerable discrepancies emerge in several upazilas, especially closer to the boundary in west-east direction. In contrast, when using MEM-RDD, Figure 9 (c) shows that differences are closer to zero. This improvement holds for most upazilas and on average, using the proposed estimator that accounts for data anonymisation produces estimates that are closer to the estimates obtained with the true coordinates.

For a detailed analysis of the displacement scheme using the DHS displacement radii as well as other once, see Appendix C. Here, the Setting 1 (DHS) corresponds to the setting used here.

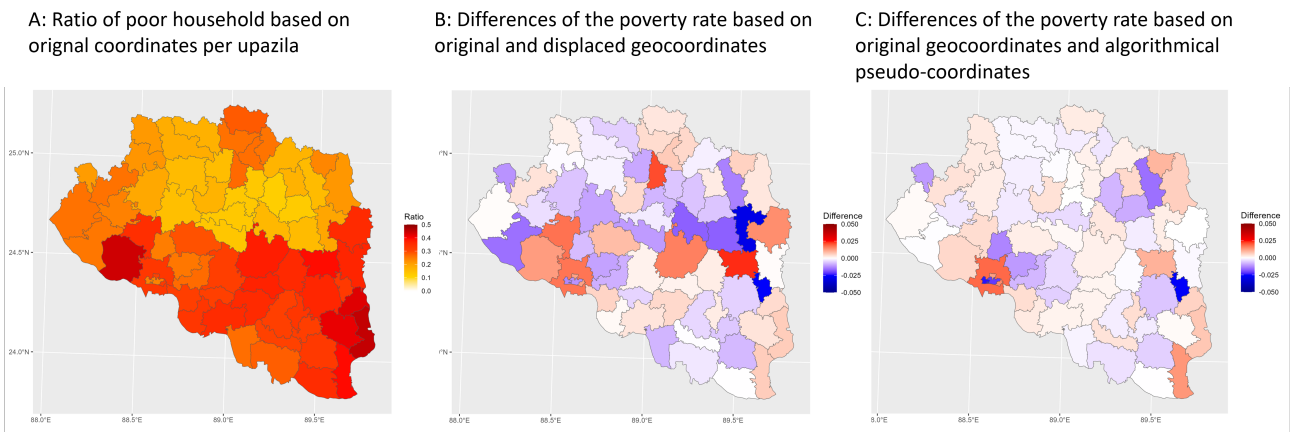

Figure 9: Proportion of poor households in each upazila (A); the effect of DHS displacement scheme on poverty estimates using the displaced coordinates (B) and the pseudo-coordinates using the proposed estimator that accounts for data anonymisation (C).

5.2 Aggregation and random displacement

In this section the analysis focuses on assessing how the results change first under aggregation to enumeration areas and then displacement of the aggregated data. Since the exact boundaries of the EAs are not publicly available, we opted to construct them synthetically using both, grid-based and polygon-based aggregation, analogously to the simulation study. For each type, we explore three different cases. For grid-based aggregation, given that the evaluation is conducted on a 250×250 meter grid, larger rounding values were required, i.e., using grids with cell sizes of 500, 1,000 and 1,500 meters.

For polygon-based aggregation, the target cluster size defines the aimed number of geocoordinates per aggregation area, i.e., 60, 120 and 240 households for this application. Followed by spatial clustering and Voronoi tessellation - for a detailed description see Section 2.3 - the polygons are constructed.

Table 4 provides an overview of the number of households aggregated in each case.

Grid-based aggregation yields equal sized rectangles and are referred to as aggregation cells or cells. On average, 113 households were grouped within the 500-meter grid, 313 within the 1,000-meter grid, and 627 within the 1,500-meter grid. These grid sizes ensure that the maximum displacement distances, which always exceed the dimensions of a single grid cell, are meaningful and feasible within the aggregation structure. Furthermore, grid-based aggregation tends to produce more pronounced extreme values of population sizes per cell, since a regular grid is applied across areas with both high and low population densities.

In contrast, the polygon-based clustering approach results in more balanced sizes. In all cases, the interquartile range is less than 25 households indicating small variation for a high number polygons. However, the polygons themselves vary in size, with smaller polygons occurring in regions of higher household density.

Grid-based aggregation							
Grid size (m)	Min	1Q	Median	Mean	3Q	Max	No. of EA
500	1	18	49	113.0	121	4112	40,692
1,000	1	42	121	313.9	313	9648	14,756
1,500	1	96	260	627.4	642	18934	7,382
Polygon-based aggregation							
Targeted size (hh)	Min	1Q	Median	Mean	3Q	Max	No. of EA
60	1	51	61	61.3	71	182	75,483
120	1	103	122	120.7	140	309	38,322
240	2	208	245	240.5	278	594	19,233

Table 4: Comparison of descriptive statistics for polygon-based and grid-based aggregation across three settings. In the grid-based approach, aggregation was performed using grid sizes of 500, 1,000, and 1,500 metres, whereas in the polygon-based approach, first spatial clustering was conducted with target cluster sizes of 60, 120, and 240 households and then Voronoi tessellation is used on the mean coordinates to obtain aggregation areas.

Figure 10 illustrates the two aggregation schemes. Plot (A) presents grid-based aggregation using a grid size of 1,000 metres, while plot (B) displays polygon-based aggregation based on a target cluster size of 120 households. As indicated in Table 4, the maximum number of aggregated households varies considerably between the two approaches. To highlight these differences, EAs containing more than 375 aggregated households are visualised in black. As expected, in the grid-based approach, the number of aggregated households is heavily influenced depending on whether the respective area is urban or rural. In regions of high population density, numerous cells exhibit high values, which are indicated in black. In contrast, polygon-based clustering results in a relatively balanced number of households per EA. However, upon closer inspection, it becomes evident that polygon sizes are considerably smaller in urban areas compared to those in rural regions. Overall, substantial differences between the grid-based and polygon-based aggregation are observed.

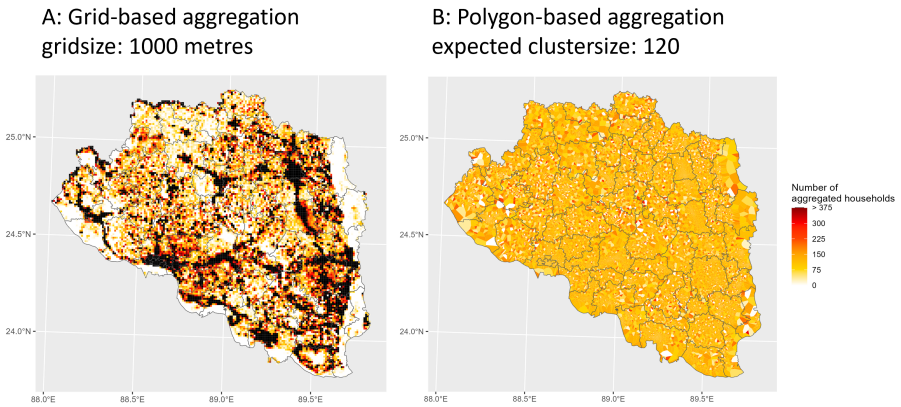

Figure 10: Comparison of the grid-based and polygon-based construction schemes for EAs. The grid-based aggregation employs a 1,000-metre resolution, whereas the polygon-based aggregation is based on a target cluster size of 120 households.

We now proceed to assess the various aggregation cases in the context of the DHS anonymisation scheme. Table 5 presents the RMISE values multiplied by 10^4 across different aggregation schemes.

The naive KDE is computed using the aggregated and displaced geocoordinates. Different aggregation types and different tuning parameters, i.e., rounding value or target cluster size, generally yield more pronounced anonymisation effects than solely displacing the geocoordinate (as in Section 5.1). Therefore, compare the RMISE of the naive estimate on (not aggregated but only) displaced data of 4.913 from Section 5.1 to the RMISE of the naive

estimates of anonymised (= aggregated + displaced) data (see Table 5 second column). Furthermore, the larger the cell size in the grid-based aggregation or the higher the targeted number of households per cluster, the greater the information loss measured by the RMISE.

Therefore, the question arises whether applying a measurement error model reduces the information loss.
The MEM-ARDD assumes knowledgment about the aggregation and displacement process, i.e., the true geometries of the aggregation process as well as the displacement radii are known. The RMISE results are given in the fourth column of Table 5. For all considered cases the information loss could be reduced.
Nevertheless, having information about the true geometries of the aggregation process is by no means guaranteed. Therefore, we introduced an approximation of the geometries based on a Voronoi tessellation using the anonymised points. These results are given fifth column of Table 5. Compared to the MEM-ARDD with full information, the RMISE is slightly higher. Nevertheless, when comparing these values to the naive estimate (second column of Table 5) a clear reduction in all cases could be achieved.
Among the MEM-ARDD, which account for both aggregation and random displacement, the MEM-RDD does not need information about the aggregation process at all, i.e., information about EA structure is omitted. Hence, also an approximation of the EAs is not needed. The RMISE when only the displacement is accounted for is given in the third column of Table 5. Compared to the MEM-ARDD having information of the EA structure the performance of the MEM-RDD differs only slightly. In some cases, the MEM-ARDD with full information outperforms the MEM-RDD, while in others the opposite holds.

Agg. Size	**Agg. and Disp. Data** (naive KDE)	**MEM under disp.** (MEM-RDD)	**MEM under agg. and disp. with Aggregation Process** (MEM-ARDD)	
			Known	**Unknown**
Grid size (m)	Grid-based			
500	9.021925	6.779919	**6.753312**	7.446810
1,000	9.743909	**8.375750**	8.385499	8.363975
1,500	10.09097	9.96455	**9.934652**	9.966229
Target size	Polygon-based			
60	7.224907	**4.773390**	4.777598	4.780384
120	7.880908	4.972655	**4.964186**	4.985212
240	8.501170	**5.215416**	5.226106	5.238818

Table 5: RMISE values using measurement error models (MEM) under different assumptions, for different aggregation types and sizes, multiplied by a factor of 10^4. The best results are highlighted in bold.

In addition to the global measure assessed by the RMISE, the two leftmost panels of Figure 11 presents the Jaccard Similarity as a local measure. For grid-based aggregation a cell size of 1,000 meters and for polygon-based aggregation a target number of 120 aggregated households are shown.
The KDE of the original geocoordinates is used as a reference. Density hotspots from the 85% to the 99% percentile (in 1% increments) of the reference are compared to the KDE of the anonymised data in orange. Furthermore, the Jaccard Similarity comparing the original KDE to the density estimates from the MEM-RDD (cyan) and the MEM-ARDD with known (dark green) and approximated geometries (light green) are depicted. Higher values indicate a greater overlap between the high-density areas of the compared density estimates. In both cases shown, a clear improvement could be achieved when considering the measurement error. The performance of the models differs only marginally; however, in more detailed graphical analysis, a slight advantage of the MEM-ARDD with known geometries can be observed.

Furthermore, the relative bias of the poverty ratio evaluated at the upazila-level for these two cases is illustrated in the two rightmost panels of Figure 11. The relative bias is calculated by comparing the poverty ratio obtained from the original coordinates with the anonymised coordinates (orange). Furthermore, it is of interest whether considering the anonymisation process as measurement error has an effect on the poverty ratios. The measurement error model is formulated as an iterative MCMC approach, with the pseudo-coordinates obtained in the sampling phase serving as the basis for estimating the poverty ratio. Multiple sets of pseudo-coordinates are summarized at upazila-level and the poverty ratio is estimated. The relative bias obtained by the comparison with the estimated poverty ratio obtained by the pseudo-samples from the MEM-RDD (cyan) as well as the MEM-ARDD with known (dark green) and approximated (light green) geometries is on average lower.

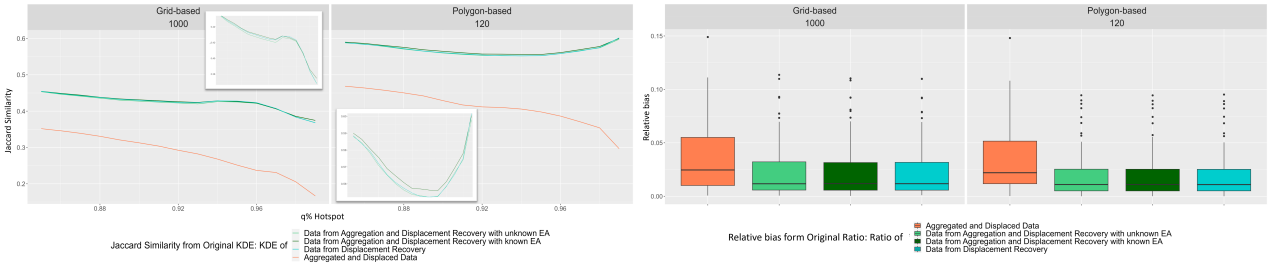

Figure 11: Jaccard Similarity and relative bias of the poverty rate of household data anonymised by the DHS scheme using grid-based or polygon-based aggregation.

In summary, considering the DHS anonymisation scheme as measurement error reduces the information loss evaluated by the considered performance measures. Among the approaches, the MEM-ARDD exhibits the best performance; however, the advantage is negligible. Given the additional information required for the MEM-ARDD, accounting for displacement alone (MEM-RDD) should prove sufficient in practice. In particular, when the precise structure of EAs is unknown, approximating them from anonymised data is not necessary, as accounting for displacement is sufficient.

6 Conclusion

Spatial anonymisation distorts population densities and associated variables. If the anonymisation scheme is known or partially known, it is possible to use this information to improve inference. For aggregated data, previous work by Groß et al. (2017) proposed an iterative method generating pseudo-samples that reflect the underlying population distribution. Building on the idea of treating the anonymisation as measurement error, a Markov Chain Monte Carlo approach was developed to handle random donut displacement as well as displacement in combination with aggregation by incorporating the corresponding anonymisation strategies as measurement error. In these iterative approaches, the geomasking of original coordinates is addressed. The proposed methods are first tested in a simulation study and then applied to household coordinates from the Rajshahi division in Bangladesh.

In the simulation study, when considering random donut displacement as measurement error, it becomes clear that the positive effects emerge under strong anonymisation. By inducing the displacement, the coordinates are more scattered and the structure of the data fades. Local and global density-based error measures were used to demonstrate that this effect can be reduced. In more complex schemes, i.e., combining aggregation and subsequent displacement, first the question of how aggregation areas are constructed arises. In applications involving official data, these areas usually correspond to administrative boundaries. Here, they have been synthetically constructed using grid-based and polygon-based mechanisms. It is important to emphasize that the combination of aggregation and displacement must be implemented with meaningful parameter choices. If the parameters are chosen inappropriately, either the aggregation has little effect, because too few coordinates have been grouped or the displacement

has no effect, because the original centroid can be clearly assigned. A nested approach for random displacement was suggested by Hossain (2023). To reduce computation time and enhance smoothness, the assumptions of the two-step approach were relaxed and a one-step approach proposed. The poroposed MEM under the consideration of RDD and ARDD show similar results in both the simulation study and the application, in particular if the tuning parameter of the displacement are chosen moderately.

The DHS anonymisation scheme is applied to household coordinates of the Rajshahi division in Bangladesh, which are generated proportional to the night time light intensity obtained from Worldpop data source. These households are equipped with the poverty and regional status using DHS data source. When only applying random donut displacement on the geocoordinates, we can clearly see the effects on the KDE and on the poverty rates per upazila. Applying the iterative approach reduces the information loss. Among the parameter for displacement of the DHS scheme, the application of a non-zero minimum distance has significant effects on the anonymisation of the data, as shown in Appendix C.
The application of both, aggregation and displacement, is again based on grid- and polygon-based aggregation to mimic EAs, which are not publicly available. The chosen settings of grid-based aggregation, showed higher effects when computing the naive estimate compared to the polygon-based aggregation. Nevertheless, these settings are hard to compare as they have different characteristics. For both EA generation approaches and their corresponding settings, no answer could be found, whether the consideration of aggregation leads to an important advantage. In polygon-based aggregation, information loss could be reduced more drastically. In all considered cases, applying the MEM reduces the bias of the poverty ratio.

In conclusion, it can be said that measurement errors introduced by the DHS anonymisation scheme should by no means be ignored. Nevertheless, the aggregation areas are too small to recommend taking the aggregations process into account. The performance measures, i.e., global and local error measures for densities and the poverty ratio, do not indicate a clear preference of the MEM-ARDD. Since displacement radii are known for DHS anonymisation scheme, taking them into account yields comparatively good results. Hence, we recommend applying MEM-RDD to data geomasked by the DHS anonymisation scheme.

Acknowledgment

The work of Lorena Gril was supported by the collaborative project AnigeD funded by the German Federal Ministry of Research, Technology and Space and the EU.

References

Allshouse, W. B. et al. (2010). "Geomasking sensitive health data and privacy protection: an evaluation using an E911 database". In: *Geocarto international* 25.6, pp. 443–452.

Altay, U. et al. (2024). "Impact of jittering on raster- and distance-based geostatistical analyses of DHS data". In: *Statistical Modelling* 25.1, pp. 55–74.

Balk, D. et al. (2004). "A spatial analysis of childhood mortality in West Africa". In: *Population, Space and Place* 10.3, pp. 175–216.

Burgert, C. R. et al. (Sept. 2013). *Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys*. SAR7. Calverton, Maryland, USA: ICF International.

Burgert-Brucker, C. R. et al. (Aug. 2016). *Guidance for Use of The DHS Program Modeled Map Surfaces*. SAR14. Rockville, Maryland, USA: ICF International.

Casella, G. and R. Berger (2024). *Statistical Inference*. 2nd. New York: Chapman and Hall/CRC, p. 565.

Clifton, K. J. and S. R. Gehrke (2013). "Application of Geographic Perturbation Methods to Residential Locations in the Oregon Household Activity Survey: Proof of Concept". In: *Proceedings of the Transportation Research Board 92nd Annual Meeting*. Washington, DC.

DHS Program (2025). *Spatial Anonymization at The DHS Program*. Accessed: 2025-01-28. URL: `https://blog.dhsprogram.com/spatial-anonymization-at-the-dhs-program/`.

Feldacker, C. et al. (2010). "The who and where of HIV in rural Malawi: Exploring the effects of person and place on individual HIV status". In: *Health & place* 16.5, pp. 996–1006.

Fox, L. et al. (2024). "Enhancing insights in sexually transmitted infection mapping: Syphilis in Forsyth County, North Carolina, a case study". In: *PLoS computational biology* 20.10, e1012464.

Gething, P. W. et al. (2015). *Creating spatial interpolation surfaces with DHS data*. DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International.

Gril, L. et al. (2025). "Kernel Heaping – Kernel Density Estimation from Regional Aggregates via Measurement Error Model". In: *The R Journal* 16.3, pp. 115–133.

Groß, M. et al. (Jan. 2017). "Estimating the Density of Ethnic Minorities and Aged People in Berlin: Multivariate Kernel Density Estimation Applied to Sensitive Georeferenced Administrative Data Protected Via Measurement Error". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 180.1, pp. 161–183.

Hampton, K. H. et al. (2010). "Mapping health data: improved privacy protection with donut method geomasking". In: *American journal of epidemiology* 172.9, pp. 1062–1069.

Hossain, J. (2023). "Statistical Estimation and Inference with Aggregated and Displaced Georeferenced Data". Doctoral Thesis. University of Southampton, p. 159.

Izenman, A. J. (1991). "Recent Developments in Nonparametric Density Estimation". In: *Journal of the American Statistical Association* 86.413, pp. 205–224.

Kounadi, O. and M. Leitner (2016). "Adaptive areal elimination (AAE): A transparent way of disclosing protected spatial datasets". In: *Computers, Environment and Urban Systems* 57, pp. 59–67.

Lohela, T. J et al. (2012). "Distance to care, facility delivery and early neonatal mortality in Malawi and Zambia". In: *PloS one* 7.12.

Lu, Y. et al. (2012). "Considering Risk Locations When Defining Perturbation Zones for Geomasking". In: *Cartographica* 47.3, pp. 168–178.

Okabe, A. and B. Boots (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. 2nd ed. Wiley.

Pande, S. et al. (2008). "Addressing diarrhea prevalence in the West African Middle Belt: social and geographic dimensions in a case study for Benin". In: *International Journal of Health Geographics* 7.1, pp. 1–17.

Perez-Heydrich, C. et al. (2016). "Influence of demographic and health survey point displacements on raster-based analyses". In: *Spatial Demography* 4.2, pp. 135–153.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Stinchcomb, D. (Oct. 2004a). "Procedures for geomasking to protect patient confidentiality". In: *Proceedings of the ESRI International Health GIS Conference*. October 17–20, 2004. Washington, DC.

— (2004b). "Procedures for geomasking to protect patient confidentiality". In: *ESRI international health GIS conference*, pp. 1–17.

Wand, M. P. and C. Jones (1994). "Multivariate plug-in bandwidth selection." In: *Computational Statistics* 9.2, pp. 97–116.

Warren, J. L. et al. (2016). "Influence of demographic and health survey point displacements on point-in-polygon analyses". In: *Spatial Demography* 4.2, pp. 117–133.

A Derivation relating random donut displacement

Figure 12 shows multiple samples for random displacing the true location for fixed δ_{min} and δ_{max}. The histograms coloured in green and yellow show that the angle and the distances are uniformly distributed over the intervals $[0, 360)$ and $[\delta_{\text{min}}, \delta_{\text{max}}]$, respectively. However, the displacement points are not uniformly distributed over the displacement area, which is shown by plotting the offset values along each axis. Although the displaced points are uniformly distributed along the angle and the distance, the resulting distribution along each coordinate, i.e., offset along the x- and y-coordinates, is not. This point is illustrated in the next two paragraphs in which the distribution of the true coordinates given the observed coordinates under random displacement is formally derived. The offset along the x- or y-axis (see in Figure 12 the blue and red histograms) represents the reflection of the potential displacement points (black) on the horizontal or vertical axis. Thereby, clear boundaries are seen when reaching the maximum displacement distance in both directions, as points can not be displaced further. The two peaks come from the non-zero minimum distance.

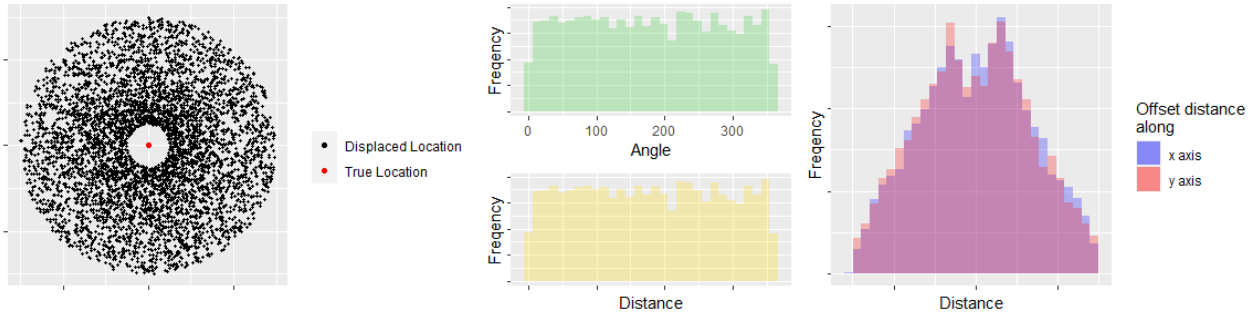

Figure 12: Random donut displacement by generating a uniformly distributed circular angle and distance between δ_{min} and δ_{max} showing that the offset on the x- and y- axis are not uniformly distributed

From Figure 12 it is evident that the distribution of the points within the donut-shape area around the true location (red) is not uniformly. Hence, the joint distribution of the offsets e_{i1} and e_{i2} are the mean focus of our analysis.

The joint distribution of e_{i1} and e_{i2} is denoted as $f_{e_{i1},e_{i2}}(e_{i1}, e_{i2})$ and is derived by transforming the variable $v = (v_1, v_2)$ to $e_k = (e_{i1}, e_{i2})$. The joint distribution of the random variable v_1 and v_2 is uniform on the unit square, as both, v_1 and v_2 are independently uniformly distributed over $[0, 1]$. Let g_1 and g_2 represent the transformation function, i.e., $g_1(v_1, v_2) = e_{k1}$ and $g_2(v_1, v_2) = e_{k2}$, respectively. The inverse transformation function for g_1 and g_2 mapping e_{k1} and e_{k2} to v_1 and v_2, respectively, are denoted by g_1^{-1} and g_2^{-1}. Hence, the joint distribution of e_{i1} and e_{i2} is given

$$f_{e_{i1},e_{i2}}(e_{i1}, e_{i2}) = f_{v_1,v_2}(g_1^{-1}(e_{i1}, e_{i2}), g_2^{-1}(e_{i1}, e_{i2})) \cdot \left|J(g_1^{-1}, g_2^{-1})\right| = 1 \cdot \left|J(g_1^{-1}, g_2^{-1})\right| = \left|J(g_1^{-1}, g_2^{-1})\right|.$$

As $\tan(2\pi v_1) = \frac{e_{k1}}{e_{k2}}$, we obtain for v_1 and v_2, respectively g_1^{-1} and g_2^{-1}

$$\|\mathbf{e}_k\|_2 = \sqrt{e_{k1}^2 + e_{k2}^2} = \delta_{min} + v_2(\delta_{max} - \delta_{min}) \quad \Rightarrow \quad v_2 = \frac{\sqrt{e_{k1}^2 + e_{k2}^2} - \delta_{min}}{\delta_{max} - \delta_{min}}$$

and

$$\frac{e_{k1}}{e_{k2}} \quad \Rightarrow \quad 2\pi v_1 = \arctan\left(\frac{e_{k1}}{e_{k2}}\right) \quad \Rightarrow \quad v_1 = \frac{1}{2\pi}\arctan\left(\frac{e_{k1}}{e_{k2}}\right),$$

if $\delta_{max} - \delta_{min} \neq 0$. Note that $||.||_2$ denotes the Euclidean distance. The partial derivatives of v_1 read as

$$\frac{\partial v_1}{\partial e_{k1}} = \frac{1}{2\pi} \cdot \frac{e_{k2}}{e_{k1}^2 + e_{k2}^2}, \quad \frac{\partial v_1}{\partial e_{k2}} = -\frac{1}{2\pi} \cdot \frac{e_{k1}}{e_{k1}^2 + e_{k2}^2},$$

and for v_2

$$\frac{\partial v_2}{\partial e_{k1}} = \frac{e_{k1}}{(\delta_{max} - \delta_{min})\sqrt{e_{k1}^2 + e_{k2}^2}}, \quad \frac{\partial v_2}{\partial e_{k2}} = \frac{e_{k2}}{(\delta_{max} - \delta_{min})\sqrt{e_{k1}^2 + e_{k2}^2}}.$$

Therefore, the Jacobian matrix of $g^{-1}(e_{k1}, e_{k2})$ reads as

$$J_{g^{-1}}(e_{k1}, e_{k2}) = \begin{pmatrix} \frac{1}{2\pi}\frac{e_{k2}}{e_{k1}^2+e_{k2}^2} & -\frac{1}{2\pi}\frac{e_{k1}}{e_{k1}^2+e_{k2}^2} \\ \frac{e_{k1}}{d_2\sqrt{e_{k1}^2+e_{k2}^2}} & \frac{e_{k2}}{d_2\sqrt{e_{k1}^2+e_{k2}^2}} \end{pmatrix}$$

and its determinant

$$f_{e_{i1},e_{i2}}(e_{i1}, e_{i2}) = \det\left(J_{g^{-1}}\right) = \frac{1}{2\pi(\delta_{max} - \delta_{min})\sqrt{e_{k1}^2 + e_{k2}^2}} \mathbb{1}_{\{\delta_{min} < ||e_k||_2 < \delta_{max}\}} \tag{12}$$

for $\delta_{max} - \delta_{min} \neq 0$. Note that $\mathbb{1}$ represents the indicator function being one, if the condition is fulfilled and zero otherwise. When considering random donut displacement in the further course of this work, the joint distribution plays an important role. The joint distribution of e_{i1} and e_{i2} can be derived by the transformation law of densities, see Casella and Berger (2024).

B Further results on the simulation study

In this section we present some further simulation results. Section B.1 provides more detail on the MCMC and discusses the relevance of the burn-in and sampling phase, while Section B.2 provides more detail on the properties of MEM-RDD. The proposed MEM-ARDD tackling aggregation and displacement as measurement error represents a further development of the two-step approach of Hossain (2023) and different properties of the methods are highlighted in a simulation.

B.1 Simulation study on MCMC using MEM-RDD

First, an analysis of the various number of sample points as outlined in Section 3.2.2, will be conducted for the four settings. When comparing the KDE of the sampling data points S_0 to the true density $f(\cdot)$ on the evaluation grid (sampling error), the data show that increasing the number of sample points leads to a closer approximation of the true density. In Table 6 the performance of the naive kernel density estimator and the density estimators using MEM-RDD is compared to the underlying density $f(\cdot)$ (equation 11) at each evaluation point x_g, $g = 1, \ldots, G$ using RMISE. The data is normalised by the sampling error. In comparison to the benchmark, f_{RDD} (obtained from MEM-RDD; Section 3.2.2) could reduce the error in each setting compared to the KDE based on the displaced data S_D. The gain from the use of the measurement error model becomes larger with increasing sample sizes. Even more important the loss due to anonymisation becomes larger for f_{S_D} with increasing sample size. Furthermore, the relative error reduction clearly increases with the number of sample points. From the data we see, that the relative standard deviation is higher in settings with strong anonymisation effects, such as Settings B or D.
Second, the burn-in and sampling phases of the MEM-RDD are analysed, as the number of iterations in each phase affects the effectiveness of the MCMC algorithm. As the number of iterations increases, both the mean deviation and the standard deviation tend to decrease. However, only a short burn-in phase is necessary to ensure stable results. An increase in iteration numbers is accompanied by an increase in the computational time required. For a detailed overview, see Table 7.

n	True Density f compared to	$\delta_{min}=0$ $\delta_{max}=0.5$	$\delta_{min}=0$ $\delta_{max}=2$	$\delta_{min}=0.5$ $\delta_{max}=1.5$	$\delta_{min}=1.5$ $\delta_{max}=2$
250	Sample KDE f_{S_0}	1	1	1	1
	Displaced KDE f_{S_D}	1.034	1.493	1.447	1.996
	Algo. KDE f_{RDD}	1.025	1.378	1.331	1.689
500	Sample KDE f_{S_0}	1	1	1	1
	Displaced KDE f_{S_D}	1.053	1.720	1.647	2.444
	Algo. KDE f_{RDD}	1.032	1.500	1.422	1.903
1000	Sample KDE f_{S_0}	1	1	1	1
	Displaced KDE f_{S_D}	1.0675	2.006	1.912	3.054
	Algo. KDE f_{RDD}	1.027	1.598	1.500	2.136
5000	Sample KDE f_{S_0}	1	1	1	1
	Displaced KDE f_{S_D}	1.137	3.011	2.830	5.128
	Algo. KDE f_{RDD}	1.025	1.764	1.626	2.646
10000	Sample KDE f_{S_0}	1	1	1	1
	Displaced KDE f_{S_D}	1.176	3.635	3.417	6.497
	Algo. KDE f_{RDD}	1.021	1.824	1.656	2.861

Table 6: RMISE for Settings A-D comparing $f(\cdot)$ with the kernel density estimates of the sample data (Sample KDE f_{S_0}), displaced points (Displaced KDE f_{S_D}), and the proposed density obtained from MEM-RDD f_{RDD} in relation to the Sample KDE. KDEs of sample and displaced data are evaluated against the iterative method using varying sample sizes n with fixed burn-in (20) and sampling (20) iterations.

Burn-in/ Sample Phase	True Density f compared to	$\delta_{min}=0$ $\delta_{max}=0.5$	$\delta_{min}=0$ $\delta_{max}=2$	$\delta_{min}=0.5$ $\delta_{max}=1.5$	$\delta_{min}=1.5$ $\delta_{max}=2$
	Sample KDE f_{S_0}	0.188 (0.000489)	0.186 (0.000457)	0.186 (0.000451)	0.186 (0.000435)
	Displaced KDE f_{S_D}	0.197 (0.000506)	0.320 (0.000425)	0.306 (0.000445)	0.457 (0.000199)
0/1	Algo. KDE f_{RDD}	0.225 (0.000506)	0.344 (0.000341)	0.329 (0.000389)	0.454 (0.000191)
1/2	Algo. KDE f_{RDD}	0.195 (0.000528)	0.302 (0.000543)	0.287 (0.000609)	0.408 (0.000349)
5/20	Algo. KDE f_{RDD}	0.193 (0.000519)	0.279 (0.000508)	0.264 (0.000560)	0.357 (0.000447)
10/50	Algo. KDE f_{RDD}	0.193 (0.000512)	0.278 (0.000502)	0.263 (0.000530)	0.355 (0.000420)
20/20	Algo. KDE f_{RDD}	0.193 (0.000519)	0.279 (0.000505)	0.264 (0.000540)	0.356 (0.000475)
20/100	Algo. KDE f_{RDD}	0.193 (0.000510)	0.277 (0.000499)	0.262 (0.000520)	0.354 (0.000410)

Table 7: Mean and variance (in parentheses) of the RMISE for settings A-D with 500 fixed sample points over 500 repetitions, comparing $f(\cdot)$, f_{S_0}, and f_{S_D} to f_{RDD} under varying MCMC burn-in and sampling phases, scaled by 10^5.

B.2 From a two-step to an one step-approach: Aggregation and random donut displacement

As short recap, an original coordinate X_i is first aggregated to centroid $M_{a,i}$, and the resulting centroids are then displaced according to a donut displacement rule. The anonymised coordinate $W_{a,i}$ is observable, for $i = 1, \ldots, n$ and $a = 1, \ldots, A$.

As previously outlined in Section 3.2.3, a two-step approach grounded in Hossain (2023) adheres closely to the underlying equation (8). Employing Bayes' theorem (2), the two-step approach involves drawing a new centroid within the displacement area surrounding the displaced point for all points aggregated and displaced **joint**ly. This step requires knowledge of the aggregation areas and their centroids. Given the anonymisation process, we certainly know the true centroid lies within the displacement area around the anonymised point. In addition, the characteristics of the aggregated and displaced data points are preserved throughout the process. The selection

of suitable centroids is performed by adapting the MEM-RDD on the centroids, i.e., the pseudo-samples can be selected from the set of centroids satisfying the condition. Based on the newly selected centroid, the data are then redistributed within the corresponding polygon or grid cell proportionally to the re-estimated density (MEM-Agg; Section 3.2.1). The procedure is repeated for each set of newly drawn centroids from the sampling phase of the MEM-RDD. It is worth noting that this approach does not guarantee that the correct centroids and subsequently the correct polygons are identified. This limitation becomes particularly problematic when the aggregation areas are large, as the algorithm may become trapped in local clusters, causing inaccurate recovery.

In Figure 13 the graphical example of Section 2 explaining the DHS anonymisation scheme is further extended to explain the different approaches of the two-step and one-step approaches. We assume that the anonymised point W (= aggregated + displaced), the set of polygons P and its responding centroids M as well as the displacement radii δ_{min} and δ_{max} are known. Figure 13 shows the two-step approach, in which a **joint** search is performed for a new centroid of the anonymised points. The centroids lying within the donut around the anonymised point W_a are identified and shown as purple hexagons. The centroid M_a (highlighted in lighter purple) associated with W_a lies within the area, but it remains unknown whether it is the true one. Based on the MEM-RDD adapted to centroids, the set of likely centroids is selected, i.e., anonymised points represented by W_a are **jointly** moved to the most likely centroid $\tilde{M}_a$. Then, points associated with $\tilde{M}_a$ are redistributed in the area. Note that these point may be redistributed outside the donut area. The points of the evaluation grid depicted as gray triangles define the set of potential pseudo-samples. It should be noted that the selected centroid does not necessarily correspond to the true one.

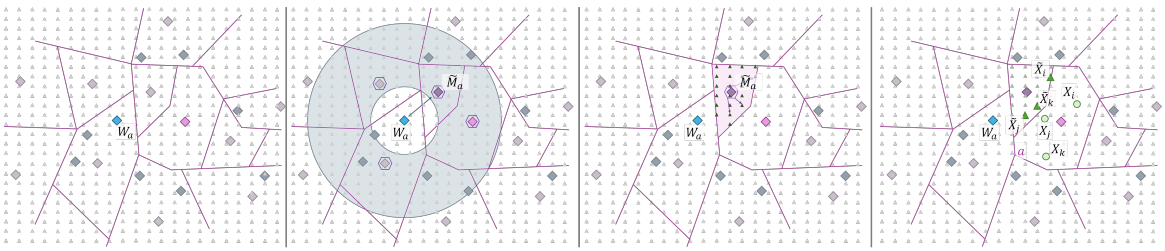

Figure 13: Exemplary setting for the joint two-step approach based on Figure 1 to show the idea on the joint reallocation of a suitable centroid and redistribution of the points within the associated area.

A natural extension of the previous approach is to relax the restriction that all points aggregated and displaced into a single location must be jointly reassigned to a new centroid (joint). Instead, each anonymised point is now treated individually and assigned to a centroid independently. Given that there is no guarantee that the correct centroid has been identified, this **separate** reassignment offers the advantage that not all points are necessarily misallocated in the same way. Once each point has been assigned to a centroid, those associated with the same centroid are distributed within the corresponding polygon. As a result, the points that were previously concentrated at a single location are more widely spread, leading to smoother kernel density estimates.

Figure 14 again refers to the example of Figure 1. Again, we assume to know the anonymised point W, the set of polygons P and its responding centroids M as well as the displacement radii δ_{min} and δ_{max}. Centroids lying within the donut area around the anonymised point W_a are identified and shown as purple hexagons. The centroid M_a (highlighted in lighter purple) associated with W_a lies within the area, but it remains unknown whether it is the true one. Based on the MEM-RDD adapted to centroids, the most likely centroids are selected, i.e., each anonymised point represented by $W_a = W_{i,a}$ are **separately** moved to the most likely centroid $\tilde{M}_{i,a}$. Note that not necessarily all points anonymised to W_a are redistributed to the same polygon and hence to the same centroid. Then, the points are redistributed in the area associated with $\tilde{M}_{i,a}$, potentially lying outside the donut area. The points of the evaluation grid depicted as gray triangles define the set of potential pseudo-samples. It should be noted that

the selected centroid does not necessarily correspond to the true one and that the points anonymised to W_a can lie within different aggregation areas and hence, some characteristics of the original aggregation area may be lost.

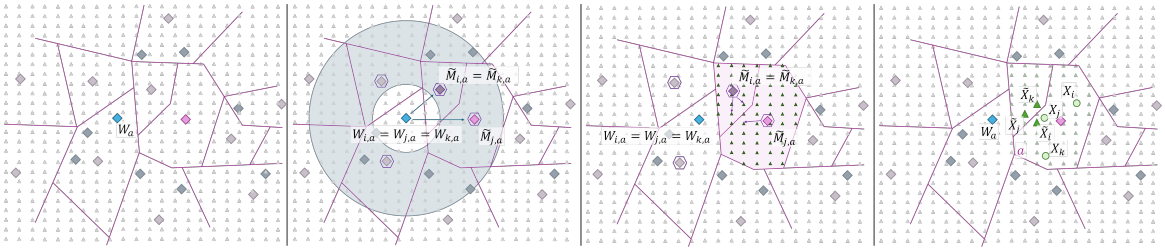

Figure 14: Exemplary setting for the separate two-step approach based on Figure 1 to show the idea on the separate reallocation to suitable centroids individually for all anonymised points. The points associated with different centroids are further redistributed within the area.

To avoid computing two nested algorithms and hence for computational efficiency, the one-step approach was developed, see Section 3.2.3. In this approach, the determination of new centroids is omitted, and the aggregated and displaced points are directly redistributed. Only those polygons whose centroids lie within the donut-shaped area surrounding the anonymised point are considered valid for redistribution. By omitting the determination of new centroids multiple times, this approach reduces computational runtime by avoiding the nesting of two loops. Similarly to the separate reassignment of centroids, the one-step approach does not account for the characteristics of the individual points that were initially aggregated into a single location and has the risk of misallocation.

Figure 15 returns to the example in Figure 1. Again, we assume that the anonymised point W, the polygons P and its responding centroids M as well as the displacement radii δ_{min} and δ_{max} are known. Centroids lying within the donut around the anonymised point W_a are identified and shown as purple hexagons. The centroid M_a (highlighted in lighter purple) associated with W_a lies within the area, but it remains unknown whether it is the true one. The points of the evaluation grid depicted as gray triangles define the set of potential pseudo-samples. The centroids identified within the donut area of the anonymised point determine the polygons that restrict the set of admissible pseudo-samples. Note that potential pseudo-samples can lie outside the donut area. Each anonymised point represented by $W_a = W_{i,a}$ is moved to the most likely pseudo-sample $\tilde{X}_i$. Note that the points anonymised to W_a can lie within different aggregation areas and hence, some characteristics of the original aggregation area may be lost.

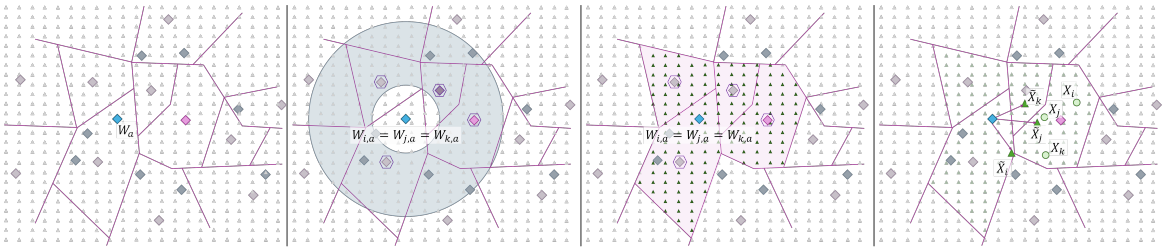

Figure 15: Exemplary setting for the one-step approach based on Figure 1 to show the idea on avoiding the selection of an centroid and subsequent redistribute, but rather redistributing the anonymised points at the individual level among the feasible set.

The numerical effects are reflected in the RMISE, see Figure 16. For all settings 500 independently generated data sets S_0 are generated and the anonymisation as well as the measurement error model explained above are applied. For all approaches, a burn-in and sampling phase of each 20 iterations is applied.
For moderate tuning parameters, all methods perform well and, in particular, outperform estimates based on the anonymised data. However, when the tuning parameters are set high and the level of anonymisation is strong, as in Setting A4, the smoothness resulting from the separate reassignment of each point becomes advantageous.

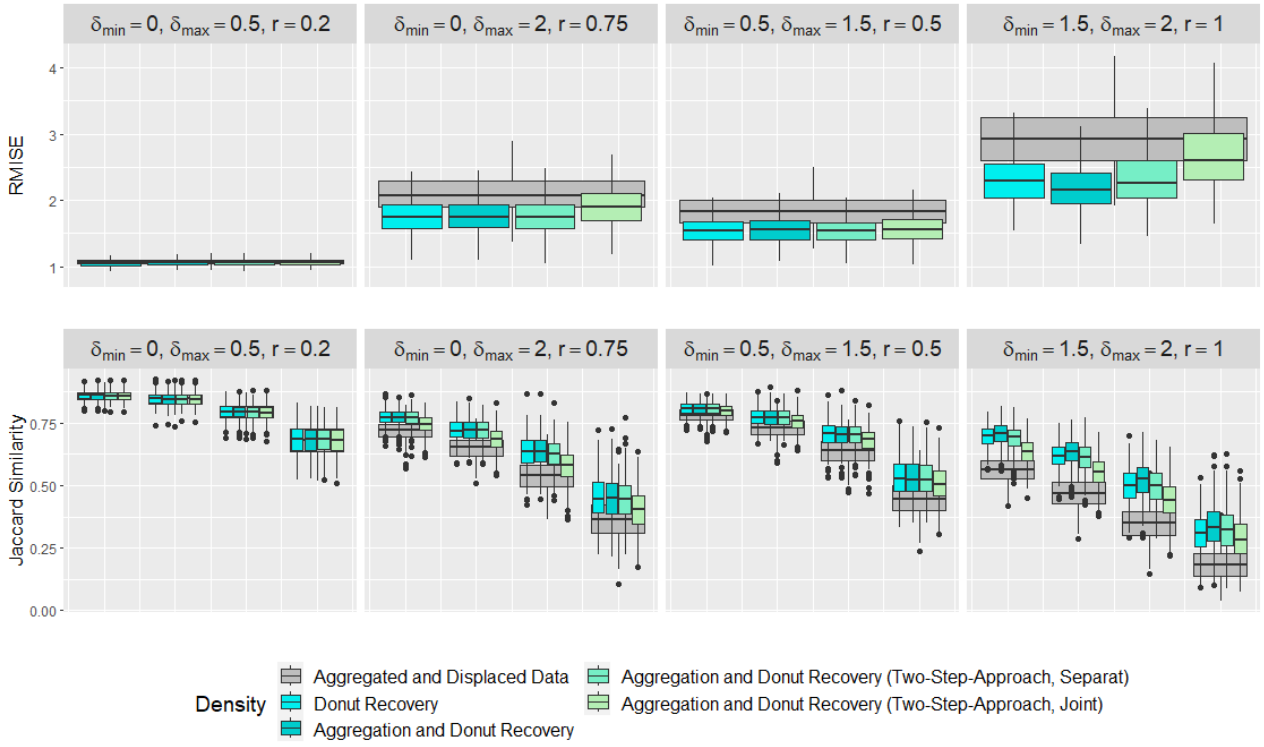

Figure 16: Performance measure of the proposed approaches using the RMISE and the Jaccard Similarity by comparing the naive estimation f_{S_D} (gray) and the estimation f_{RDD} (cyan) and f_{ARDD} (darkcyan) representing the one-step approach to the data generation process for four fixed settings for grid-based aggregation. Moreover, the two-step approach rooted in Hossain (2023) reassigning aggregation points jointly (aquamarine) and its extension to separate assignments (seagreen) are visualized.

Moreover, if the aggregation process is unknown but the tuning parameters of the random donut displacement are known, the MEM-RDD can be applied. The cyan boxplots in Figure 16 show that performing MEM-RDD produces results comparable to both the separate two-step and the one-step approaches. This outcome is clearly rooted in the design of the MEM-RDD, as each anonymised point is also reassigned on point level within the donut-shaped displacement area. However, it is important to note that this approach does not explicitly account for the underlying aggregation structure. As a result, reassigned points may potentially fall outside the theoretically feasible region, since this approach does not use the information about whether the centroids lie within the donut area or not.

A visual analysis is based on Setting A4 of the simulation study in Section 4.2. Here, using the data generation process described in equation (11), sample data S_0 of size $n = 500$ are drawn. An example of the kernel density estimation based on one set of sample data S_0 is shown in Figure 17 plot A. The KDE of anonymised points, i.e., aggregated by a rounding value of 1 and displaced by $\delta_{min} = 1.5$ and $\delta_max = 2$, is shown in Figure 17 plot B. Here, grid-based aggregation and random donut displacement are applied at the original data points. Note that these tuning parameters have strong effects on the KDE of the anonymised points. It is clear that the anonymisation process significantly distorts the usability of the data and severely impairs the underlying spatial structure of the three agglomeration areas. Compare the data generation process (11). The three agglomeration regions are hardly visible and due to the anonymisation it seems that multiple local clusters are given. Strong anonymisation effects clearly destroy the underlying data structure and motivate an iterative procedure under the knowledgment of the tuning parameters.

In Figure 17 plot C the resulting estimate using the **joint** two-step originating from Hossain (2023), who first proposed a measurement error model under both aggregation and random displacement. The limitations of this

approach, i.e., the possibility to reliably identify the correct centroid and polygon, especially in large aggregation areas where the algorithm may become trapped in local clusters, are clearly visible in the figure. Hence, due to the design of the algorithm, certain centroids remain overly prominent and do not blend smoothly into the overall spatial pattern.

Relaxing the assumption of redistributing the points solely in one polygon leads to the **separate** two-step approach. In Figure 17 plot D the smoothness of the solution can be clearly seen. However, this goes with the the risk of misallocation and characteristics of certain EAs can not be jointly transferred.

Further developing the **separate** two-step approach to improve computational efficiently leads the one-step approach, visualized in Figure 17 plot E. As these algorithms have the same feasible set, the results are similar showing that the three main aggregation centres remain identifiable, thereby enhancing the detection of local clusters.

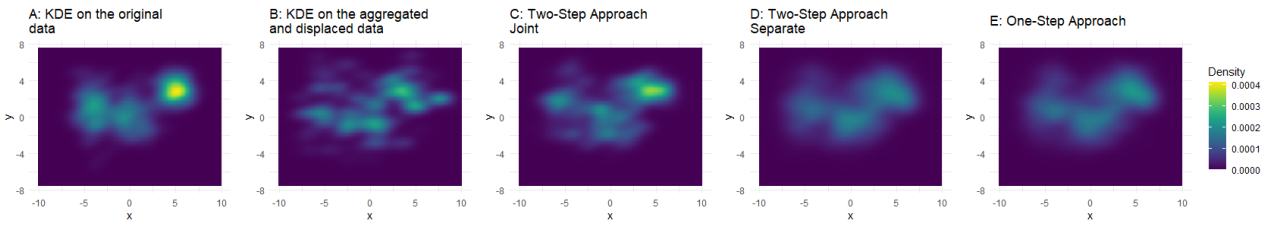

Figure 17: Comparison of density estimates of a data sample and the impact of strong spatial anonymisation using tuning parameters of Setting A4. Based on the anonymised data, the density of different algorithmic approaches, i.e., two-step approach using separate and joint reassignment and one-step approach, are visualized.

C Random donut displacement in Rajshahi division

In addition to the displacement parameters used in the DHS anonymisation scheme, described in Section 2.2, we aim to investigate the effects of alternative displacement parameters using the Bangladesh dataset. Note that here **no aggregation** is applied on the data in order to avoid making further case distinctions regarding the aggregation process. Hence, not applying aggregation and using only the displacement parameters of the DHS anonymisation scheme is referred to as *DHS displacement scheme*. While the DHS displacement scheme assumes a minimum displacement radius of zero, we explore how setting a non-zero minimum displacement distance influences the results. Furthermore, we extend the analysis to examine the effects of larger maximum displacement radii. Hence, three additional settings were developed and applied to household density data from the Rajshahi division in Bangladesh. In addition to the standard DHS displacement (Setting 1), Setting 2 introduces a minimum displacement distance of 1 km, while maintaining the original maximum displacement parameters, i.e., 2 km in urban and 5, respectively 10 km for 1% of the households in rural areas. In Settings 3 and 4, the displacement parameters were approximately doubled to simulate a more extensive anonymisation process. Specifically, the maximum displacement distance was set to 5 km in urban areas, and to 10 km in rural areas, with 1% of rural points displaced up to 20 km. Setting 3 retained a minimum displacement of 0 km, while Setting 4 implemented a minimum displacement of 2.5 km.

RMISE in comparison to true population density of Rajshahi					
Setting	δ_{min} (m)	δ_{max} in urban areas (m)	δ_{max} in rural areas (m)	Anonymised data	MEM-RDD
1 (DHS)	0	2,000	5,000 (10,000)	$4.913148 \cdot 10^{-4}$	$4.818337 \cdot 10^{-4}$
2	1,000	2,000	5,000 (10,000)	$6.912589 \cdot 10^{-4}$	$4.926492 \cdot 10^{-4}$
3	0	5,000	10,000 (20,000)	$6.420011 \cdot 10^{-4}$	$5.781869 \cdot 10^{-4}$
4	2,500	5,000	10,000 (20,000)	$9.772397 \cdot 10^{-4}$	$5.825330 \cdot 10^{-4}$

Table 8: RMISE values for different displacement distances in Rajshahi

When examining Table 8, which compares the RMISE, and Figure 18, which illustrates the Jaccard Similarity, it becomes evident that the presence of a non-zero minimum displacement radius has a substantial impact on the displaced data and, consequently, on the performance of the MEM-RDD. A comparison between Setting 1 and 2, as well as Setting 3 and 4, which differ only in terms of the minimum displacement distance, reveals a notable increase in RMISE. Specifically, the RMISE of the displaced data increases from $4.91 \cdot 10^{-4}$ in Setting 1 to $6.91 \cdot 10^{-4}$ in Setting 2, under maximum displacement distances corresponding to the DHS scheme. Under increased maximum displacement parameters, the RMISE increases from $6.42 \cdot 10^{-4}$ in Setting 3 (with no minimum distance) to $9.77 \cdot 10^{-4}$ in Setting 4 (with a fixed minimum radius). In all cases, MEM-RDD was able to reduce the RMISE. Notably, the algorithm reduced the RMISE to levels comparable to those observed when no minimum displacement radius was applied.

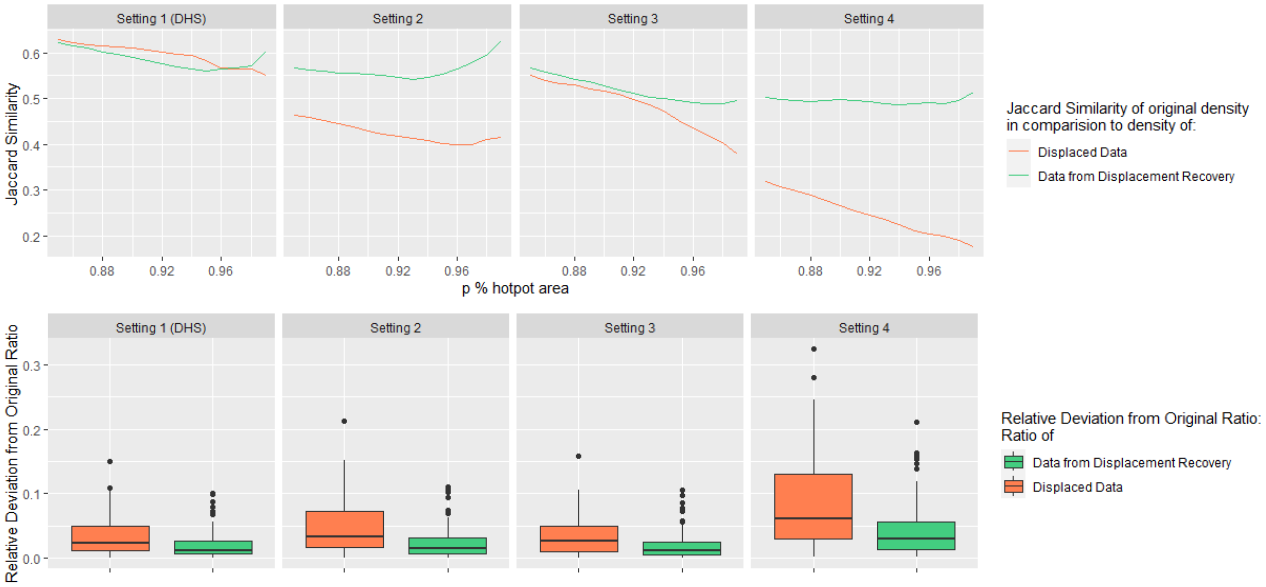

Figure 18: Jaccard Similarity and relative bias of different settings inducing random donut displacement on the household data of the Rajshahi division in Bangladesh

This effect is even more clearly visible in the Jaccard Similarity, see Figure 18. Comparing Setting 1 and 2, as well as Setting 3 and 4, demonstrates again that the introduction of a non-zero minimum distance leads to a significant reduction in overlapping of high-density hotspots. Applying the MEM-RDD, however, similarity levels comparable to those with zero minimum displacement were restored, which is evidenced by the large gap between the naive estimate (coral) and the result of the algorithm (green) in Figure 18. It is important to note that, under the DHS displacement scheme, the MEM-RDD improves the Jaccard Similarity predominantly in regions of very high population density, while in other areas it yields results comparable to those of the original displacement. It is also worth noting that while increasing the maximum displacement distance does have a visible effect, it is considerably less pronounced. This can be attributed to the nature of the displacement process. Although the distance and angle are sampled uniformly, the resulting spatial offset along the axis is not uniformly distributed, as illustrated in Figure 12. Hence, there is a higher probability that the displacement point lies nearer at the minimum distance and therefore, the introduction of a non-zero minimum distance has great effects. In addition, it is possible to see the effects of the consecutive anonymisation on the hot spots when comparing the Jaccard Similarity of Figure 11, applying the DHS anonymisation scheme (including aggregation) and Figure 18 Plot 1 in the first row, applying only the DHS displacement scheme (without aggregation). The aggregation of households to EAs has huge effects on the identification of local hot spots.

If the coordinates are additionally linked to a variable, e.g., whether a household is classified as poor, then spatial displacement can have substantial effects on the distribution of that attribute, cf. Section 5.2. Figure 18 illustrates the relative bias of the proportion of poor households per upazila, comparing the original ratios to the ratio obtained from the displaced data or from the pseudo-coordinates of MEM-RDD, respectively. The bias caused by displacement (coral) can be substantially reduced by the measurement error model (green). A pattern emerges that is consistent with the findings based on RMISE and Jaccard Similarity: the impact of a non-zero minimum displacement distance is clearly visible when comparing Settings 1 and 2, as well as Settings 3 and 4. In contrast, simply increasing the maximum displacement radius has a negligible effect on the accuracy, compare Settings 1 and 3.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business & Economics - Freie Universität Berlin

2026 erschienen:

2026/1 Hundsdoerfer, Jochen und Maren Löwe: How Do Value Added Taxes Affest Wages and Labor?
FACTS

2026/2 Gril, Lorena und Ulrich Rendtel: Mapping High-Income Taxpayers in Berlin Using Kernel-Smoothed Proportions from Aggregated Georeferenced Data
Economics