geographical analysis

Check for updates

Geographical Analysis (2024) **0**, n/a

Exploring the Tradeoff Between Privacy and Utility of Complete-count Census Data Using a Multiobjective Optimization Approach

Yue Lin[1,2] and Ningchuan Xiao[1]

[1]Department of Geography, The Ohio State University, Columbus, Ohio, USA, [2]Center for Spatial Data Science, University of Chicago, Chicago, Illinois, USA

Privacy and utility are two important objectives to consider when releasing census data. However, these two objectives are often conflicting, as protecting privacy usually necessitates introducing noise into the data, which compromises data utility. Determining the appropriate level of privacy protection presents a significant challenge in the data release. Therefore, it is necessary to investigate the tradeoff between privacy and utility before making a final decision on the level of privacy protection. In this article, we propose a multiobjective optimization framework to generate multiple optimal solutions that satisfy the two objectives of privacy and utility, as well as to analyze the tradeoff between privacy and utility for decision-making. This framework relocates individuals susceptible to revealing their identities to protect their privacy. We maximize the number of individuals relocated while maximizing the utility of the data after relocations. The proposed framework is tested using synthetic population data in Franklin County, Ohio. Our experimental results show that the framework can efficiently generate a collection of optimal solutions and can be used to effectively balance privacy and utility.

Introduction

Complete-count census data enumerates every individual and household of the United States and includes information on their demographic, social, and housing characteristics. Such data has been widely used in a variety of social, economic, health, and demographic applications (Donnelly 2019; Walker 2023). In line with the Census Bureau's commitment to promoting public participation and ensuring "everyone counts," measures have been taken to protect individual privacy by removing personal identifiers such as names and addresses from data releases. However, even without these identifiers, privacy concerns can still arise because the data contains information such as locations and demographic attributes that can be used to disclose the identities of individuals (Sweeney 2000; Lin and Harvey 2015; Lin and Xiao 2023a).

Correspondence: Yue Lin, Center for Spatial Data Science, University of Chicago, 1155 E 60th St, Chicago, IL 60637, USA. Email: liny2@uchicago.edu

Submitted: March 17, 2023. Revised version accepted: January 11, 2024.

doi: 10.1111/gean.12388

An illustrative case is found in the research conducted by Abowd and Hawes (2023) on the 2010 U.S. Census data, which demonstrate that using a combination of census block, sex, and single year of age can differentiate a person from others in 44% of the national population.

Many methods have been proposed to enhance the privacy protection of census data, and these methods often introduce noise into data so that the chance (or risk) of disclosure can be reduced (Shlomo, Tudor, and Groom 2010; Abowd 2018; Abowd et al. 2022; Ruggles and Magnuson 2023). For example, the U.S. Census Bureau employs a method called data swapping to protect the 2010 Census data, which exchanges the locations of people who have high disclosure risks with those in different locations, thereby introducing noise to individual locations to protect privacy (Dalenius and Reiss 1982). While introducing noise helps to protect privacy, doing so will inevitably reduce the usefulness of data and therefore compromises data utility. Purdam and Elliot (2007), for instance, demonstrate that data swapping can significantly impact the utility of the published census data as well as the validity of findings derived from analyzing such data. The term "data utility" can encompass various aspects, such as completeness, relevance, consistency, and other requirements of users or applications (Veregin 1999). But within the privacy literature, data utility often assumes a more specific definition, referring to a numerical assessment of the usefulness of the released data to users that involves quantifying the disparity between the privacy-preserved data and the original data (Duncan, Keller-McNulty, and Stokes 2004; Li and Li 2009). In our research, we adhere to this specific convention.

The privacy concerns revolving census data are essentially a location privacy issue, as identity disclosure with census data involves not only attribute information such as age and race, but also geographic information at multiple levels such as census block, block group, and tract. In addition, census data exhibits unique characteristics as a spatial data set, where individuals or households in close locational proximity often share similarities. The literature on location privacy has seen the development of privacy-preserving techniques applicable to a broad spectrum of geographic data beyond census data, including disease locations, crime incidents, and individual GPS waypoints. These methods consider not only privacy protection but also the utility of geographic data. For example, Wieland et al. (2008) propose an optimization method to relocate sensitive disease-related points to protect patients' privacy while minimizing the expected distance displaced required to achieve a certain privacy level. In other research, some incorporate strategies into method design to eliminate the displacement distance for point data to maintain utility (Seidl, Jankowski, and Tsou 2016; Zurbarán et al. 2018; Houfaf-Khoufaf, Touya, and Le Guilcher 2021; Lin 2023). Additionally, there are studies that assess the utility of privacy-preserved data as an indicator of the effectiveness of privacy-preserving methods (Kounadi and Leitner 2016; Wang, Kim, and Kwan 2022). Recent years have also witnessed a new branch of research dedicated to eliminating false identifications, thus avoiding the relocation of points to improbable locations (e.g., moving households to water bodies) in order to enhance data utility (Richter 2018; Seidl, Jankowski, and Clarke 2018; Polzin and Kounadi 2021). These research endeavors offer invaluable insights for addressing the privacy issues associated with census data by prioritizing the importance of both privacy protection and data utility.

The public release of census data can be thought of as a decision-making problem, where a solution to such a problem should be evaluated by two objectives: the level of privacy protection and the utility of the data. These two objectives, however, are often conflicting because privacy protection necessitates the introduction of noise that reduces data utility. In addition, a complex network of actors are involved in this problem, including decision-makers such as administrative leadership, data management teams, and legal and compliance teams, as well as stakeholders

such as data subjects (e.g., civilians) and data users (e.g., communities, researchers, planners, and marketers). As with many decision-making problems that involve multiple and conflicting objectives as well as diverse actors (Brill Jr. 1979; Brill Jr., Chang, and Hopkins 1982; Deb 2001), there may not be a single optimal or best solution to maximize both privacy protection and data utility. Instead, there often exists a set of optimal solutions where no alternative is available to improve one objective without degrading the other. The practical release of census data often requires the decision-makers and stakeholders to negotiate and weigh priorities and demands in order to determine the most preferred solution. In order to make sound decisions, it is important for them to be able to survey a wide range of optimal solutions to the problem.

The decision-making problem of releasing census data that balances between privacy protection and data utility will be referred to as the privacy-utility tradeoff (PUT) problem in this article. The literature on generating multiple optimal solutions to the PUT problem has been limited. While practical approaches such as data swapping exist, they are not designed to produce optimal solutions. Other methods typically fix either privacy protection or data utility and then optimize the other to obtain one singular optimal solution (Abowd et al. 2022; Lin and Xiao 2023b). In the 2020 U.S. Census, for example, the Census Bureau uses a privacy parameter to maintain the desired level of privacy protection and then applies a series of optimization models to derive an optimal solution that preserves the level of privacy protection while maximizing data utility (Abowd et al. 2022). However, without considering other optimal solutions, this singular solution may not fully reflect how privacy and utility trade off, and doing so also lacks transparency to explain why the particular level of privacy protection is chosen over others.

The purpose of this article is to develop a modeling framework for generating and evaluating multiple optimal solutions to the PUT problem. While the decision-making process in general is a fascinating topic that deserves much attention, we set the scope of this article on a methodological ground in the mathematical formulation of the PUT problem and illustrate how to explore the tradeoffs revealed by solving such a problem. The proposed modeling framework is developed based on multiobjective optimization, which has been widely used to generate optimal solutions to decision-making problems that have multiple conflicting objectives (Deb 2001; Cohon 2004; Xiao, Bennett, and Armstrong 2007; Branke et al. 2008). In the remainder of this article, Section The Privacy-Utility Tradeoff Problem describes the PUT problem. Section Methods details the proposed modeling framework by formally describing the optimization model and the measurement of privacy protection and data utility. In Section Computational Experiments, the proposed framework is applied to data in Franklin County, Ohio to demonstrate its effectiveness in generating and evaluating optimal solutions to the PUT problem. We conclude the article in Section Discussion and Conclusions.

The privacy-utility tradeoff problem

Individual-level data contains the geographic location and attributes of each individual in a population. The geographic location is typically collected at the street address level, but for confidentiality reasons, the address-level location is often replaced by a small area, such as a census block, which is presumed to contain multiple individuals rather than a single person. We denote the number of geographic locations in the individual-level data as n. Fig. 1a illustrates such an individual-level data set in a table form, which includes two ($n = 2$) block-level locations (replacing each individual's address-level location) and three attributes (voting age, race, and ethnicity) of 10 individuals. Based on the attributes in the individual-level data, we form a

Person	Block	Voting age	Race	Ethnicity
1	1	Age < 18	White	Non-Hispanic
2	1	Age < 18	White	Hispanic
3	1	Age ≥ 18	White	Non-Hispanic
4	1	Age ≥ 18	White	Non-Hispanic
5	1	Age ≥ 18	White	Non-Hispanic
6	1	Age ≥ 18	Black	Hispanic
7	2	Age < 18	White	Non-Hispanic
8	2	Age < 18	White	Non-Hispanic
9	2	Age < 18	White	Non-Hispanic
10	2	Age < 18	White	Non-Hispanic

(a)

	Block 1	Block 2
1: <Age < 18, White, Non-Hispanic>	1	4
2: <Age ≥ 18, White, Non-Hispanic>	3	0
3: <Age < 18, White, Hispanic>	1	0
4: <Age ≥ 18, White, Hispanic>	0	0
5: <Age < 18, Black, Non-Hispanic>	0	0
6: <Age ≥ 18, Black, Non-Hispanic>	0	0
7: <Age < 18, Black, Hispanic>	0	0
8: <Age ≥ 18, Black, Hispanic>	1	0

(b)

Figure 1. Census data at the individual level. (a) Table form; (b) Matrix form.

combination of attribute values using one value from each of these attributes (e.g., <Age<18, White, Non-Hispanic>), and the number of all possible combinations is denoted as m. In our example, there are eight possible combinations ($m = 8$) containing one value from each of the three attributes, as listed on the left of Fig. 1b. The individual-level data can then be transformed into an $m \times n$ matrix $\mathbf{X} = \{x_{k,i}\}$, where $x_{k,i}$ denotes the number of individuals at location i ($1 \le i \le n$) who can be characterized using combination k ($1 \le k \le m$) of attribute values. For example, the data in Fig. 1a can be written equivalently as an 8×2 matrix $\mathbf{X}$ (Fig. 1b).

Many privacy laws and regulations restrict the publication of individual-level data with small area locations, because this type of data includes a full set of attributes that, when combined with the small area locations, can still be used to reveal the identities of individuals. In our example, person 1 who is counted in the first element of $\mathbf{X}$ ($x_{1,1} = 1$) can be uniquely identified using a combination of attribute values <Age<18, White, Non-Hispanic> together with the block location, which leads to a disclosure. The probability of disclosure for an individual, also referred to as the *individual disclosure risk*, is determined by the value of the matrix element that counts the individual. For individuals counted in matrix elements of ones (e.g., $x_{1,1}$), the individual disclosure risk is at its maximum, or one. As the value of a matrix element increases, the corresponding individual disclosure risk decreases. For example, each of the persons counted in elements of fours, such as $x_{1,2}$, has a decreased individual disclosure risk of one in four.

Individual-level data is often aggregated to reduce the disclosure risk for public use. There are two typical types of aggregation. The first is spatial aggregation, which combines multiple small areas into a single large unit. In our example, spatial aggregation can be accomplished by combining blocks 1 and 2 into a single geographic unit. In this way, there will be five individuals sharing the same attribute values <Age<18, White, Non-Hispanic> in the new unit, and the individual disclosure risk for person 1 is reduced from one to one in five. However, spatial aggregation can limit data availability for applications such as retail site selection that require small area information.

This limitation of spatial aggregation can be addressed, at least partly, by the second type of aggregation, attribute aggregation, which reduces the number of attributes for each individual. Attribute aggregation is the focus of this article: by "aggregated data," we mean data that have undergone attribute aggregation. In our example, the individual-level data in Fig. 1a has three attributes (voting age, race, and ethnicity), and it can be aggregated to data in Fig. 2a that counts the number of individuals by two attributes (voting age and race). We use p ($p < m$) to denote the number of possible combinations of attribute values in the aggregated data, where each combination contains one value from each attribute used for aggregation (e.g., <Age<18, White>). The aggregated data can be represented using a $p \times n$ matrix $\mathbf{Y} = \{y_{k',i}\}$, where $y_{k',i}$ is the number of individuals at location i ($1 \le i \le n$) who can be characterized using combination

Block	<Age < 18, White>	<Age ≥ 18, White>	<Age < 18, Black>	<Age ≥ 18, Black>
1	2	3	0	1
2	4	0	0	0

(a)

	Block 1	Block 2
1: <Age < 18, White>	2	4
2: <Age ≥ 18, White>	3	0
3: <Age < 18, Black>	0	0
4: <Age ≥ 18, Black>	1	0

(b)

Figure 2. Census data at the aggregate level. (a) Aggregated data that counts the number of individuals by voting age and race in each block. There are four possible combinations, each with one value from sex and race, respectively ($p = 4$). (b) A 4×2 matrix $\mathbf{Y}$ that represents the aggregated data in (a). (a) Table form; (b) Matrix form.

k' ($1 \leq k' \leq p$) of attribute values. For example, the aggregated data in Fig. 2a can be written equivalently as the matrix $\mathbf{Y}$ in Fig. 2b. In our example, after attribute aggregation, the individual disclosure risk for person 1 (first row in Fig. 1a) is reduced from one to one in two, while the block-level location for each individual is preserved. It should be noted that, while the individual disclosure risk may decrease for some persons when data is aggregated, it may remain high for others. In our example, individuals may still be uniquely identified using information from aggregated data and have an individual disclosure risk of one (e.g., the individual counted in $y_{4,1}$).

To protect individual privacy for both individual-level and aggregated data, we need to determine the individuals with high risks of disclosure. We define a λ-element as a non-zero matrix element with its value smaller than or equal to an integer λ. For the example in Fig. 1b, all elements of one are 1-elements, and all elements of one and two are 2-elements. Each individual counted in a λ-element has an individual disclosure risk of at least one in λ. When a small λ value is set, individuals counted in the λ-elements typically have high individual disclosure risks and therefore need to be protected. We protect privacy by transforming the individual-level data $\mathbf{X}$ so that, when a small λ value is set, individuals counted in the λ-elements (at-risk individuals) are assigned to other locations (Fig. 3). In theory, privacy protection can be achieved by transforming both individual-level and aggregated data. We only transform the individual-level data here because aggregation using the transformed data does not produce new λ-elements that require further protection, and doing so can ensure consistency across different sets of data aggregated from the same transformed individual-level data.

We use a *transition probability*, or the probability of assigning an individual from the original location to a candidate location, to indicate how at-risk individuals will be relocated. The primary rationale for utilizing transition probabilities instead of a binary policy of relocating or not relocating an at-risk individual is that, by basing the actual relocation on probabilities, randomness is introduced to mitigate the risks associated with reverse engineering the original data. The candidate location to be assigned from the original location can be either the original or a new location. The sum of the probabilities of assigning an individual to all the candidate locations is one. When the candidate location is the original location, it means that the at-risk individuals will not be relocated (e.g., arrow A in Fig. 3). High probabilities of assigning at-risk individuals to new locations (e.g., arrow B in Fig. 3) generally mean strong privacy protection.

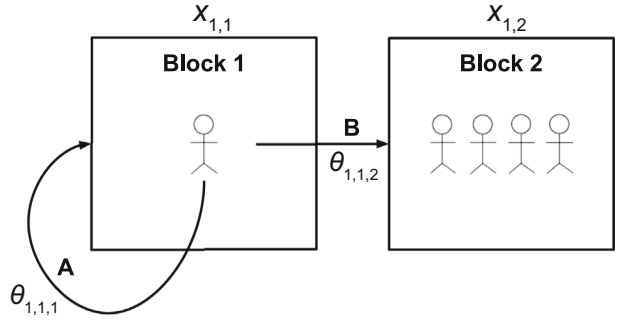

Figure 3. Relocating at-risk individuals to protect their privacy. The individual in block 1 counted in the first row of **X** in Fig. 1b ($x_{1,1}$) is considered at risk. This individual could be assigned to one of two locations, block 2 or the original location, as represented by two arrows, and the transition probabilities ($\theta_{1,1,1}$ and $\theta_{1,1,2}$) are used to indicate the likelihood of each assignment.

However, assigning individuals to new locations introduces noise in the data. As individuals are assigned to new locations, the number of λ-elements decreases, while the values of other elements in the same rows of **X** increase. For example, if the individual counted in $x_{1,1}$ in Fig. 1b is assigned from block 1 to block 2, values in the first row of **X** will change from $[1, 4]$ to $[0, 5]$. This introduces noise and reduces the utility of the data.

We aim to find optimal transition probabilities to the PUT problem. Each at-risk individual has a set of transition probabilities for being assigned to different candidate locations. The transition probabilities for all at-risk individuals form a solution to the PUT problem. Each solution in this context is evaluated by two objectives: level of privacy protection and data utility after relocations (we will formally define these two objectives in the next section). All of the solutions can therefore be placed on a two-dimensional coordinate system, with the two axes being privacy protection and data utility, respectively. These solutions together form a solution space, in which each point represents a unique solution. Given the conflicting objectives, Fig. 4 provides an illustration of such a solution space, where all solutions are bounded by the axes and the curve on the upper-right side of the shaded area. No solutions can exist outside the solution space. The upper-right curve is known as the Pareto front (Pareto 1971), and solutions on this front are optimal because they all have at least one, or in our case, exactly one objective function value that is higher than any other solution. Each solution that is not on the front has both objective function values smaller than at least one solution on the front, and is thus referred to as a dominated or inferior solution. For this reason, optimal solutions are also called nondominated solutions. For the example in Fig. 4, solutions A through D are nondominated, while solution E is dominated by B.

Methods

Multiobjective optimization is concerned with finding optimal solutions to decision-making problems with multiple conflicting objectives. We propose a modeling framework based on multiobjective optimization for identifying optimal solutions to the PUT problem and evaluating the solutions for decision-making (Fig. 5). This framework begins by assessing the risks of disclosure for the original data, which can be at either individual or aggregate levels. If the

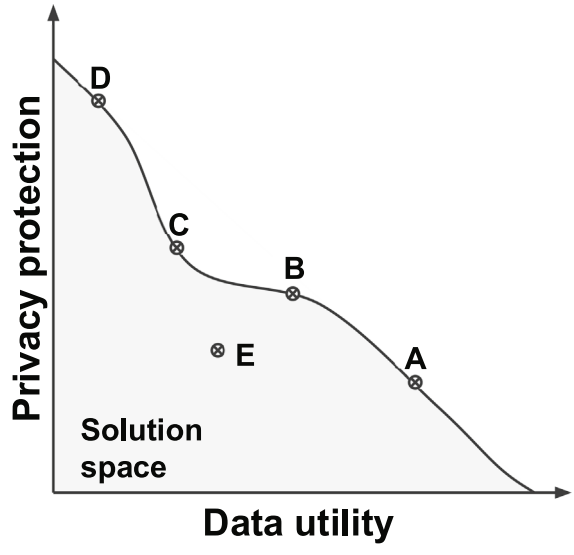

Figure 4. Solution space and the Pareto front to a PUT problem.

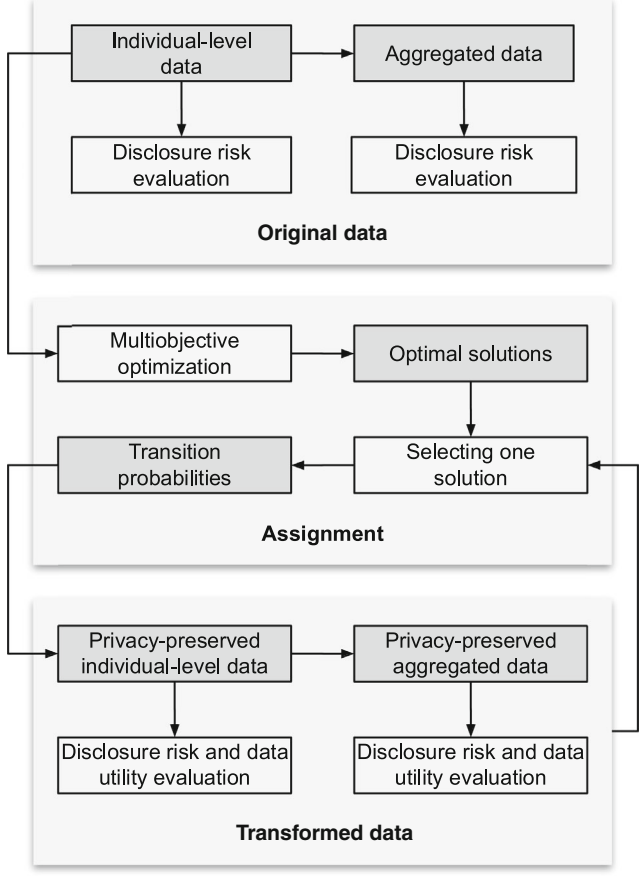

Figure 5. A flowchart of the proposed modeling framework.

original disclosure risks are deemed to be high, the PUT problem will be formulated as a multiobjective optimization problem to generate multiple nondominated or optimal solutions. Decision-makers and stakeholders will then select one of these solutions for further evaluation after negotiating based on their privacy and utility preferences. The transition probabilities from the selected solution will be used to determine how to transform the data by relocating at-risk individuals. By evaluating the disclosure risk and utility of the transformed data, decision-makers and stakeholders can determine whether to use the selected solution for the official data release or explore other options from the set of optimal solutions that better align with their priorities.

Multiobjective optimization

A multiobjective optimization model is formulated to determine the optimal transition probabilities. Several parameters are required as input by the user. First, the value of λ for λ-elements needs to be specified to determine whether an individual counted in $\mathbf{X}$ is at risk and should be relocated. The second parameter is the capacity of each candidate location, or the maximum number of individuals that can be assigned to a candidate location from other locations. Limiting the capacity for each candidate location can prevent overcrowding and excessive noise that may arise from assigning too many people to a single location. The capacity of each candidate location can be set differently based on its population size. For the purpose of demonstration, we use a capacity of 20 for each candidate location in this article. Third, it is necessary to decide the locations to which at-risk individuals can be relocated. A candidate location is said to cover an individual if it is different from his or her original location and is feasible for this individual to be assigned to. We define in this article that at-risk individuals counted in each element $x_{k,i}$ are covered by a candidate location j that is not the original location i, has no at-risk individuals, and has a nonzero count in the kth row of $\mathbf{X}$ (i.e., $x_{k,j} > \lambda$). Other principles to define the coverage are possible, such as those considering the spatial proximity between at-risk individuals and candidate locations, as introduced in the work by Lin and Xiao (2023b). The input parameters are summarized below.

$\lambda =$ integer used to determine theλ-elements in $\mathbf{X}$,

$m =$ number of rows in $\mathbf{X}$,

$n =$ number of columns in $\mathbf{X}$,

$K =$ set of all combinations of attribute values (rows of $\mathbf{X}$; indexed by k),

$I_k =$ set of locations for individuals counted in λ-elements of row k in $\mathbf{X}$ (indexed by i),

$J =$ set of all candidate locations (indexed by j),

$r_j =$ capacity of j,

$$t_{k,i,j} = \begin{cases} 1 & \text{if location } j \text{ covers the individuals counted in row } k \text{ of } \mathbf{X} \text{ at location } i, \\ 0 & \text{otherwise.} \end{cases}$$

The decision variables are a set of transition probabilities, each denoted as:

$\theta_{k,i,j} =$ transition probability of assigning an individual counted in row k of $\mathbf{X}$ (i.e., the kth combination of attribute values) from location i to j.

Objective functions

Two linear objective functions are developed to formulate the multiobjective optimization problem as a linear problem. The first objective is to maximize the sum of at-risk individuals assigned to new locations, where each at-risk individual receives a weight so that those with high individual disclosure risks can be relocated with a high priority. The expected sum of at-risk individuals being relocated is calculated as

$$\sum_{k\in K}\sum_{j\in J}\sum_{i\in I_k} t_{k,i,j}\theta_{k,i,j}x_{k,i}. \tag{1}$$

To prioritize the relocation of people with high individual disclosure risks (i.e., people counted in small elements of **X**), we define a weight as the inverse of a super-linear (e.g., polynomial or exponential) function of $x_{k,i}$ that grows faster than a linear function. In this way, assigning individuals counted in a small element of **X** is preferred over assigning those counted in a large element. As we maximize the weighted sum of at-risk individuals assigned to new locations, those counted in small elements will be relocated before those counted in large elements. Let $w_{k,i}$ denote the weight of any at-risk individual included in element $x_{k,i}$. We formally define the first objective function as the expected weighted sum of at-risk individuals assigned to new candidate locations:

$$P = \sum_{k\in K}\sum_{j\in J}\sum_{i\in I_k} w_{k,i}t_{k,i,j}\theta_{k,i,j}x_{k,i}. \tag{2}$$

Table 1 illustrates the effects of different types of weight $w_{k,i}$ using our example in Fig. 1b. We consider individuals counted in 3-elements to be at risk and relocate them to protect their privacy. Two example solutions are presented. In the first solution, individuals counted in both $x_{1,1}$ and $x_{2,1}$ have a probability of 0.5 to be assigned to new locations (i.e., $\theta_{1,1,1} = \theta_{2,1,1} = 0.5$, $\theta_{1,1,2} = \theta_{2,1,2} = 0.5$), and individuals counted in $x_{3,1}$ and $x_{8,1}$ remain at their original locations (i.e., $\theta_{3,1,1} = \theta_{8,1,1} = 1$, $\theta_{3,1,2} = \theta_{8,1,2} = 0$). In the second solution, the individual counted in $x_{1,1}$ has a probability of 0.9 to be assigned to a new location (i.e., $\theta_{1,1,1} = 0.1$, $\theta_{1,1,2} = 0.9$), while those counted in $x_{2,1}$ have a probability of 0.1 to be assigned to new locations (i.e., $\theta_{2,1,1} = 0.9$, $\theta_{2,1,2} = 0.1$). Individuals counted in $x_{3,1}$ and $x_{8,1}$ still remain at their original locations in the second solution. To prioritize the relocation of people with high individual disclosure risks, the person counted in $x_{1,1} = 1$ should be relocated before those counted in $x_{2,1} = 3$. We should therefore favor the second solution over the first. Of the five types of weight that we examine, the ones that yield the desired result are all inverses of super-linear functions ($1/x_{k,i}^2$, $1/x_{k,i}^3$, and

Table 1. The Privacy Protection Measures (P) for Different Types of Weight $w_{k,i}$.

		Privacy protection (P)	
$w_{k,i}$		Solution 1	Solution 2
Constant	1	2	1.2
Inverse-linear	$1/x_{k,i}$	1	1
Inverse-quadratic	$1/x_{k,i}^2$	0.67	0.93
Inverse-cubic	$1/x_{k,i}^3$	0.56	0.91
Inverse-exponential	$1/e^{x_{k,i}}$	0.26	0.35

Note: We assume $t_{k,1,1} = 0, t_{k,1,2} = 1\forall k$.

$1/e^{x_{k,i}}$). Without loss of generality, in this article, we illustrate the use of the inverse-quadratic weight for each individual counted in element $x_{k,i}$:

$$w_{k,i} = \frac{1}{x_{k,i}^2}. \tag{3}$$

In practical applications of the model, any types of inverse super-linear functions can be utilized, and they yield equivalent transition probabilities $\theta_{k,i,j}$ for relocating individuals.

The second objective of the optimization problem is to maximize the utility of transformed data. We first define the inverse concept of utility – noise – as the ratio of the absolute difference between the original and transformed data to the original data. When one at-risk individual with combination k of attribute values is moved out of location i, the absolute difference between the ith row in the original data and transformed data is one, and thus the ratio is $\frac{1}{x_{k,i}}$. Meanwhile, this individual is moved to location j, resulting in an absolute change of one at row k and column j of $\mathbf{X}$, and the ratio of the change is therefore $\frac{1}{x_{k,j}}$. We always assume that an individual can only be moved to a nonzero element, meaning $x_{k,j} > 0$. The total amount of noise as the result of assigning one individual counted in the kth row of $\mathbf{X}$ from location i to j, denoted as $e_{k,i,j}$, is the sum of the two ratios:

$$e_{k,i,j} = \frac{1}{x_{k,i}} + \frac{1}{x_{k,j}}. \tag{4}$$

The expected noise introduced by relocating all individuals counted in the kth row of $\mathbf{X}$ from location i to j can be derived by multiplying the expected sum of relocated individuals ($t_{k,i,j}\theta_{k,i,j}x_{k,i}$) by the amount of noise resulting from relocating one individual ($e_{k,i,j}$), and its sum is the total expected noise introduced to $\mathbf{X}$:

$$E = \sum_{k\in K}\sum_{j\in J}\sum_{i\in I_k} e_{k,i,j}t_{k,i,j}\theta_{k,i,j}x_{k,i}. \tag{5}$$

We define the expected utility of transformed data as one minus the noise E averaged over all matrix elements, which serves as the second objective function:

$$U = 1 - \frac{E}{m \times n}, \tag{6}$$

The value of U has a range of zero to one, with one indicating that the transformed data is identical to the original data and therefore has the highest utility, and zero indicating the opposite.

Model formulation

We formulate the multiobjective optimization problem based on two well-known spatial optimization problems: assignment problem (Munkres 1957) and covering problem (Daskin 2013), with each transition probability $\theta_{k,i,j}$ being a decision variable to be determined:

$$\max \quad P = \sum_{k\in K}\sum_{j\in J}\sum_{i\in I_k} w_{k,i}t_{k,i,j}\theta_{k,i,j}x_{k,i}\ , \tag{7}$$

$$\max \quad U = 1 - \frac{1}{m \times n}\sum_{k\in K}\sum_{j\in J}\sum_{i\in I_k} e_{k,i,j}t_{k,i,j}\theta_{k,i,j}x_{k,i}\ , \tag{8}$$

$$\text{subject to} \quad \theta_{k,i,i} + \sum_{j\in J} t_{k,i,j}\theta_{k,i,j} = 1 \quad \forall k, i\ , \tag{9}$$

$$t_{k,i,j} - \theta_{k,i,j} \geq 0 \quad \forall k, i, j \neq i \ , \tag{10}$$

$$\sum_{k \in K} \sum_{i \in I_k} t_{k,i,j} \theta_{k,i,j} x_{k,i} \leq r_j \quad \forall j \ , \tag{11}$$

$$0 \leq \theta_{k,i,j} \leq 1 \quad \forall k, j, i \ . \tag{12}$$

Objective 7 maximizes the expected weighted sum of at-risk individuals assigned to new candidate locations, whereas objective 8 maximizes the expected utility of transformed data. Constraints 9 state that each at-risk individual must either remain at the original location or be assigned to a candidate location that covers the person. Constraints 10 ensure that at-risk individuals cannot be relocated to new candidate locations that do not cover them. Constraints 11 ensure that the expected sum of at-risk individuals assigned to each candidate location does not exceed its capacity. Constraints 12 define the range of decision variables.

Disclosure risk evaluation

Disclosure risk and utility of the transformed data are evaluated after deriving multiple optimal solutions using the optimization model. The objective function value U (equation (6)) can be used to evaluate data utility, because it has a definite range of zero to one and can be compared across different privacy-preserved data sets. However, to formulate a linear optimization problem in order to efficiently find optimal solutions, as well as to prioritize the relocation of people with high individual disclosure risks, the objective function value P (equation (2)) is computed with a user-defined weight $w_{k,i}$ and may not be appropriate for use as a well-defined measure of privacy protection or disclosure risks. We therefore develop two measures for the evaluation of disclosure risks. Specifically, we here show how to evaluate disclosure risks for individual-level data. Similar methods used to evaluate the aggregated data are discussed in Appendix A.1.

We begin by representing the transformed individual-level data using an $m \times n$ matrix $\tilde{\mathbf{X}} = \{\tilde{x}_{k,i}\}$, where $\tilde{x}_{k,i}$ is the expected number of individuals at location i ($1 \leq i \leq n$) who can be characterized using the kth ($1 \leq k \leq m$) combination of attribute values. Based on transition probabilities $\theta_{k,i,j}$, the value of $\tilde{x}_{k,i}$ can be calculated as:

$$\tilde{x}_{k,i} = \sum_{j=1}^{n} \theta_{k,j,i} x_{k,j}. \tag{13}$$

For any person counted in element $\tilde{x}_{k,i}$, we compute the individual disclosure risk as a probability suggested by Shlomo and Skinner (2010); Shlomo (2014):

$$p_{k,i} = \frac{\theta_{k,i,i}}{\tilde{x}_{k,i}}. \tag{14}$$

The probability of disclosure for the transformed data $\tilde{\mathbf{X}}$, rather than per individual, is referred to as the *global disclosure risk*, which is denoted as τ and can be calculated as the average of the individual disclosure risks ($p_{k,i}$) over all its elements:

$$\tau = \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{i=1}^{n} p_{k,i}. \tag{15}$$

We use the global disclosure risk as a risk measure because it represents the average level of disclosure risk per person across the data set. The global disclosure risk for the original data $\mathbf{X}$, referred to as the baseline global disclosure risk, can be estimated by setting $\theta_{k,i,i} = 1$ and $\tilde{x}_{k,i} = x_{k,i}$ (i.e., no individuals are assigned to new candidate locations):

$$\tau^* = \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{i=1}^{n} \frac{1}{x_{k,i}}. \tag{16}$$

Another risk measure used in this article is the probability of finding a true 1-element, namely a matrix element of one before and after transformation (Bethlehem, Keller, and Pannekoek 1990; Fienberg and Makov 1998; Dale and Elliot 2001; Skinner and Elliot 2002; Lin and Xiao 2023a). We refer to this risk measure as the *population uniqueness rate*. For the transformed individual-level data $\tilde{\mathbf{X}}$, the population uniqueness rate (ϕ) is estimated as

$$\phi = \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{I}(\theta_{k,i,i} = 1 \text{ç} \tilde{x}_{k,i} = 1), \tag{17}$$

where $\mathbb{I}(\cdot)$ is the indicator function which equals one when the input condition is true, and zero otherwise. For the original data $\mathbf{X}$, we always have $\theta_{k,i,i} = 1$, and $\tilde{\mathbf{X}}$ is identical to the original data $\mathbf{X}$ (i.e., $\tilde{x}_{k,i} = x_{k,i}$). The corresponding population uniqueness rate, referred to as the baseline population uniqueness rate, can be calculated as follows:

$$\phi^* = \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{I}(x_{k,i} = 1). \tag{18}$$

Computational experiments

A set of computational experiments is performed to examine the effectiveness of the proposed modeling framework and to analyze the privacy-utility tradeoffs. The framework requires the input of an individual-level data set that covers the entire population of a region, but such data is typically not publicly available due to individual privacy protection. A synthetic individual-level data set is therefore used for our computational experiments (Lin and Xiao 2022, 2023c). Specifically, this data set contains 1,163,414 individuals in 284 census tracts of Franklin County, Ohio, and each individual has three attributes: voting age (V), ethnicity (E), and race (R). We generate this data set based on the 2010 U.S. Summary File 1 (SF1), which contains the data aggregated from 100% individual census responses from Franklin County residents. The synthetic data are compared to the original SF1 data as well as an external data set called the American Community Survey Public Use Microdata Sample (ACS PUMS). The results show that the synthetic data is consistent with both the SF1 and the ACS PUMS, with correlation coefficients of 1 and 0.99, respectively.

Three data matrices are created based on the synthetic population data. The first matrix, also called VER, is the individual-level data matrix (i.e., $\mathbf{X}$ in Fig. 1b). Each element in the matrix represents the number of individuals in each census tract that share the same values of voting age, ethnicity, and race. The individual-level data are aggregated into two sets of data, which are represented by two data matrices called ER and R, respectively. Each element in ER represents the population count in each census tract with the same values of ethnicity and race, and each element in R represents the population count in each census tract with the same race. We calculate

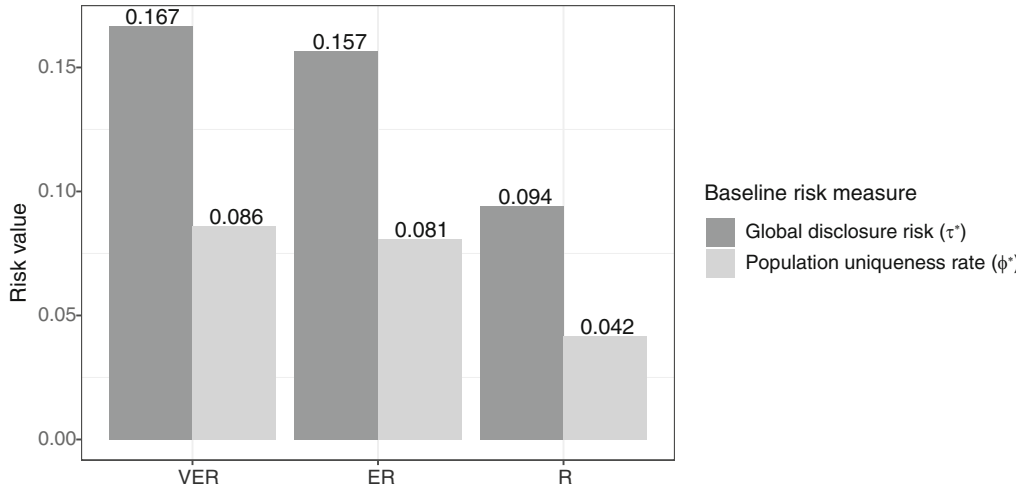

Figure 6. Baseline risk measures for individual-level (VER) and aggregated (ER and R) data.

the baseline risk measures for VER, ER, and R, as shown in Fig. 6. Specifically, VER has a baseline global disclosure risk (τ^*) of 0.167, which means that releasing the individual-level data is likely to disclose the identities of more than 16% of the population in Franklin County. Even for data that is aggregated, such as ER and R, the values of τ^* can still reach 0.154 and 0.094, which are close to or exceed 0.1. The baseline population uniqueness rates (ϕ^*), which measure the portion of the population at the highest risk of disclosure, are lower compared to τ^*. However, considering Franklin County's large population, over 800 individuals can be counted in the 1-elements of VER and ER and can be uniquely identified if releasing the data. Because of the relatively high risk measures observed in the three data sets (VER, ER, and R), it is necessary to conduct privacy protection so that at-risk individuals will be relocated to different locations. The transformed VER, ER, and R after privacy protection are denoted as $\tilde{\text{VER}}$, $\tilde{\text{ER}}$, and $\tilde{\text{R}}$, respectively.

The ε-constraint method (Haimes, Lasdon, and Wismer 1971) is used to solve the multiobjective optimization problem and obtain a set of optimal solutions to transform our data for privacy protection. The use of the ε-constraint method consists of two steps. The first step is to determine the range of each objective function (P and U). Given a λ value, P reaches its maximum value while U has its minimum value using the transition probabilities derived from optimizing a single objective of max P. In this case, all at-risk individuals are assigned to new locations, which provides maximum privacy protection while leading to minimum data utility. When no individuals are assigned to new locations, we have $\theta_{k,i,i} = 1$ and $\theta_{k,i,j} = 0, \forall i \neq j$. In this case, P has its minimum value while U reaches its maximum value (i.e., minimum protection but maximum utility). The ranges of P and U for varying λ values are presented in Table 2.

The second step of the ε-constraint method is to transform the multiobjective optimization problem into a series of single objective optimization problems. Specifically, we retain P as the objective function and add U as a constraint that specifies $U \geq \varepsilon$, where ε is a constant that falls within the range of U. Given a λ, the value of ε is systematically changed from the minimum to the maximum value of U using

$$\varepsilon = U_{\max} - q(U_{\max} - U_{\min}), \tag{19}$$

Table 2. Ranges of P and U for λ Values of One, Two, and Three

		Range	
λ	Objective functions	Min	Max
1	P	0	684
	U	0.91	1
2	P	0	921.5
	U	0.83	1
3	P	0	1,034.17
	U	0.78	1

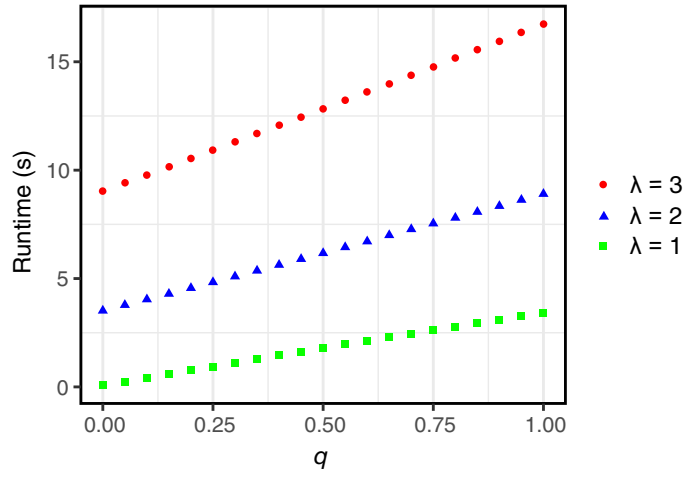

Figure 7. Cumulative runtime to obtain optimal solutions using the ϵ-constraint method.

where $U_{\min}$ and $U_{\max}$ are the minimum and maximum values of U, and q changes from zero to one with an increase of 0.05. The 21 different values of ϵ lead to 21 single objective optimization problems. We solve these optimization problems using a linear programming solver called Gurobi (Gurobi Optimization, LLC 2021). The solutions to these single objective optimization problems are optimal solutions to the original multiobjective optimization problem. Fig. 7 shows the cumulative runtime for solving the 21 optimization problems on a computer with AMD Ryzen 5 5600X 6-Core Processor (3.70 GHz) and 32GB RAM. The runtime increases as the value of λ increases. However, the runtime to obtain all 21 optimal solutions is less than 20s for each λ value tested.

Tradeoffs between privacy and utility

Fig. 8 illustrates the 21 optimal solutions for each λ value. There is an obvious tradeoff between privacy and utility when λ is fixed: increasing the privacy protection (P) clearly leads to a decrease in data utility (U) because of the increasing noise introduced into the data. The tradeoff is also demonstrated by the observation that as the value of λ increases, the maximum value of P increases, but at the expense of a reduction in the minimum value of U. It is also observed that the curves representing the Pareto fronts for different λ values overlap with each other. For example, the curve formed by red dots for $\lambda = 3$, which stretches from 0 to 921.5 in P

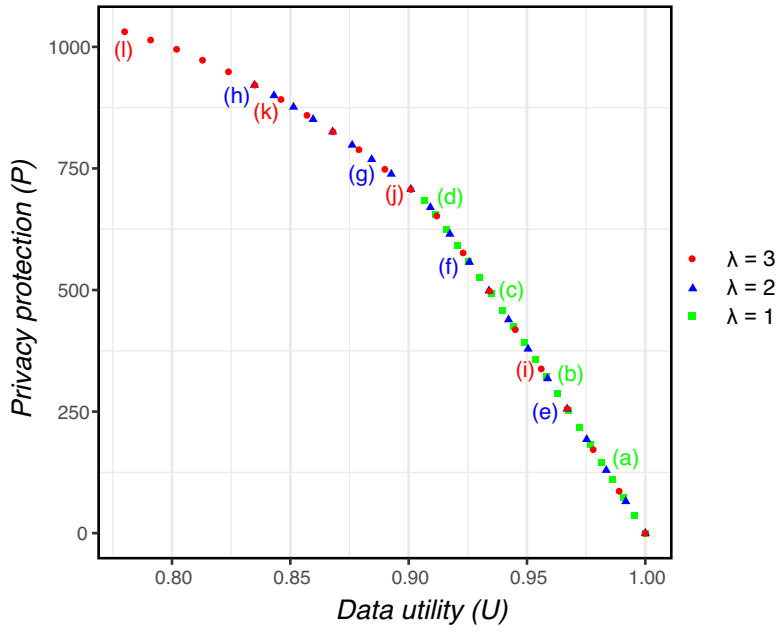

Figure 8. Optimal solutions to the PUT problem. The labeled solutions correspond to the subplots in Fig. 11.

and from 0.83 to 1 in U, overlaps with the blue curve for $\lambda = 2$. This overlap is due to the practice of prioritizing the relocation of at-risk individuals using the weight $w_{k,i}$ in equation (2), which results in individuals included in smaller element counts being relocated before those included in larger counts. When λ is set to three, individuals relocated before U reaches 0.83 and P reaches 921.5 all come from 2-elements, causing the curve for $\lambda = 3$ before U reaches 0.83 to overlap with the curve for $\lambda = 2$. When U falls below 0.83, individuals included in elements of three start to be relocated, resulting in an increase in the P value for $\lambda = 3$ after reaching 921.5.

Fig. 9 presents the risk and utility measures for the transformed individual-level (VẼR) and aggregated (ẼR and R̃) data. The global disclosure risk (τ) and population uniqueness rate (ϕ) for all the transformed data are lower than the corresponding baseline risk measures (τ^* and ϕ^*). when U is smaller than one ($\tau^* = \tau$ and $\phi^* = \phi$ when $U = 1$). However, data utility is subject to tradeoffs for reduced disclosure risks. When λ is fixed, the utility measure decreases as τ or ϕ decreases. As we increase the value of λ, the minimum value of τ decreases, at the expense of a decrease in the minimum value of U. This means that raising the λ value may allow us to further reduce the global disclosure risks because additional people will be relocated, but doing so may also lead to a further decrease in data utility. Nonetheless, the value of λ has no effect on the minimum ϕ that can be obtained, because the latter is only affected by the 1-elements and is always consistent with the minimum ϕ when λ is set to 1. We also find that the curves for different λ values overlap in each subplot of Fig. 9. This is consistent with the patterns found in Fig. 8 and indicates that individuals counted in small elements are relocated before those counted in large elements.

The multiobjective optimization model maximizes the number of at-risk individuals assigned to new locations instead of directly optimizing the risk measures defined in

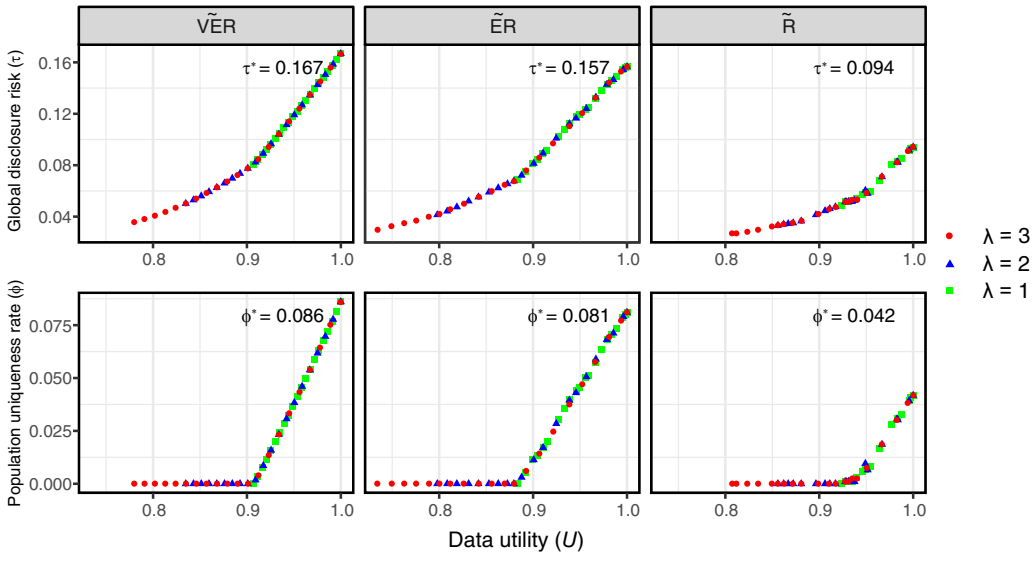

Figure 9. Risk and utility measures for the transformed individual-level (VẼR) and aggregated (ẼR and R̃) data.

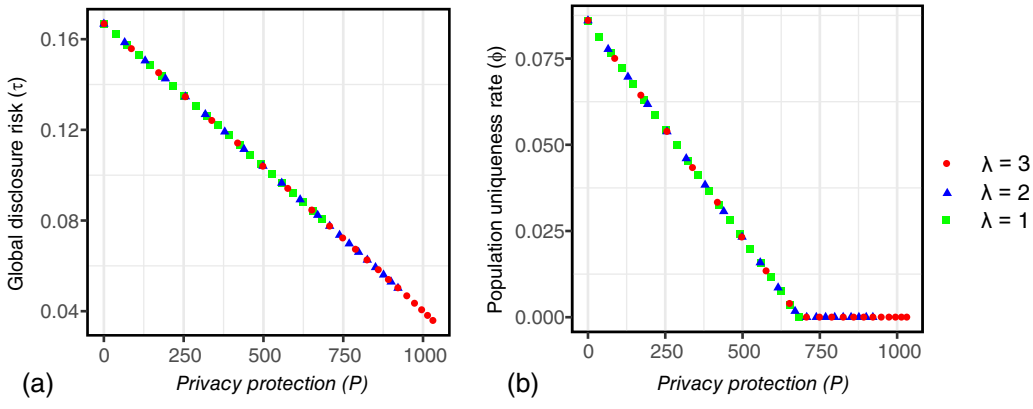

Figure 10. Relationships between the objective function P and the risk measures (τ and ϕ).

Section Disclosure Risk Evaluation. This is because we want to formulate a linear optimization problem in order to efficiently find optimal solutions. We here examine the relationship between the objective function P and the risk measures (τ and ϕ), as depicted in Fig. 10. All data points in Fig. 10a are on a decreasing straight line, implying that increasing the privacy protection (P) necessarily reduces the global disclosure risks. In Fig. 10b, the data points for various λ values lie on a decreasing straight line before P reaches 684, which is when all individuals counted in 1-elements are assigned. This implies that before finishing assigning individuals counted in 1-elements, the population uniqueness rate decreases as we increase the privacy protection as indicated by P. Overall, these results demonstrate that the objective function P can effectively represent the risk measures in formulating the multiobjective optimization problem.

Figure 11. Expected population change for each census tract with respect to the selected 12 optimal solutions. Each map, labeled as (a) through (l), corresponds to a solution labeled in Fig. 8. Each black line connects a tract that contains at-risk individuals to the candidate tract that receives the maximum expected number of individuals among all candidate tracts. We only show this particular candidate tract and all the lines that converge on it. (a) $\lambda = 1, q = 0.25$; (b) $\lambda = 1, q = 0.5$; (c) $\lambda = 1, q = 0.75$; (d) $\lambda = 1, q = 1$; (e) $\lambda = 2, q = 0.25$; (f) $\lambda = 2, q = 0.5$; (g) $\lambda = 2, q = 0.5$; (h) $\lambda = 2, q = 1$; (i) $\lambda = 3, q = 0.25$; (j) $\lambda = 3, q = 0.5$; (k) $\lambda = 3, q = 0.75$; (l) $\lambda = 3, q = 1$.

Effects on population totals

Assigning individuals to new locations can change the population totals in different locations, which is likely to affect the subsequent applications of data such as population-based policymaking. Two questions are investigated to explore the impact on population totals: how the change in population totals varies across space, and how to explain the variation? To answer these questions, we select a subset of the optimal solutions with different combinations of λ and q. Each q value is used to compute a value of ϵ (equation (19)), which yields a pair of P and U values on the Pareto front. Specifically, for each λ value, we choose from q values ranging from 0.25 to 1, with an interval of 0.25. Based on the selected solutions (also labeled in Fig. 8), we derive the transformed data and compute the expected population change with respect to the original data. Fig. 11 shows the census tracts with net population increase, net population decrease, and unchanged population. When λ is fixed (each row), increasing the value of q reduces the number of tracts with unchanged population. This is because increasing q reduces

Table 3. All Explanatory Variables and Their Summary Statistics

Variable	Min	Mean	Max
Percent of Whites (%)	4.20	67.28	100
Percent of Black or African Americans (%)	0	25.73	91.80
Percent of American Indians and Alaska Natives (AIANs) (%)	0	0.27	1.51
Percent of Asian Americans and Pacific Islanders (AAPIs) (%)	0	3.46	33.69
Percent of individuals with other races (%)	0	3.26	23.71
Percent of individuals aged below 18 (%)	0	24.01	43.43
Percent of individuals aged 18 and above (%)	56.57	77.18	100
Percent of Non-Hispanics (%)	69.97	96.78	100
Percent of Hispanics (%)	0	3.23	30.03
Total population	2	4,007	15,479

the lower bound of data utility, and therefore increases the number of at-risk individuals who are expected to be assigned. In addition, when we fix the value of q (each column), increasing the λ value reduces the number of tracts with an unchanged population, indicating an increase in the number of individuals assigned to different locations.

The effects of privacy protection on population totals vary across space, as illustrated in Fig. 11. Identifying the factors that determine the population change in each census tract is critical in order to understand the consequences of implementing different optimal solutions. Potential factors to determine the change include the demographic composition and population size of each census tract (Table 3), because they both may affect the number of at-risk individuals who require privacy protection. We use stepwise regression (Draper and Smith 1998) to investigate how these factors, or explanatory variables, affect the population change in each census tract with $\lambda = 3$ and $q = 1$ (Fig. 11l) that serves as the response variable. The goal of stepwise regression is to select explanatory variables to fit the best regression model that explains changes in a response variable. Starting with a null model with an intercept term, explanatory variables are added to the model one at a time, based on which variable is the most statistically significant, until no new variables are available. Following the addition of each new variable, all explanatory variables in the model are checked to see if they are significant and can be removed.

The best model resulting from stepwise regression contains four of the ten explanatory variables, as shown in Table 4. These variables are highly significant statistically, with p-values under 0.01, indicating their effectiveness in explaining the variations in population change. The percents of Whites and Non-Hispanics have negative coefficients, while the percent of American Indians and Alaska Natives (AIANs) has a positive coefficient. Because tracts with high percents of Whites and Non-Hispanics and low percents of AIANs are typically less diverse demographically, we can infer that tracts with low demographic diversity are prone to population decrease as a result of privacy protection. One possible explanation for the decrease is that in tracts with low demographic diversity, it can be difficult to have many individuals from minority populations (e.g., AIAN) who share the same demographics, which leads to their relocation to other locations for privacy protection that causes the population decrease. It is also observed that the total population has a positive coefficient, which suggests that tracts with small population sizes are also susceptible to population decline due to privacy protection. This may be because these tracts have a small number of individuals who share the same demographics, and some individuals may be relocated due to privacy protection.

Table 4. Explanatory Variables in the Best Model Resulted from Stepwise Regression

Variable	Coefficient	*P*-value
(Intercept)	−0.00	1.000
Percent of Whites	−0.16	0.002
Percent of AIANs	0.19	0.001
Percent of Non-Hispanics	−0.25	0.000
Total population	0.53	0.000

Discussion and conclusions

The privacy protection of census data often entails relocating individuals to different geographic locations, either directly or indirectly. For example, in the U.S. Census data releases from 1970 to 2010, the data swapping approach has been employed to exchange the location of at-risk individuals with those not considered at risk from a different geographic unit (Dalenius and Reiss 1982). In the 2020 Census, the Census Bureau applies a new disclosure avoidance method known as the TopDown Algorithm (TDA) that adds noise to census tables to protect privacy (Abowd et al. 2022). The TDA produces the same outcomes as relocating individuals by altering the population counts in each geographic unit while maintaining the total population of all geographic units. This article answers an essential question for methods that relocate people to protect privacy: what are the possible optimal ways of relocating individuals while considering data utility? Importantly, the multiobjective optimization approach proposed in this article has two significant advantages. The first is that it ensures the optimality of solutions to the PUT problem. In the practical release of census data, decision-makers would prefer an optimal solution that maximizes either privacy or utility when the other is held constant, rather than running into the risk of adopting a dominated or inferior solution in the solution space. Being able to explore possible optimal solutions will substantially aid the decision-making process. Another advantage of our method is that it can be used to efficiently generate different alternatives. Having more information is preferable to less when solving decision-making problems like the PUT problem. Accepting or rejecting a single optimal solution provided by methods such as the TDA may lead to an uninformed decision because it lacks knowledge of the full range of possibilities. Multiobjective analysis can address this issue by presenting a range of choices rather than only one of the optimums.

In this article, our approach to protecting at-risk individuals involves moving them from a less densely populated region to a more densely populated region. It may be argued that an alternative approach, which moves individuals from a more populated region to a less densely populated one, may also reduce the presence of at-risk individuals and contribute to privacy preservation. The reason why we did not pursue this alternative approach is that, when individuals are moved out of densely populated regions, those who remain in the originally densely populated areas may become the new at-risk individuals. This could potentially result in weaker privacy protection compared to the approach we propose in this study. Another potential concern about our approach is the possibility of reverse engineering the original data using the published data. If transition probabilities are not made public, reverse engineering is improbable, similar to the privacy protection approach of data swapping used in the 2010 U.S. Census. However, if, for transparency reasons, the transition probabilities are made public as part of the method's reproducible details, there could still be a probabilistic risk of reverse engineering. Further research could explore the implications of this. In addition, it should be noted that the outcomes

of our approach are dependent on how we define coverage, which specifies the locations to which at-risk individuals can be relocated. In this study, we base our coverage principle on the assumption that there exist geographic units with populations exceeding the k threshold set by users. If this assumption is not met, alternative principles can be considered. For example, one might relax the principle and allow at-risk individuals to be relocated to units with other at-risk individuals. This adjustment would not affect the use of our approach and could expand its applicability to other scenarios.

While this article does not include a formal comparison of the effectiveness of our method with other privacy-preserving techniques applied to complete-count census data, existing literature has conducted relevant assessments and comparisons using identical synthetic data and metrics. For example, Lin and Xiao (2023a) evaluate the population uniqueness rate for the TDA using synthetic data in Franklin County, which reveals that the TDA may introduce significant noise to the data without effectively minimizing population uniqueness. In another study, Lin and Xiao (2023b) compare the TDA to the Pareto front solution with maximum privacy protection. The results demonstrate that the latter can provide better privacy protection while introducing less noise compared to the former. These research endeavors offer valuable insights into the effectiveness of our optimization method compared to established approaches in census data privacy. Notably, the ability of our method to generate multiple effective and optimal solutions further distinguishes it as a promising approach. With that being said, we acknowledge the limited scope of comparisons in the current literature and recognize the necessity for further studies to enhance our understanding in this domain.

Concerns may arise that this approach is not differentially private and can subject to privacy risks. To address this concern, it is essential to recognize that the Census Bureau's adoption of differentially private TDA in 2020 was primarily driven by internal reconstruction experiments that showed the potential for combining multiple census tables to recover a substantial portion of the original individual-level data of the national population and identify these individuals (Abowd et al. 2022). Our approach, though not differentially private, relocates individuals using individual-level data and then generates aggregated data. This method ensures that even if someone were to attempt to reconstruct the individual-level data, it would not match the original data, thereby preventing reconstruction-abetted identification. In this vein, our approach can be used to effectively address the concerns raised in the reconstruction experiments. It should also be mentioned that various research indicates that TDA's strength in introducing excessive noise to the data to avoid reconstruction may be unnecessary (Muralidhar and Domingo-Ferrer 2023a, 2023b), and our research provides an alternative perspective on this matter.

The relocation of individuals for privacy protection, particularly in official statistics like census data, should undergo careful evaluation for potential impacts on future data applications. In this study, we investigate the effects of relocating individuals on the population total of each geographic unit. This is crucial for census data since these totals inform policy decisions such as congressional redistricting and federal funding allocation. The population shifts resulting from privacy protection may introduce bias in public policymaking, as evidenced by previous research on the impact of the TDA (Kenny et al. 2021; Cohen et al. 2022). It is therefore critical to involve a variety of stakeholders, rather than just technicians, in evaluating solutions that align with their priorities and demands to ensure that the final resolution to the PUT problem reflects the public interest. One limitation of our study is that we only examine the factors that affect population change for one optimal solution. In practical decision-making, analyzing all possible optimal solutions of interest would be useful to gain a better understanding of the impacts.

Computational efficiency is a practical issue in multiobjective programming, especially for large-scale problems. The computational experiments in this study use a tract-level data set as the test data, and it takes less than 20s to generate a set of 21 optimal solutions using the exact linear programming solver. If the data used is at a low geographic level such as the census block level, the runtime of the model is expected to increase significantly as the number of the decision variables grows. Existing literature has shown that heuristic methods are promising for efficiently finding high-quality solutions to large-scale multiobjective optimization problems (Reeves 1995). Among existing heuristics, genetic algorithms (GAs) have proven to be effective for solving problems with similar formulations to ours such as the covering problems (Xiao, Bennett, and Armstrong 2002; Tong, Murray, and Xiao 2009; Bao et al. 2015). The use of GAs will be explored in the future to ensure reasonable runtime for solving large problems.

Conflict of interest Statement

The authors report there are no competing interests to declare.

Data availability statement

The data that supports the findings of this study is openly available on GitHub at https://github.com/linyuehzzz/synthetic-populations.git.

Appendix A

The aggregated data matrix $\mathbf{Y}$ can be derived by premultiplying the individual-level data matrix $\mathbf{X}$ by a $p \times m$ query matrix $\mathbf{A} = \{a_{k',k}\}$, where $a_{k',k}$ equals one if the k'th ($1 \leq k' \leq p$) row of $\mathbf{Y}$ includes the counts in the kth row of $\mathbf{X}$ ($1 \leq k \leq m$). With the query matrix, the probability of assigning an individual counted in the k'-th row of $\mathbf{Y}$ from location i to j can be calculated as

$$\theta_{k'ij} = \frac{\sum_{k=1}^{m} a_{k',k} x_{k,i} \theta_{k,i,j}}{\sum_{k=1}^{m} a_{k',k} x_{k,i}}. \tag{A1}$$

The transformed aggregated data matrix can be represented using a $q \times n$ matrix $\tilde{\mathbf{Y}} = \{\tilde{y}_{k',i}\}$, where $\tilde{y}_{k',i}$ is the expected number of individuals at location i who can be characterized using a combination of attribute values k'. The value of $\tilde{y}_{k',i}$ can be calculated as

$$\tilde{y}_{k',i} = \sum_{k=1}^{m} a_{k',k} \tilde{x}_{k,i}. \tag{A2}$$

We derive the following risk measures for the aggregated data by replacing the matrix elements $x_{k,i}$ and $\tilde{x}_{k,i}$ with $y_{k',i}$ and $\tilde{y}_{k',i}$ as well as the transition probabilities $\theta_{k,i,j}$ with $\theta_{k',i,j}$ in equations (15)–(18):

$$\tau = \frac{1}{q \times n} \sum_{k'=1}^{q} \sum_{i=1}^{n} \frac{\theta_{k',i,i}}{\tilde{y}_{k',i}}, \tag{A3}$$

$$\tau^* = \frac{1}{q \times n} \sum_{k'=1}^{q} \sum_{i=1}^{n} \frac{1}{y_{k',i}}, \tag{A4}$$

$$\phi = \frac{1}{q \times n} \sum_{k'=1}^{q} \sum_{i=1}^{n} \mathbb{I}(\theta_{k',i,i} = 1 \text{ç} \tilde{y}_{k',i} = 1) \,, \tag{A5}$$

$$\phi^* = \frac{1}{q \times n} \sum_{k'=1}^{q} \sum_{i=1}^{n} \mathbb{I}(y_{k',i} = 1) \,. \tag{A6}$$

Here, τ is the global disclosure risk for the transformed aggregated data, τ^* is the baseline global disclosure risk for aggregated data, ϕ is the population uniqueness rate for the transformed aggregated data, and ϕ^* is the baseline population uniqueness rate for aggregated data.
The utility measure used for aggregated data is computed as

$$U = 1 - \frac{1}{q \times n} \sum_{k'=1}^{q} \sum_{i=1}^{n} \frac{\left| y_{k',i} - \tilde{y}_{k',i} \right|}{y_{k',i}}. \tag{A7}$$

References

Abowd, J. M. (2018). "The US Census Bureau Adopts Differential Privacy." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867. New York, NY: ACM.

Abowd, J. M., and M. B. Hawes. (2023). "Confidentiality Protection in the 2020 US Census of Population and Housing." *Annual Review of Statistics and Its Application* 10, 119–44.

Abowd, J. M., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, B. Moran, W. Sexton, M. Spence, and P. Zhuravlev. (2022). "The 2020 Census Disclosure Avoidance System TopDown Algorithm." *Harvard Data Science Review* (Special Issue 2). doi:10.1162/99608f92.529e3cb9.

Bao, S., N. Xiao, Z. Lai, H. Zhang, and C. Kim. (2015). "Optimizing Watchtower Locations for Forest Fire Monitoring Using Location Models." *Fire Safety Journal* 71, 100–9.

Bethlehem, J. G., W. J. Keller, and J. Pannekoek. (1990). "Disclosure Control of Microdata." *Journal of the American Statistical Association* 85(409), 38–45.

Branke, J., K. Deb, K. Miettinen, and R. Slowiński. (2008). *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Berlin, Germany: Springer Science & Business Media.

Brill, E. D., Jr. (1979). "The Use of Optimization Models in Public-Sector Planning." *Management Science* 25(5), 413–22.

Brill, E. D., Jr., S.-Y. Chang, and L. D. Hopkins. (1982). "Modeling to Generate Alternatives: The HSJ Approach and an Illustration Using a Problem in Land Use Planning." *Management Science* 28(3), 221–35.

Cohen, A., M. Duchin, J. Matthews, and B. Suwal. (2022). "Private Numbers in Public Policy: Census, Differential Privacy, and Redistricting." *Harvard Data Science Review* (Special Issue 2). doi:10.1162/99608f92.22fd8a0e.

Cohon, J. L. (2004). *Multiobjective Programming and Planning*. Mineola, NY: Dover Publications.

Dale, A., and M. Elliot. (2001). "Proposals for 2001 Samples of Anonymized Records: An Assessment of Disclosure Risk." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164(3), 427–47.

Dalenius, T., and S. P. Reiss. (1982). "Data-Swapping: A Technique for Disclosure Control." *Journal of Statistical Planning and Inference* 6(1), 73–85.

Daskin, M. S. (2013). "Covering Problems." In *Network and Discrete Location: Models, Algorithms, and Applications*, 2nd ed., 124–92, edited by M. S. Em Daskin. Hoboken, NJ: John Wiley & Sons, Ltd.

Deb, K. (2001). *Multi-Objective Optimization Using Evolutionary Algorithms*. New York, NY: John Wiley & Sons, Inc.

Donnelly, F. (2019). *Exploring the US Census: Your Guide to America's Data*. California: SAGE Publications.

Draper, N. R., and H. Smith. (1998). *Applied Regression Analysis*. New York, NY: John Wiley & Sons.

Duncan, G., S. Keller-McNulty, S. Stokes. (2004). Database security and confidentiality: examining disclosure risk versus data utility through the RU confidentiality map. Tech. Rep. 142, National Institute for Statistical Sciences, Research Triangle Park, NC.

Fienberg, S. E., and U. E. Makov. (1998). "Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data." *Journal of Official Statistics* 14(4), 385.

Gurobi Optimization, LLC. (2021). *Gurobi Optimizer Reference Manual*. https://www.gurobi.com.

Haimes, Y. Y., L. S. Lasdon, and D. A. Wismer. (1971). "On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1(3), 296–7.

Houfaf-Khoufaf, W., G. Touya, and A. Le Guilcher. (2021). "Geographically Masking Addresses to Study Covid-19 Clusters." *Cartography and Geographic Information Science*, 1–15. doi: 10.1080/15230406.2021.1977709

Kenny, C. T., S. Kuriwaki, C. McCartan, E. T. Rosenman, T. Simko, and K. Imai. (2021). "The Use of Differential Privacy for Census Data and its Impact on Redistricting: The Case of the 2020 US Census." *Science Advances* 7, eabk3283.

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.

Li, T., and N. Li. (2009). "On the Tradeoff between Privacy and Utility in Data Publishing." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 517–26. New York, NY: ACM.

Lin, A., and F. Harvey. (2015). "Spatial Variation of Privacy Measured through Individual Uniqueness Based on Simple us Demographics Data." In *Advances in Spatial Data Handling and Analysis*, 289–98. Cham, Switzerland: Springer.

Lin, Y. (2023). "Geo-Indistinguishable Masking: Enhancing Privacy Protection in Spatial Point Mapping." *Cartography and Geographic Information Science* 50(6), 608–23.

Lin, Y., and N. Xiao. (2022). "Developing Synthetic Individual-Level Population Datasets: The Case of Contextualizing Maps of Privacy-Preserving Census Data." AutoCarto 2022, the 24th International Research Symposium on Cartography and GIScience.

Lin, Y., and N. Xiao. (2023a). "Assessing the Impact of Differential Privacy on Population Uniques in Geographically Aggregated Data: The Case of the 2020 u.s. Census." *Population Research and Policy Review* 42(5), 81.

Lin, Y., and N. Xiao. (2023b). "A Computational Framework for Preserving Privacy and Maintaining Utility of Geographically Aggregated Data: A Stochastic Spatial Optimization Approach." *Annals of the American Association of Geographers* 113(5), 1035–56.

Lin, Y., and N. Xiao. (2023c). "Generating Small Areal Synthetic Microdata from Public Aggregated Data Using an Optimization Method." *The Professional Geographer* 75(6), 905–15.

Munkres, J. (1957). "Algorithms for the Assignment and Transportation Problems." *Journal of the Society for Industrial and Applied Mathematics* 5(1), 32–8.

Muralidhar, K., and J. Domingo-Ferrer. (2023a). "Database Reconstruction Is Not So Easy and Is Different from Reidentification." *Journal of Official Statistics* 39(3), 381–98.

Muralidhar, K., and J. Domingo-Ferrer. (2023b). "A Rejoinder to Garfinkel (2023)–Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial us Census: Still a Viable Option." *Journal of Official Statistics (JOS)* 39(3), 411–20.

Pareto, V. (1971). In *Manual of Political Economy*, edited by A. Montesano, A. Zanni, L. Bruni, J. S. Chipman, and M. McLure. Oxford, UK: Oxford University Press.

Polzin, F., and O. Kounadi. (2021). "Adaptive Voronoi Masking: A Method to Protect Confidential Discrete Spatial Data." In *11th International Conference on Geographic Information Science (Giscience 2021) - Part II*, 1:1–1:17, edited by K. Janowicz and J. A. Verstegen. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Purdam, K., and M. Elliot. (2007). "A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records." *Environment and Planning A* 39(5), 1101–18.

Reeves, C. (1995). *Modern Heuristics Techniques for Combinatorial Problems*. New York, NY: McGraw-Hill.

Richter, W. (2018). "The Verified Neighbor Approach to Geoprivacy: An Improved Method for Geographic Masking." *Journal of exposure science & environmental epidemiology* 28(2), 109–18.

Ruggles, S., and D. L. Magnuson. (2023). ""it's None of their Damn Business": Privacy and Disclosure Control in the us Census, 1790–2020." *Population and Development Review* 49(3), 651–79.

Seidl, D. E., P. Jankowski, and M.-H. Tsou. (2016). "Privacy and Spatial Pattern Preservation in Masked Gps Trajectory Data." *International Journal of Geographical Information Science* 30(4), 785–800.

Seidl, D. E., P. Jankowski, and K. C. Clarke. (2018). "Privacy and False Identification Risk in Geomasking Techniques." *Geographical Analysis* 50(3), 280–97.

Shlomo, N. (2014). "Probabilistic Record Linkage for Disclosure Risk Assessment." *International Conference on Privacy in Statistical Databases*, 8744, 269–82.

Shlomo, N., and C. Skinner. (2010). "Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata." *The Annals of Applied Statistics* 4(3), 1291–310.

Shlomo, N., C. Tudor, and P. Groom. (2010). In *Privacy in Statistical Databases*. *PSD 2010*. Lecture Notes in Computer Science Vol 6344, 41–51, edited by J. Domingo-Ferrer and E. Magkos. Berlin: Springer.

Skinner, C., and M. J. Elliot. (2002). "A Measure of Disclosure Risk for Microdata." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(4), 855–67.

Sweeney, L. Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000. https://dataprivacylab.org/projects/identifiability/paper1.pdf

Tong, D., A. Murray, and N. Xiao. (2009). "Heuristics in Spatial Analysis: A Genetic Algorithm for Coverage Maximization." *Annals of the Association of American Geographers* 99(4), 698–711.

Veregin, H. (1999). "Data Quality Parameters." In *Geographical Information Systems, Principles and Technical Issues*, 177–89, edited by P. Longley and M. Goodchild. New York, NY: John Wiley & Sons, Inc.

Walker, K. (2023). *Analyzing us Census Data: Methods, Maps, and Models in R*. Boca Raton, FL: CRC Press.

Wang, J., J. Kim, and M.-P. Kwan. (2022). "An Exploratory Assessment of the Effectiveness of Geomasking Methods on Privacy Protection and Analytical Accuracy for Individual-Level Geospatial Data." *Cartography and Geographic Information Science* 49(5), 385–406.

Wieland, S. C., C. A. Cassa, K. D. Mandl, and B. Berger. (2008). "Revealing the Spatial Distribution of a Disease while Preserving Privacy." *Proceedings of the National Academy of Sciences* 105(46), 17608–13.

Xiao, N., D. A. Bennett, and M. P. Armstrong. (2002). "Using Evolutionary Algorithms to Generate Alternatives for Multiobjective Site-Search Problems." *Environment and Planning A* 34(4), 639–56.

Xiao, N., D. A. Bennett, and M. P. Armstrong. (2007). "Interactive Evolutionary Approaches to Multiobjective Spatial Decision Making: A Synthetic Review." *Computers, Environment and Urban Systems* 31(3), 232–52.

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), 1257–74.