UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

**Expert meeting on Statistical Data Confidentiality**

26–28 September 2023, Wiesbaden

# An overview of data protection strategies for individual-level geocoded data

Maike Steffen, Konstantin Körner, Jörg Drechsler

Institute for Employment Research (IAB)

maike.steffen@iab.de

### *Abstract*

In response to a growing need for small-scale geographic information in various research areas, data-collecting institutions are increasingly georeferencing individual-level data. However, due to confidentiality concerns, external researchers typically have very limited access to these data if at all, resulting in a substantial loss of informational value. A growing body of literature on data protection strategies for geocoded data attempts to find solutions for the tradeoff between privacy protection and utility preservation of the individual-level data. The purpose of this paper is to systematically collect and review the literature in the field and to offer a classification of existing methods. Various strategies for estimating the utility and the remaining risk of disclosure for the protected data are also discussed.

# 1 Introduction

Geocoded data have become increasingly relevant in various research areas since they offer insights that can only be acquired considering spatial context. The granular information enables researchers to include fine geographic patterns and spatial variation of individual characteristics in their analyses. The detailed geographical information facilitates studying such diverse topics as neighborhood effects, mobility patterns, or the spread of diseases to name only a few of the possible applications. Moreover, the geo-coordinates are not subject to changes over time as it is the case with administrative borders, which often hampers longitudinal analyses. Finally, the availability of detailed geographical information allows to easily merge information from various data sources.

However, access to detailed geocoding information is currently limited as it is well known that detailed geographical information is highly identifying (De Montjoye et al., 2013). To still enable access to this valuable source of information, various strategies have been proposed in the literature to protect confidentiality while still maintaining the utility of the collected information. This paper aims to give an overview of the various approaches. We also provide an overview of metrics that have been used to assess the disclosure risk and the utility of the protected data.

The remainder of the paper is organized as follows. In Section 2, we review the three most popular approaches for protecting geocoded data: aggregation, geographic masking, and data synthesis. In Section 3, we discuss various tools which are used to assess the risk and utility of the protected data. Section 4 concludes the article.

# 2 Data Protection Strategies

Two general strategies are commonly applied to reduce the risk of disclosure when disseminating data to the public: information reduction and perturbation. Information reduction limits the amount of detail that is available in the data. This can range form discretizing continuous variables (e.g., reporting age in five-year intervals) over coarsening categorical variables (e.g., reporting only the first two digits of a hierarchical classification code such as the NACE code) to removing entire variables. Perturbation approaches try to preserve the level of detail contained in the original data. They reduce the risk of disclosure by slightly altering the microdata on the record level. Examples include noise infusion, top-coding, or swapping.

Both strategies are also used when disseminating detailed geo-information. Aggregation as a form of information reduction is probably the most widely adopted strategy to reduce the risk of reidentification. We will review different aggregation strategies in more detail in Section 2.1. The early influential paper by Armstrong et al. (1999) lists two alternative strategies to aggregation that rely on perturbation: affine transformations and geographic masking. Affine transformations are methods that displace, rescale, or rotate the entire vector of original locations. Since they are completely deterministic, these methods are relatively easy to reverse engineer. They also lead to a substantial loss of information since the transformation of the original locations are data independent and thus spatial clustering effects found in the original data can be destroyed. Furthermore, external geographical information can no longer be linked to the transformed data in a reasonable way (Zandbergen, 2014). For these reasons, these methods have never been widely adopted and we will only review geographic masking in more detail in Section 2.2.

In recent years, synthetic data approaches have emerged as another perturbation strategy. With synthetic data, original values are replaced with synthetic values drawn from a model fitted to the original data. We will review synthetic data approaches for disseminating detailed geo-information in Section 2.3.

## 2.1 Aggregation

As discussed earlier, aggregation is the most widely adopted strategy to reduce risks from reidentification. Aggregation does not alter the information, that is, the number of observations per aggregated unit remains accurate and the location of individuals may be coarsened but will not be replaced by fake locations. However, it does lead to a loss of information and thereby reduces the range of applications the data can be used for. Broadly, there are two general aggregation strategies: aggregation within pre-defined areas, such as grid cells or administrative areas, and more spatially flexible microaggregation, which ensures that each aggregation cell contains a predefined number of records.

The use of aggregation within pre-defined areas is by far the most commonly adopted approach, and guidelines to assign observations to standardized grid cells have been developed (e.g., INSPIRE, 2014). Using standardized formats comes with the advantage that additional spatial information such as climate, health, or economic data can be easily linked using these grid cells (Klumpe et al., 2020). At the same time, it is a rather inflexible strategy. If the uniformly sized grid cells are sufficiently small, they allow detailed analyses, but may not protect confidentiality adequately in sparsely populated cells. If they are large enough to protect confidentiality even in rural areas, there is a high information loss in urban areas. To address this issue, grid cell sizes can be adapted to the population density (e.g., Lagonigro et al., 2017). This approach, however, renders the linking of external grid cell data more difficult. Some researchers (e.g., Groß et al., 2017, 2020) have proposed to improve the utility of the aggregated data by applying a smoothing function based on kernel density estimators, which randomly reassigns the individuals to point locations within the aggregation cell. This strategy can, for example, be beneficial if the goal is to compute distance measures or for plotting the data on a map.

Microaggregation techniques allow to flexibly adapt the size of the aggregation area to the desired level of protection (Domingo-Ferrer and Torra, 2005; Castro et al., 2022). Research on microaggregation in the context of geographic data mainly focuses on anonymizing digital trace data (see, e.g., Domingo-Ferrer and Trujillo-Rasua, 2012; Rebollo-Monedero et al., 2011), but the approach has also been adopted to achieve strong privacy guarantees for geocoded data based on the concept of differential privacy (Soria-Cormas and Drechsler, 2013). While microaggregation can protect privacy consistently, it creates irregular polygons that are somewhat difficult to interpret and cannot easily be linked to external geographic data.

## 2.2 Geographic Masking

Geographic masking relies on randomly displacing the original location to protect confidentiality. A variety of methods have been developed in this field. The simplest form of geographic masking assigns new locations by drawing a circle with fixed radius around the original location and randomly picking a new location on that circle (Zandbergen, 2014). With such a fixed displacement distance, the risk of re-engineering the original locations from the masked data can be relatively high (Zandbergen, 2014), hence random perturbation within a predefined maximum distance from the original location is more commonly used (see Armstrong et al., 1999; Kwan et al., 2004; Zandbergen, 2014; Hampton et al., 2010). This increases the level of protection as the actual displacement distance is unknown to the end user even if the masking approach is disclosed.

Various strategies how to randomly draw the displacement distance have been proposed in the literature. One strategy is to use a uniform distribution within the radius of a circle centered on the original value (Armstrong et al., 1999; Zimmerman and Pavlik, 2008). Since this allows for the masked location to be very close or even equal to the original location, an alternative method called donut masking that provides higher confidentiality protection has been suggested (Hampton et al., 2010; Allshouse et al., 2010; Kounadi and Leitner, 2015). This masking method requires a minimum displacement distance additionally to the maximum displacement distance, forming a donut shape around the original location. An alternative approach to increase the displacement distance is N-Rand masking (Wightman et al., 2011), which also uses perturbation within a circle but draws $N$ potential displacement locations. The location that is furthest away from the original location is then selected as the final displacement location.

Instead of displacing the original locations within a circle with fixed radius and using a uniform distribution, some authors have suggested drawing the distance and direction of displacement from a bivariate Gaussian probability distribution (Cassa et al., 2006, 2008; Zimmerman and Pavlik, 2008). Compared to drawing from a uniform distribution, using a Gaussian distribution renders a displacement close to the original location more likely and therefore has little effect on spatial clusters (Cassa et al., 2006). Of course, a negative consequence is an increased risk of disclosure as most of the masked locations will be close to the original location. A variant of this method therefore uses a bimodal Gaussian distribution to approximate donut masking (Zandbergen, 2014). Note that, although unlikely, extremely high displacement distances can drawn from a normal distribution for a small fraction of the locations (Armstrong et al., 1999).

If population density in the data varies substantially, perturbation with fixed maximum distance (or fixed variance for the bivariate Gaussian approach) may lead to an unnecessarily large alteration of spatial information in highly populated areas where shorter displacement distances may suffice, and to privacy risks where population density is low and locations should be displaced more. This can be addressed by taking population density into account, such that the radius of the displacement area is larger in less densely populated areas (Kwan et al., 2004; Cassa et al., 2006; Hampton et al., 2010; Lu et al., 2012; Zurbarán et al., 2018). This results in masked data that are more similar to the original data in urban areas while offering a higher level of confidentiality protection in rural areas. With the bivariate Gaussian approach, the variance of the distribution can be set to be inversely proportional to the square of the population density (Cassa et al., 2006).

However, as illustrated in Allshouse et al. (2010), using externally provided population density data on an administrative area level as a benchmark, as done for example in Cassa et al. (2006); Hampton et al. (2010), may not sufficiently protect confidentiality in areas with high population distribution heterogeneity. As a remedy, the authors suggest tripling the displacement distance in areas with heterogeneous population distribution. Kounadi and Leitner (2016) argue that, when information is available at the point level, the actual distance to the $k$th nearest neighbor should be used to determine displacement distance rather than using external population density data at the administrative-area level.

In recent years, some authors proposed masking techniques that displace the original locations taking the actual position of the surrounding locations into account, such as Voronoi masking or location swapping (Seidl et al., 2015; Zhang et al., 2017). Voronoi Masking, developed by Seidl et al. (2015), is based on Voronoi polygons (Voronoi, 1908), which are shapes built around each single location with boundaries marking the half of the distance to the next location in any direction. A Voronoi polygon surrounding a point location contains all locations that are closer to this location than they are to any neighboring point locations in the data. In the masking process, each original location is moved to the closest point along the boundaries of its polygon, placing it in the middle between two actual locations. Seidl et al. (2019) find that this decreases map users' beliefs in being able to re-identify households. The locations are, on average, moved less in areas with higher density of the original points. At the same time, a group of at least two locations that are remote but close to each other will likely be displaced less than would be the case using random perturbation methods, and multiple locations may be relocated to the same masked location.

Since many masking approaches do not account for geographic characteristics or whether units exist at the masked location, they may generate unrealistic locations, such as within water bodies or parks. Zhang et al. (2017) propose a location swapping approach to address these concerns. This method draws a circle or donut around the original location with varying distances based on population density. Then, the original location is swapped with another location with similar geographic characteristics within the specified area. They find that location swapping yields higher values of $k$-anonymity (defined in Section 3.1) than random perturbation using the same displacement area. However, we note that when applying random perturbation techniques with a maximum displacement distance, and especially in scarcely populated areas, the actual level of $k$ achieved can be lower than the level implied by commonly applied techniques to measure $k$ and, thus, we generally do not recommend using this measure to assess the level of protection (we will discuss this problem in more detail in Section 3.1).

To address the problem with distance based perturbation techniques, Kounadi and Leitner (2016) propose adaptive areal elimination masking that guarantees a minimum $k$-anonymity for every location. This method merges predefined shapes, e.g., administrative areas, until the number of locations per polygon is $k$ or higher. The locations are then aggregated or randomly perturbed within each polygon. While this guarantees to achieve the desired level of $k$-anonymity, most polygons will contain (substantially) more than $k$ units and therefore spatial patterns can be altered excessively.

## 2.3 Synthetic Data

An alternative to the information reduction and masking methods discussed in the previous sections is to replace the true observations with draws from a statistical model, i.e., to generate synthetic data. Such datasets aim to preserve distributional properties and the spatial structure of the original data. Since these patterns are preserved at a much smaller spatial level compared to other anonymization techniques, authors such as Quick et al. (2018); Lawson et al. (2012), and Bradley et al. (2017) argue that synthetic data is able to reduce the risk of ecological fallacies (i.e., misleading inferences from the protected data, see Freedman, 1999). Two general approaches are distinguished in the literature: fully and partially synthetic data. With fully synthetic data (Rubin, 1993), all records in the released data are synthetic. Since synthesizing all variables in a dataset can be challenging for large scale surveys, Little (1993) suggested synthesizing only those variables that are either sensitive or that could be used for re-identification. See Drechsler (2011); Drechsler and Haensch (2023) for a detailed overview on the topic.

The approach has also been adopted in recent years for protecting data containing detailed geographical information. Two general strategies can be distinguished in the literature. Several papers do not synthesize the geographical information. Instead, they specifically account for the spatial structure of the data when synthesizing other variables in the dataset to improve the utility of the synthetic data. While these papers focus on protecting sensitive information in the data, i.e., reducing the risk of attribute disclosure, other approaches directly synthesize the geographical information, hence reducing the risk of reidentification. We will separately review the two strategies in the remainder of this section.

*2.3.1 Synthesizing non-geographic variables while preserving the spatial information.* Sakshaug and Raghunathan (2010) is one of the early papers that specifically adjust common synthesis strategies to preserve the detailed spatial information. The authors propose using mixed effects modeling strategies. Mixed effects synthesis models are a natural way to preserve the geographical clustering effect. These models are especially popular in the literature on small area estimation. The authors later (2014) extended their approach by incorporating area level covariates in the model, which allows to generate synthetic data even for small areas not included in the original sample. Zhou et al. (2010) offer a more rigorous treatment of the spatial information problem by modeling all variables as spatial processes and applying spatial smoothing when modeling the variables. They show that their method introduces bias for non-linear regression models and propose a strategy for choosing the smoothing function to keep this bias small. Yet another synthesis strategy is described in Quick et al. (2018), which uses a differential smoothing synthesizer for locations of home sale in San Francisco. Their approach is a two-step process. First, they model the log-transformed home sale prices using an unrestricted hierarchical model. Second, they identify spatial outliers based on the distances to their nearest neighbors, then fit a restricted hierarchical model to provide additional smoothing for higher protection. In a related approach, Quick and Waller (2018) also use a hierarchical Bayesian model that preserves spatial, temporal, and between age-groups dependencies. They synthesize county-level heart disease deaths to complete public use data, which would be suppressed at units with cases lower than 10. More recently, Koebe et al. (2023) suggest publishing two different versions of georeferenced data. The first version includes the original location, but all other attributes are synthesized using a Gaussian copula model. The second version omits the geographic identifier, but leaves the other attributes at their original values.

*2.3.2 Synthesizing the geographical information.* The first successful implementation of geographical synthesis was discussed in Machanavajjhala et al. (2008). The authors propose a strategy for synthesizing the place of living for all individuals working in the U.S. The synthesizer is used to generate the underlying data for an application called OnTheMap provided by the U.S. Census Bureau. This application graphically visualizes commuting patterns on a detailed geographical level. The authors used a Dirichlet/Multinomial model for synthesis and adjusted the Dirichlet priors such that they were able to prove that their synthesizer guaranteed some formal level of privacy called $\varepsilon - \delta$-probabilistic differential privacy (see Machanavajjhala et al. (2008) for details). However, the multinomial model used in this paper offers low utility if the population sizes or event rates are very heterogenious. To address this limitation, Quick (2021) suggests relying on Poisson models–popular

in the disease mapping literature–for differentially private data synthesis. He later extended the approach by incorporating public knowledge to further improve the utility of the synthesizer (Quick, 2022).
Another synthesis strategy proposed by Wang and Reiter (2012) is to treat the detailed geocoding information as a continuous variable and use CART models to sequentially synthesize the longitude and latitude of the geocodes. This approach was later compared in Drechsler and Hu (2021) with two other synthesis strategies for the geocodes: using a Dirichlet Process of Mixtures of Products of Multinomials (Si and Reiter, 2013; Hu et al., 2018, DPMPM) and CART models treating the geocoding information as categorical variables. The authors find that the categorical CART models offer the highest utility, but also the highest risk of disclosure. When trying to increase the level of protection, they find it to be more effective to synthesize additional variables instead of aggregating the geocoding information to a higher grid level.
Burgette and Reiter (2013) generate a partially synthetic dataset in which they synthesize the location of US census tract identifiers using a Bayesian multinomial model with a group of Dirichlet processes priors and a multiple shrinkage prior distribution. This framework is chosen because it shrinks the parameters toward a small number of learned locations, which increases the utility of the data. Paiva et al. (2014) use areal level spatial models (often called disease mapping models in the literature) to synthesize the geographical information. Although they start with exact geographies, their methods require defining fine grids over the spatial domain, then using the conditional autoregressive (CAR) model of Besag et al. (1991) to model the distribution of grid-counts. When synthesizing exact geographies, they recommend first to synthesize grid cells for each individual, and second to randomly assign each individual a location within the grid cells. The approach is computationally intensive and can be challenging to apply if the number of categorical variables or the number of levels within the variables is large. The authors also note that their partially synthetic data do not preserve the spatial pattern because the independent draws from the underlying Poisson model can imply that close geographic units in the original data might be far apart in the synthetic data. This caveat is considered by Quick et al. (2015) who extend the spatial modeling process of geo-coordinates using marked point process models, which simultaneously model the location and the variables (Liang et al., 2008; Taddy and Kottas, 2012). Specifically, the authors propose to model the data in three steps: (i) specify multinomial models for the categorical variables in the data, (ii) use a log-Gaussian Cox process to model the geographical location within each cell specified by cross classifying all categorical variables, and (iii) specify a normal regression for continuous variables given the categorical variables and location. The authors point out that estimating this model can be computationally intractable and suggest several steps and simplifying assumptions to reduce the computational burden.

# 3 Risk and Utility Assessment

Data dissemination always faces two conflicting goals: minimizing the risk of disclosure and maintaining the usefulness of the data. Therefore, it is crucial to always evaluate data protection strategies for both of these dimensions. In this section we review strategies that have been proposed in the literature to measure the utility and the level of protection for geocoded data that underwent some form of disclosure protection.

## 3.1 Risk Evaluation

The most commonly applied measure for evaluating the disclosure risk of masked geodata is spatial $k$-anonymity. It is related to the classical definition as proposed by Sweeney (2002), which states that $k$-anonymity is achieved if a record is indstinguishable from $k-1$ other records in the dataset based on a set of prespecified variables (e.g. age, sex, education). Specifically, spatial $k$-anonymity is reached if a location is indistinguishable from at least $k-1$ other locations. However, in practice it is interpreted in many different ways (Cassa et al., 2006; Allshouse et al., 2010; Hampton et al., 2010; Kounadi and Leitner, 2016; Zhang et al., 2017; Hasanzadeh et al., 2020).

There are two main definitions of $k$-anonymity for masked geodata. First, some researchers define spatial $k$-anonymity as the number of locations around the **original** point within a circle with radius equal to the displacement distance (Hampton et al., 2010; Allshouse et al., 2010). The second definition is to measure $k$-anonymity as the number of locations around the **masked** location that are within a circle with radius equal to the displacement distance (Lu et al., 2012; Zhang et al., 2017; Hasanzadeh et al., 2020). Note, however, that both approaches can overestimate the level of $k$, when random perturbation within a circle or donut is applied. This can be amplified if the maximum displacement distance depends on the population density (Allshouse et al., 2010) or is determined by the distance to the $k^{th}$ nearest neighbor. To illustrate, imagine one household located in an area with few observations or low population density which borders an urban area. If the displacement radius for this household is chosen to reach a certain level of $k$-anonymity, its maximum displacement distance will be relatively large reaching the outer areas of the urban area. A location in the urban area, on the contrary, has many neighbors in close proximity and will thus, taking $k$-anonymity as the objective, be displaced within a smaller area that does not include all possible displacements of the rural location. In this example, the rural location may be the only one that can be displaced far into the rural area. As a consequence an ill-intentioned user of the released data can be confident that a masked record in certain rural areas can only stem from one of the few observations in the rural area. Thus, neither counting the cases within a circle around the original point nor counting the cases within a circle around the masked point provides adequate information how well these points are protected. Kounadi and Leitner (2016) empirically demonstrate that to achieve the desired level of $k$-anonymity for close to 100% of the locations, the maximum distance of displacement needs to be substantially larger than the distance to the $k^{th}$ nearest neighbor.

Beyond the (often flawed) risk assessment based on spatial $k$-anonymity, strategies for measuring the remaining risk of disclosure are surprisingly limited. Some authors discuss general aspects that impact the risk of disclosure. For example, Cassa et al., 2008 point out that risks of reidentification increase when multiple protected versions of the same georeferenced dataset are published. The original locations can then be approximated by averaging of the masked locations (assuming the same records can be uniquely identified in the different datasets). The more versions of the data are published, the higher the accuracy of this approximation. As Zimmerman and Pavlik (2008) point out, the risk is particularly high when the locations are labelled or details on the masking approach are disclosed such as the maximum displacement radius.

A classical risk assessment strategy that has been used in some applications is to mount a *record linkage attack*. With these types of attacks, the intruder is assumed to possess some information about the units contained in the database (e.g., age, marital status, and employment status) and uses this information to identify units in the database. Risk measures based on record linkage attacks typically try to estimate how likely it is that such an attack will lead to a correct identification in the protected dataset. In the context of geocoded data, it is typically assumed that one of the attributes that is known to the attacker is the (approximately) exact location of the target record. Simulated record linkage attacks have for example been used in Drechsler and Hu (2021) (and implicitly in Koebe et al., 2023) to assess how well the different synthesis strategies protect the geographical information. Drechsler and Hu (2021) use risk measures originally proposed in Reiter and Mitra (2009) to specifically estimate reidentification risks for partially synthetic data. With this approach it is assumed that the attackers possess some background knowledge for a set of target records they wish to identify in the data. Based on this knowledge, they estimate the probability of a match for each unit in the released file. A match is declared for the record that has the highest average matching probability across the synthetic datasets. The risk is evaluated by means of these matches using two different measures. The first one calculates the expected number of correctly declared matches, i.e., the expected match risk. The second one calculates the number of correct unique matches, i.e., the true match rate.

Another strategy to evaluate the level of protection specifically for partially synthetic data approaches was used in Quick et al. (2018). The authors focus on spatial outliers in the original data. For those records, they generate a large number of synthetic values by repeatedly drawing from the synthesis model. They then look at histograms of the generated values. If the spatial synthesis model is overfitting, the draws from the model will be centered around the true value with limited variability potentially indicating an unacceptable risk of disclosure. Using a related idea, Quick et al. (2015) and Quick and Waller (2018) compare synthesized values with the true, confidential values. In light of privacy protection, the objective is here to obtain different values. Given that they propose releasing two versions of the same dataset (see Section 2.3), Koebe et al. (2023) measure the risk of correctly re-identifying the sensitive small-area identifiers (zip codes) in the unprotected data without geoinformation using information from the synthetic data. They train random forest models on the dataset in which the geolocations have been protected. The trained model is then run on the original data to predict the locations. The fraction of successful predictions denotes the risk measure.

## 3.2 Utility Evaluation

While offering a sufficient level of protection should always be the primary goal of any disclosure limitation strategy, it is crucial to also measure its impacts on utility. In the geocoding context, the utility is typically assessed by measuring to what extent the spatial structure of the data is maintained. The list of metrics that is used for this purpose in the literature is almost as large as the disclosure avoidance literature itself. Here, we only focus on the utility assessment based on spatial pattern retention. A more general discussion on utility evaluations can be found for example in Domingo-Ferrer et al. (2012). In the following, we will classify the various approaches into four broad categories: (1) point locations and density measures; (2) cluster analysis; (3) spatial autocorrelation; and (4) land use assessment.

*3.2.1 Point Locations and Density Measures.* Utility evaluations often start by graphically comparing the population densities of the confidential data and the protected data. A simple approach is to visually compare the locations on a map (e.g., Kwan et al., 2004). However, unless the original data is non-confidential, this approach can only be used internally, as the plots of the original data might spill sensitive information otherwise. A more versatile approach is to estimate the population density using kernel density estimation (Shi et al., 2009; Gatrell et al., 1996). The kernel density estimator creates a smooth density surface which allows to graphically compare the densities of the original and masked data on a heatmap (e.g., Kwan et al., 2004; Zandbergen, 2014). The heatmaps can be used to either visualize the density levels for each dataset separately or to directly display the discrepancies between the two densities. Beyond visualizing the population densities (e.g., Gatrell et al., 1996) the approach can also be used to measure spatial discrepancies in any other variable contained in the data. For example, Seidl et al. (2015) show differences in total warm water consumption among others.

*3.2.2 Clustering.* Another common approach to evaluate the utility of the protected dataset is to assess whether the data show similar clustering behavior as the original data. A descriptive statistic that is often used to describe clustering in a point pattern is Ripley's $K$ function (see, e.g., Kwan et al., 2004; Zhang et al., 2017; Quick et al., 2015; Seidl et al., 2015; Drechsler and Hu, 2021). It is defined as expected number of points within a predefined radius around the location of interest normalized by the average point density across the entire geographical area covered in the data (Ripley, 1976; Kwan et al., 2004). It assesses to which extent a point pattern deviates from spatial homogeneity (Drechsler and Hu, 2021). Based on the $K$ function, the more easily interpretable $L$ function can be computed. It takes values close to zero for homogeneously distributed data, while positive values indicate heterogeneity or clustering. Closely related, the cross-$K$ function and its analog for the $L$ statistic assess the clustering of one point pattern relative to another point pattern, for example the underlying population distribution (Kwan et al., 2004).

As an alternative measure, Zhang et al. (2017) apply an average nearest-neighbor analysis to quantify how well the spatial pattern of the original data is preserved. Specifically, they compute a nearest-neighbor index that consists of the average distances from each unit to its nearest neighbor (measured in, e.g., Euclidean or Manhattan distance). An index value similar to that of the original data indicates comparable clustering intensity. In a related approach, Lu et al. (2012) apply a nearest-neighbor index that compares the average distance to the nearest neighbor with the expected distance assuming a uniform distribution of the locations. Values below one indicate clustering. Seidl et al. (2015) use a nearest-neighbor hierarchical clustering analysis to compare the number of clusters on the first level (clusters of individual data points) in the data (see also Levine, 2006; Kounadi and Leitner, 2015). They also compare standard deviational ellipses between the original and the protected data. These ellipses cover the area that is within, say, one or two standard deviations from the center of

the cluster (Kounadi and Leitner, 2015). They facilitate understanding the two-dimensional clustering behavior. Another measure to assess clustering and to identify hotspots is the Gi* statistic proposed by Getis and Ord (1992); Ord and Getis (1995). The Gi* statistic can be used to test the null hypothesis of spatial independence. Rejecting the null hypothesis indicates clustering (Getis and Ord, 1992). Kounadi and Leitner (2015) develop an indicator that combines nearest-neighbor hierarchical clustering and the Gi* statistic.
In health research, SatScan (Kulldorff, 1997) is a popular software tool for disease mapping. It can be used to identify spacial and temporal clustering in the data (Kulldorff et al., 2005). Several authors (Olson et al., 2006; Cassa et al., 2006; Hampton et al., 2010) use the software to compare the sensitivity and specificity of the underlying cluster detection approach run on the original and protected data.
Finally, some researchers use the original and masked dots to identify a data-dependent geographical area. The utility of the protected data is assessed by measuring the overlap of this area between the two datasets. For example, Hasanzadeh et al. (2017) propose an approach that compares the similarity of individuals' frequently visited points. Specifically, they extend the residential points to home areas, where the edges mark locations that are visited frequently. Large overlaps of the home areas of the protected and the confidential data indicate high similarity of individuals' neighborhoods in both datasets.

*3.2.3 Spatial Autocorrelation.* While clustering analysis focuses on identifying the number and size of clusters in the data, spatial autocorrelation more generally assesses the spatial dependence in a point pattern. Both approaches are closely related. A prevalent measure for spatial autocorrelation is Moran's I (e.g., Ord and Getis, 1995; Lu et al., 2012; Seidl et al., 2015). It tests whether the null hypothesis that the spatial autocorrelation is zero can be rejected. If this is the case, spatial autocorrelation can be assumed. Another common measure to compare spatial autocorrelation between datasets is the empirical semivariogram. (Matheron, 1963; Quick et al., 2018; Seidl et al., 2015)). It visualizes the homogeneity of non-geographic variables as a function of the distance between the locations. An output graph that increases and then flattens with further distance indicates positive spatial autocorrelation.

*3.2.4 Land use.* Another widely used approach to measure the utility of masked geodata is to compare the geography of the masked point-coordinates with their original counterparts. Quick and Waller (2018) and Zhang et al. (2017) consider, for instance, land cover categories or the proximity to roads. Regarding land cover rates, they compare whether the point-locations are in the same raster of either urban or rural areas. In an optimal scenario, the protected data would have the same share of points in urban areas as the original. Analogously, this applies to the proximity to roads, where the authors measure the closest distance of each point to the next road. The distances are compared using cumulative distribution functions (cdfs). The closer the two cdfs from the original and the protected data, the higher the utility of the protected data. Related works (e.g., Hasanzadeh et al., 2020) also evaluate other geographic characteristics such as the greenness of the surroundings.

# 4 Conclusion

Broad access to detailed geo-information can enhance the understanding of our society in numerous ways. Thus, it is not surprising that many data disseminating agencies are currently discussing how to provide access to these data for external researchers without compromising the confidentiality of the units contained in the data. Optimizing the trade-off between offering high utility granular information and sufficient data protection has been the subject of various methods for disclosure protection. In this paper, we have reviewed the literature on protection strategies for georeferenced microdata. Its main strands can be divided into coarsening the geo-information, masking it by altering, perturbing, or swapping the original locations, and disseminating synthetic data instead of the original data. We also discussed the different methods that are used to evaluate the risk and utility of the protected data. When assessing the risk of disclosure, we found that many papers rely on different notions of $k$-anonymity. We discussed a key concern with these notions, namely that for many of the distance based masking techniques, disclosure risks are underestimated based on this procedures as the obtained value of $k$ tends to be much larger than the true number of indistinguishable records. We therefore strongly advice against using spatial $k$-anonymity in this context. Regarding the utility evaluation, we conclude that there are many useful approaches discussed in the literature and that it would be an interesting avenue for future research to consolidate the plethora of different measures.

## References

Allshouse, W. B., M. K. Fitch, K. H. Hampton, D. C. Gesink, I. A. Doherty, P. A. Leone, M. L. Serre, and W. C. Miller (2010). Geomasking sensitive health data and privacy protection: an evaluation using an e911 database. *Geocarto international 25*(6), 443–452.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine 18*(5), 497–525.

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics 43*, 1–20.

Bradley, J. R., C. K. Wikle, and S. H. Holan (2017). Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society Series B: Statistical Methodology 79*(3), 815–832.

Burgette, L. F. and J. P. Reiter (2013). Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian analysis (Online) 8*(2).

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

Cassa, C. A., S. C. Wieland, and K. D. Mandl (2008). Re-identification of home addresses from spatial locations anonymized by gaussian skew. *International journal of health geographics 7*, 1–9.

Castro, J., C. Gentile, and E. Spagnolo-Arrizabalaga (2022). An algorithm for the microaggregation problem using column generation. *Computers & Operations Research 144*, 105817.

De Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports 3*(1), 1–5.

Domingo-Ferrer, J., L. Franconi, S. Giessing, E. Nordholt, K. Spicer, P. de Wolf, and A. Hundepool (2012). *Statistical Disclosure Control*. Wiley Series in Survey Methodology. Wiley.

Domingo-Ferrer, J. and V. Torra (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery 11*, 195–212.

Domingo-Ferrer, J. and R. Trujillo-Rasua (2012). Microaggregation-and permutation-based anonymization of movement data. *Information Sciences 208*, 55–80.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*, Volume 201. Springer Science & Business Media.

Drechsler, J. and A.-C. Haensch (2023). 30 years of synthetic data. *arXiv preprint arXiv:2304.02107*.

Drechsler, J. and J. Hu (2021). Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data. *Journal of Survey Statistics and Methodology 9*(3), 523–548.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy. *International Encyclopedia of the social & Behavioral sciences 6*(4027-4030), 1–7.

Gatrell, A. C., T. C. Bailey, P. J. Diggle, and B. S. Rowlingson (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, 256–274.

Getis, A. and J. K. Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis 24*(3), 189–206.

Groß, M., A.-K. Kreutzmann, U. Rendtel, T. Schmid, and N. Tzavidis (2020). Switching between different non-hierachical administrative areas via simulated geo-coordinates: a case study for student residents in berlin. *Journal of Official Statistics 36*(2), 297–314.

Groß, M., U. Rendtel, T. Schmid, S. Schmon, and N. Tzavidis (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. *Journal of the Royal Statistical Society Series A: Statistics in Society 180*(1), 161–183.

Hampton, K. H., M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. L. Serre, and W. C. Miller (2010). Mapping health data: improved privacy protection with donut method geomasking. *American journal of epidemiology 172*(9), 1062–1069.

Hasanzadeh, K., A. Broberg, and M. Kyttä (2017). Where is my neighborhood? a dynamic individual-based definition of home ranges and implementation of multiple evaluation criteria. *Applied geography 84*, 1–10.

Hasanzadeh, K., A. Kajosaari, D. Häggman, and M. Kyttä (2020). A context sensitive approach to anonymizing public participation gis data: From development to the assessment of anonymization effects on data quality. *Computers, Environment and Urban Systems 83*, 101513.

Hu, J., J. P. Reiter, and Q. Wang (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis 13*(1), 183–200.

INSPIRE (2014). Data specification on geographical grid systems – technical guidelines. Technical Report D2.8.I.2, European Commission.

Klumpe, B., J. Schröder, and M. Zwick (Eds.) (2020). *Qualität bei zusammengeführten Daten*. Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute. Springer VS Wiesbaden.

Koebe, T., A. Arias-Salazar, and T. Schmid (2023). Releasing survey microdata with exact cluster locations and additional privacy safeguards. *Humanities and Social Sciences Communications 10*(1), 1–13.

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

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

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods 26*(6), 1481–1496.

Kulldorff, M., R. Heffernan, J. Hartman, R. Assunçao, and F. Mostashari (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine 2*(3), e59.

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

Lagonigro, R., R. Oller, J. C. Martori, et al. (2017). A quadtree approach based on european geographic grids: reconciling data privacy and accuracy.

Lawson, A. B., J. Choi, B. Cai, M. Hossain, R. S. Kirby, and J. Liu (2012). Bayesian 2-stage space-time mixture modeling with spatial misalignment of the exposure in small area health data. *Journal of agricultural, biological, and environmental statistics 17*, 417–441.

Levine, N. (2006). Crime mapping and the crimestat program. *Geographical analysis 38*(1), 41–56.

Liang, S., B. P. Carlin, and A. E. Gelfand (2008). Analysis of minnesota colon and rectum cancer point patterns with spatial and nonspatial covariate information. *The annals of applied statistics 3*(3), 943.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407.

Lu, Y., C. Yorke, and F. B. Zhan (2012). Considering risk locations when defining perturbation zones for geomasking. *Cartographica: The International Journal for Geographic Information and Geovisualization 47*(3), 168–178.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008). Privacy: Theory meets practice on the map. In *2008 IEEE 24th international conference on data engineering*, pp. 277–286. IEEE.

Matheron, G. (1963). Principles of geostatistics. *Economic geology 58*(8), 1246–1266.

Olson, K. L., S. J. Grannis, and K. D. Mandl (2006). Privacy protection versus cluster detection in spatial epidemiology. *American Journal of Public Health 96*(11), 2002–2008.

Ord, J. K. and A. Getis (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis 27*(4), 286–306.

Paiva, T., A. Chakraborty, J. Reiter, and A. Gelfand (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in medicine 33*(11), 1928–1945.

Quick, H. (2021). Generating poisson-distributed differentially private synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(3), 1093–1108.

Quick, H. (2022). Improving the utility of poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to cdc wonder. *Journal of Survey Statistics and Methodology 10*(3), 596–617.

Quick, H., S. H. Holan, and C. K. Wikle (2015). Zeros and ones: a case for suppressing zeros in sensitive count data with an application to stroke mortality. *Stat 4*(1), 227–234.

Quick, H., S. H. Holan, and C. K. Wikle (2018). Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *Journal of the Royal Statistical Society Series A: Statistics in Society 181*(3), 649–661.

Quick, H., S. H. Holan, C. K. Wikle, and J. P. Reiter (2015). Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics 14*, 439–451.

Quick, H. and L. A. Waller (2018). Using spatiotemporal models to generate synthetic data for public use. *Spatial and Spatio-Temporal Epidemiology 27*, 37–45.

Rebollo-Monedero, D., J. Forné, and M. Soriano (2011). An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. *Data & Knowledge Engineering 70*(10), 892–921.

Reiter, J. P. and R. Mitra (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality 1*(1).

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability 13*(2), 255–266.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Sakshaug, J. W. and T. E. Raghunathan (2010). Synthetic data for small area estimation. In J. Domingo-Ferrer and E. Magkos (Eds.), *Privacy in Statistical Databases*, Berlin, Heidelberg, pp. 162–173. Springer Berlin Heidelberg.

Sakshaug, J. W. and T. E. Raghunathan (2014). Generating synthetic data to produce public-use microdata for small geographic areas based on complex sample survey data with application to the national health interview survey. *Journal of Applied Statistics 41*(10), 2103–2122.

Seidl, D. E., P. Jankowski, and A. Nara (2019). An empirical test of household identification risk in geomasked maps. *Cartography and Geographic Information Science 46*(6), 475–488.

Seidl, D. E., G. Paulus, P. Jankowski, and M. Regenfelder (2015). Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography 63*, 253–263.

Shi, X., J. Alford-Teaster, and T. Onega (2009). Kernel density estimation with geographically masked points. In *2009 17th International Conference on Geoinformatics*, pp. 1–4. IEEE.

Si, Y. and J. P. Reiter (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of educational and behavioral statistics 38*(5), 499–521.

Soria-Cormas, J. and J. Drechsler (2013). Evaluating the potential of differential privacy mechanisms for census data. In *UNECE Work Session on Data Confidentiality*.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems 10*(05), 557–570.

Taddy, M. A. and A. Kottas (2012). Mixture modeling for marked poisson processes. *Bayesian Analysis 7*(2).

Voronoi, G. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelles Journal) 1908*(134), 198–287.

Wang, H. and J. P. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. *The annals of applied statistics 6*(1), 229.

Wightman, P., W. Coronell, D. Jabba, M. Jimeno, and M. Labrador (2011). Evaluation of location obfuscation techniques for privacy in location based information systems. In *2011 IEEE Third Latin-American Conference* *on Communications*, pp. 1–6. IEEE.

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

Zhang, S., S. M. Freundschuh, K. Lenzer, and P. A. Zandbergen (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science 44*(1), 22–34.

Zhou, Y., F. Dominici, and T. A. Louis (2010). A smoothing approach for masking spatial data. *The Annals of Applied Statistics 4*(3), 1451–1475. DOI: 10.1214/09-AOAS325.

Zimmerman, D. L. and C. Pavlik (2008). Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical analysis 40*(1), 52–76.

Zurbarán, M., P. Wightman, M. Brovelli, D. Oxoli, M. Iliffe, M. Jimeno, and A. Salazar (2018). Nrand-k: Minimizing the impact of location obfuscation in spatial analysis. *Transactions in GIS 22*(5), 1257–1274.