Published in final edited form as:

Stat Med. 2014 May 20; 33(11): 1928–1945. doi:10.1002/sim.6078.

Imputation of confidential data sets with spatial locations using disease mapping models

Thais Paiva^a, Avishek Chakraborty^b, Jerry Reiter^{a,*}, and Alan Gelfand^a

^aDepartment of Statistical Science, Duke University, Durham, NC, USA

^bDepartment of Statistics, Texas A&M University, College Station, TX, USA

Abstract

Data that include fine geographic information, such as census tract or street block identifiers, can be difficult to release as public use files. Fine geography provides information that ill-intentioned data users can use to identify individuals. We propose to release data with simulated geographies, so as to enable spatial analyses while reducing disclosure risks. We fit disease mapping models that predict areal-level counts from attributes in the file, and sample new locations based on the estimated models. We illustrate this approach using data on causes of death in North Carolina, including evaluations of the disclosure risks and analytic validity that can result from releasing synthetic geographies.

Keywords

Confidentiality; disclosure; geography; imputation; synthetic

1. Introduction

Many government agencies, research centers, and principal investigators collect data that they intend to disseminate broadly. These organizations, henceforth all called agencies, often are ethically and even legally obligated to protect data subjects' identities and sensitive attributes. This is particularly challenging when agencies seek to include fine-scale geographic variables on the files. For example, including exact addresses could enable ill-intentioned users to match names to addresses using public records, thereby revealing data subjects' identities. Even modestly coarse geography, like street block or census tract of residence, can be risky when demographic or other readily available attributes are on the file, which when combined may result in identifications.

To reduce disclosures, most agencies aggregate geographies to high levels before sharing data, if they share geography at all [1]. For example, agencies following the safe harbor provisions of the U. S. Health Insurance Portability and Accountability Act (HIPAA), which regulates sharing of personal health information, are required to release geographic units comprising at least 20,000 people. As another example, the U.S. Census Bureau does not

release geographic identifiers below aggregates of at least 100,000 people in public use files of census data. While such aggregation preserves analyses at the level of aggregation, it can disable small area estimation, mask local spatial dependencies, and create ecological inference fallacies at lower levels of aggregation. Other strategies for protecting geography include adding random noise to locations (e.g., [2, 3]) or swapping individuals across locations (e.g., [4, 5]).

An alternative framework for protecting geographies was proposed by Wang and Reiter [6]: replace actual locations with locations simulated from statistical models. Specifically, Wang and Reiter [6] treat the precise latitude and longitude of each location as a bivariate outcome to be predicted from the other attributes on the file. After fitting a prediction model—regression trees in their illustrative example—they generate new, replacement locations for each individual on the file. To account for the uncertainty introduced by simulation and thereby enable estimation of variances, they recommend that agencies generate $m > 1$ versions of the data sets for dissemination. Such data sets can protect confidentiality, since identification of units and their sensitive data can be difficult when the geographies in the released data are not actual, collected values. And, when the simulation models faithfully reflect the relationships in the collected data, the shared data can preserve spatial associations, avoid ecological inference problems, and facilitate small area estimation. A related approach was used by Machanavajjhala *et al.* [7], who use multinomial regressions to synthesize the street blocks where people live conditional on the street blocks where they work and other block-level attributes.

The approach in [6] requires that the agency knows the latitude and longitude of each location. These may not be available, at least not immediately and without additional cost for geocoding. Further, in many settings, the spatial distribution of attributes can be multimodal and complex, so that it is difficult to identify good-fitting bivariate regression approaches. Motivated by these limitations, and with a goal of accurately modeling the spatial distribution of locations, we propose to use areal level spatial models, often referred to as disease mapping models [8, 9, 10, 11], as engines for generating simulated locations. The basic idea is to (i) tile the spatial surface in ways intended to ensure adequate confidentiality protection, (ii) estimate disease mapping models that predict observed, areal-level counts from attributes on the file, and (iii) use the estimated models to sample multiple, new locations for each individual based on its attribute pattern. This approach applies most naturally for areal geographies like census tracts or street blocks, but it also can be applied with finer-grain coordinates like point locations after an initial aggregation.

We focus exclusively on methods for altering geography, leaving attributes at their original values. We note, however, that agencies might decide instead or in addition to alter the attributes on the file to strengthen the confidentiality protection [12, 13, 14]. As examples, Zhou *et al.* [15] use spatial smoothing to mask non-geographic attributes in a Medicare database, leaving original locations unperturbed; and, the Census Bureau swaps the attribute data for individuals in neighboring areas when creating the public use microdata files for the decennial census. Such methods could be applied after the generation of synthetic geographies; see [6] for further discussion.

The remainder of the article is organized as follows. In Section 2, we present the areal spatial modeling approach for generating synthetic geography. In Section 3, we describe several metrics for assessing the disclosure risks in the released synthetic data sets. We also review how one obtains point and interval estimates from such data sets. In Section 4, we illustrate the approach by generating multiply-imputed, partially synthetic versions of a spatially-referenced data set describing causes of death in Durham, North Carolina. Finally, in Section 5, we conclude with discussion of implementation of the approach.

2. Areal Spatial Models for Data Synthesis

To provide context for the approach, we introduce a scenario that motivated our investigations. Suppose a state public health agency seeks to release counts of lung cancer incidence by sex, race, and age (categorized) for each street block in the state. The agency owns the appropriate data, but it cannot release the exact counts in the blocks because of confidentiality promises.

More formally, let $D = (S, X)$ comprise data on n individuals, where $S = (s_1, \dots, s_n)$ includes each individual's location and X is the $n \times p$ matrix of each individual's non-spatial attributes. As in the motivating scenario, let the p attributes $(X_1, \dots, X_p)$ be discrete-valued. For $k = 1, \dots, p$, let d_k be the number of levels in X_k . Without loss of generality, assume the values in each $X_k = (1, \dots, d_k)$. Let $b = 1, \dots, B$ index each distinct attribute pattern in X , where $B \leq \prod_{k=1}^p d_k$. For each (b, k) , let $x_k^{(b)}$ be the value of X_k in b .

We seek to model where people with certain demographic patterns are likely to reside. To do so, we assume that the area of interest can be divided into a grid comprising G cells. The grid cells may comprise pre-existing areal units, such as collections of census tracts or street blocks. Alternatively, for point-resolved geography, they may be imposed by the agency for reasons related to computational convenience and, as we shall discuss in Section 3, reduction of confidentiality disclosure risks. Let each grid cell be indexed with $i = 1, \dots, G$. For each (i, b) , let $c_i^{(b)}$ be the number of observations in cell i with attributes b . For $k = 1, \dots, p$, let $Z_k^{(b)}$ be a $d_k \times 1$ vector comprising a one at position $x_k^{(b)}$ and zeros elsewhere. We propose to estimate spatial models of the form,

$$\begin{aligned} c_i^{(b)} &\sim \text{Poisson}(\lambda_i^{(b)}) \\ \log \lambda_i^{(b)} &= \mu + \sum_{k=1}^p \alpha_k' Z_k^{(b)} + \theta_i + \sum_{k=1}^p \phi_{ik}' Z_k^{(b)} + \epsilon_i^{(b)}. \end{aligned} \quad (1)$$

Here, μ is the overall intercept; each $\alpha_k = (\alpha_{k1}, \dots, \alpha_{kd_k})$ is a $d_k \times 1$ vector of main effects for attribute k ; θ_i is a grid-specific spatial effect; and each $\phi_{ik} = (\phi_{ik1}, \dots, \phi_{ikd_k})$ is a $d_k \times 1$ vector of grid-specific spatial effects for attribute k . For identifiability, we set each α_{k1} and each ϕ_{ik1} equal to zero. We note that one also can include interactions among the attributes, as well as flexible functions of (grid-level) continuous attributes. The spatial effects allow $\lambda_i^{(b)}$ to vary by grid cell and attribute pattern. The $\epsilon_i^{(b)}$ is an error term that allows for additional flexibility in the modeling. The model implies that, within any grid cell, the spatial intensities are assumed homogeneous over all points in that cell; in Section 5 we

suggest related approaches based on point process modeling that do not make this assumption. We note that the model in (1) is akin to the areal-level spatial ANOVA model in [16].

To induce spatial correlation among neighboring grid cells, we use the intrinsic CAR model [17] as the prior distribution for $\theta = (\theta_1, \dots, \theta_G)$. Specifically, we assume for all i that

$$\theta_i | \theta_{-i} \sim N \left(\bar{\theta}_i, \sigma_\theta^2 / n_i \right). \quad (2)$$

Here, θ_{-i} includes the values of θ_j for all $j \neq i$, and $\bar{\theta}_i$ is the average of the n_i values of θ_j for cells j that are neighbors of cell i . We define neighbors to be grid cells that share vertices. Using an analogous notation, for $\{ikj : i = 1, \dots, G, k = 1, \dots, p, j = 2, \dots, d_k\}$, we assume

$$\phi_{ikj} | \phi_{-i,kj} \sim N \left(\bar{\phi}_{ikj}, \sigma_{\phi_{kj}}^2 / n_i \right). \quad (3)$$

Following [17], to ensure identifiability we constrain the elements of θ and ϕ_{kj} so that

$$\sum_{i=1}^G \theta_i = 0 \text{ and } \sum_{i=1}^G \phi_{ikj} = 0 \text{ for all } (kj).$$

We include $\epsilon_i^{(b)}$ to capture residual variation in the Poisson rates that is not explained by the covariates, thereby adding flexibility to the modeling. We assume that

$$\epsilon_i^{(b)} \sim N \left(0, \sigma_\epsilon^2 \right). \quad (4)$$

For prior distributions on non-zero hyperparameters, we use

$$\mu \sim N \left(0, v_\mu \right) \quad (5)$$

$$\alpha_{kj} \sim N \left(0, v_{\alpha_k} \right) \quad \text{for all } (kj) \quad (6)$$

$$1/\sigma_\theta^2 \sim \text{Gamma} \left(a_\theta, b_\theta \right) \quad (7)$$

$$1/\sigma_{\phi_{kj}}^2 \sim \text{Gamma} \left(a_{\phi_k}, b_{\phi_k} \right) \quad \text{for all } (kj) \quad (8)$$

$$1/\sigma_\epsilon^2 \sim \text{Gamma} \left(a_\epsilon, b_\epsilon \right). \quad (9)$$

Here, we recommend setting all variances $v_{\dots}$ to large values, e.g., around 100, to represent vague prior information. We recommend typical vague prior specifications for $1/\sigma_\theta^2$ and each $1/\sigma_{\phi_k}^2$, for example setting $(a_\theta = b_\theta = a_{\phi_k} = b_{\phi_k}) = .1$ for all (k, b) . In the application described in Section 4.1, results were not sensitive to other common hyperparameter choices (in particular, $a_\theta = a_{\phi_k} = 2$ and $b_\theta = b_{\phi_k} = 1$). For $1/\sigma_\epsilon^2$, we recommend following the suggestions of Banerjee *et al.* [17, eq. 5.48] when specifying (a_ϵ, b_ϵ) . They advise to

account for the difference in the precision dimensions between the pure error term and the CAR model when setting these hyperparameters. We present an example of this specification in Section 4.

Posterior distributions of the parameters of this model, and hence of the λ s, can be estimated via Markov Chain Monte Carlo algorithms. In the applications in Section 4, we sample from full conditionals not in closed form using adaptive rejection sampling [18]. Details of the algorithm are available from the first author.

After estimating the posterior distributions of $\lambda = \{\lambda_i^{(b)}\}$, the agency can generate synthetic locations for the n individuals on the file. To begin, the agency takes a single draw of λ , say $\lambda^{(l)}$, from its posterior distribution. For all (i, b) , the agency computes

$$p_i^{(lb)} = \lambda_i^{(lb)} / \sum_{i=1}^G \lambda_i^{(lb)}. \quad (10)$$

For each individual with attribute pattern b , the agency randomly and independently samples its grid cell according to the probabilities in $(p_1^{(lb)}, \dots, p_G^{(lb)})$. When convenient, the sampled grid cells can serve as one set of synthetic locations. Alternatively, the agency can sample finer coordinates from inside the grid cell, for example sampling uniformly from feasible geographic locations (e.g., capable of being residences) inside the cell. The result is one set of synthetic locations, $\tilde{S}^{(l)} = (\tilde{s}_1^{(l)}, \dots, \tilde{s}_n^{(l)})$, which when attached to X results in one partially synthetic data set, $\tilde{D}^{(l)} = (\tilde{S}^{(l)}, X)$. The agency repeats the process independently m times to obtain m sets of synthetic locations, $\tilde{S} = (\tilde{S}^{(1)}, \dots, \tilde{S}^{(m)})$, and corresponding data sets, $\tilde{D} = (\tilde{D}^{(1)}, \dots, \tilde{D}^{(m)})$, which are released to the public. In practice, to obtain approximately independent realizations of the synthesis process, the agency either (i) can run m MCMC chains initiated at dispersed starting values and use the final draw of λ for each chain, or (ii) run one long MCMC chain thinned so that autocorrelations among estimated components of λ are approximately zero.

It is worth noting that two records close in space in the original data will not necessarily be close in space in the synthetic data, since their synthetic locations (grid cells and coordinates) are independently generated from the estimated Poisson models. Arguably, such possible movement is necessary to reduce disclosure risks sufficiently.

A key feature of this modeling is the choice of the grid size. Intuitively, a very thin grid (with a large number of grid cells) allows for greater heterogeneity in the tiled rate surface. When such heterogeneity is an important feature in the data, this should allow the pattern of synthetic geographies to mimic the observed pattern more faithfully. However, a very thin grid also could result in synthetic locations that are too close to the original ones, which could fail to protect confidentiality. Conversely, a very coarse grid tends to improve protection at cost of reduced data quality. This suggests that agencies can benefit from examining trade offs in disclosure risk and data quality for multiple candidate grid sizes. This requires quantifying disclosure risks and data quality, which we now discuss.

3. Disclosure Risk and Data Utility

In this section, we describe an approach to assessing disclosure risks of the partially synthetic data with simulated geographies. We focus on computing the probabilities that individuals' true areal geographies can be learned from the released data, building on ideas developed by Duncan and Lambert [19] and applied subsequently by several authors (e.g., [20, 21, 22, 23]). We use functions of these probabilities to create risk metrics for both attribute disclosure (an intruder learns the value of true areal geography) and identification disclosure (an intruder learns the identity of some record).

We also review the inferential methods for partially synthetic data [24]. These methods enable comparisons of inferences based on the original and synthetic data, which represent our primary approach to evaluate the utility of the synthetic data.

3.1. Risk Measures

To learn geographies, we assume that the intruder utilizes all information at her or his disposal. This includes information released about the synthetic data model, which we denote with M . For example, M could include mathematical descriptions corresponding to (1) – (9), including the definitions of the grid cells. Alternatively, M could include the code used to fit the models without parameter estimates (releasing parameter estimates could leak too much information about S). The intruder also may possess auxiliary information about the geographies of the records on the file, which we denote with A . For example, A could include the geographies of some subset of individuals on the file, or it could be empty.

Using this information, the intruder seeks to determine the probable values of s_t for some record t . We assume the intruder knows that t is in the sample (but does not know its location). This assumption can be relaxed; see Drechsler and Reiter [22] for a general strategy to do so. The intruder need not affiliate a particular row in $\tilde{D}$ with t to determine probable values of s_t ; indeed, for patterns b such that $\sum_i c_i^{(b)} > 1$ unique affiliation is impossible. Thus, for any particular t and potential location s , the intruder seeks to estimate

$$\begin{aligned} \rho_t^s &= P(s_t = s | \tilde{D}, A, M) = c P(\tilde{S} | s_t = s, X, A, M) P(s_t = s | X, A, M) \\ &= c \left(\int P(\tilde{S} | s_t = s, X, A, M, \lambda) P(\lambda | s_t = s, X, A, M) d\lambda \right) P(s_t = s | X, A, M) \end{aligned} \quad (11)$$

over all feasible s , where c is a normalizing constant. For any record with a unique attribute pattern, i.e., a b such that $\sum_i c_i^{(b)} = 1$, $\{\rho_t^s\}$ represents the posterior distribution of s_t for that particular record. The interpretation of ρ_t^s is more subtle for records with b such that $\sum_i c_i^{(b)} > 1$ and depends on the nature of A . For example, if an intruder knows the locations of all records in the sample with attribute pattern b except t , then $\{\rho_t^s\}$ again represents the posterior distribution of s_t for one particular record. With other forms of A , intruders may be able to interpret $\{\rho_t^s\}$ only as a distribution for all records with attribute pattern b .

We assume that the intruder selects the s yielding the maximum ρ_t^s as a best guess for s_t . Arguably the most that an intruder can learn from $\tilde{D}$ (beyond X) is the grid cell to which the

individual belongs, since finer-grain locations within any cell are randomly sampled within the cell. Hence, we suppose the intruder's goal in computing (11) is to find the correct grid cell. Thus, we let S and $\tilde{S}$ in (11) comprise grid cells.

Conceptually, $P(s_t = s | X, A, M)$ represents the intruder's prior beliefs about the grid cell of individual t , and $\tilde{S}$ serves to sharpen those beliefs. Effectively, the intruder takes guesses at the true grid cell of individual t according to the prior beliefs. Guesses that result in relatively low probability of generating $\tilde{S}$ (given X , A , and M) are downweighted compared to guesses that result in relatively high probability of generating $\tilde{S}$. By evaluating the probabilities of generating $\tilde{S}$ over all possible s , the intruder determines the *a posteriori* best estimate.

Of course, it is impossible for agencies to know any particular intruder's prior beliefs. Instead, agencies can adopt the recommendation of Skinner [25] and evaluate risks under reasonable prior distributions. For example, the agency can use a uniform distribution over all grid cells in the population that include individuals with the same attribute pattern b as individual t . This reflects vague prior knowledge about s_t . In the absence of population counts by attribute patterns per grid cell, the agency can allow the support to include the entire grid.

Similarly, it is impossible for the agency to know the auxiliary information possessed by intruders. One approach, which we adopt here, is to evaluate risks under a “worst case” scenario by assuming that the intruder knows the grid cells of all individuals except one particular t , i.e., the intruder knows $s_{t'}$ for all $t' \neq t$. Call this set S_{-t} . In addition to offering risk estimates for intruders with very strong prior knowledge, setting $A = S_{-t}$ greatly facilitates computation as we describe in Section 3.1.1.

3.1.1. Computational Methods—The form of (11) when $A = S_{-t}$ suggests a Monte Carlo approach to estimation of ρ_t^s . First, for any proposed value s , the agency replaces s_t with s to form a new set of locations, $S_t^s = (s_t = s, S_{-t})$, attached to original X . Second, treating (S_t^s, X) as if it were the collected data, the agency samples many values of λ . Third, for each sampled λ , the agency computes the probability of generating the released $\tilde{S}$, and averages these probabilities. The agency repeats these three steps for all values of s , which allows computation of the normalizing constant in (11) and hence ρ_t^s for all s .

To draw new λ s for each (S_t^s, X) , one approach is to re-estimate the model in (1) – (9). This is computationally intensive, however, as the agency needs to estimate G models per t . Instead, we suggest using the sampled values of λ from $p(\lambda/D)$ as proposals for an importance sampling algorithm [26, Chapter 3]. As a brief review of importance sampling, suppose we seek to estimate the expectation of some function $g(\lambda)$, where $\lambda \sim f(\lambda)$. Further suppose that we have available a sample $(\lambda^{(1)}, \dots, \lambda^{(L)})$ from a convenient distribution $f^*(\lambda)$ that differs from $f(\lambda)$. We can estimate $E_f(g(\lambda))$ using

$$E_f(g(\boldsymbol{\lambda})) \approx (1/L) \sum_{j=1}^L g(\boldsymbol{\lambda}^{(j)}) \frac{f(\boldsymbol{\lambda}^{(j)}) / f^*(\boldsymbol{\lambda}^{(j)})}{\sum_{j=1}^L f(\boldsymbol{\lambda}^{(j)}) / f^*(\boldsymbol{\lambda}^{(j)})}. \quad (12)$$

We note that (12) only requires that $f(\boldsymbol{\lambda})$ and $f^*(\boldsymbol{\lambda})$ be known up to normalizing constants.

We implement importance sampling algorithms to approximate the integral in (11). For any proposed $s_t = s$, we set $g(\boldsymbol{\lambda}) = cP(\tilde{S}|S_t^s, X, M)$ and seek to approximate its expectation with respect to $f(\boldsymbol{\lambda}) = P(\boldsymbol{\lambda}|\mathbf{S}_t^s, \mathbf{X}, \mathbf{M})$. To facilitate computation, we work with each set of synthetic locations $\tilde{S}^{(l)}$ separately, since

$$P(\tilde{S}|S_t^s, X, M) = \prod_{l=1}^m P(\tilde{S}^{(l)}|S_t^s, X, M). \quad (13)$$

Given a sampled value of $\boldsymbol{\lambda}$, we have

$$P(\tilde{S}^{(l)}|S_t^s, X, M, \boldsymbol{\lambda}) = \prod_{i=1}^G \prod_{b=1}^B (p_i^{(lb)})^{\tilde{c}_i^{(lb)}}, \quad (14)$$

where $\tilde{c}_i^{(lb)}$ is the count of synthetic points with attribute pattern b in grid cell i from set l , and $p_i^{(lb)}$ is computed as in (10) with the sampled $\boldsymbol{\lambda}$. We next set $(\boldsymbol{\lambda}^{(1)}, \dots, \boldsymbol{\lambda}^{(L)})$ equal to L draws of $\boldsymbol{\lambda}$ already available from the estimated posterior distribution based on D ; hence, we set $f^*(\boldsymbol{\lambda}) = f(\boldsymbol{\lambda}|S, X, M)$. Following (1) – (9), the only differences in the kernels of $f(\boldsymbol{\lambda})$ and $f^*(\boldsymbol{\lambda})$ include (i) the components of the likelihood associated with the counts on the grid cells s and s_t for attribute pattern b and (ii) the normalizing constants for each density. Hence, after computing the normalized ratio in (12), we are left with the expression,

$$P(\tilde{S}^{(l)}|S_t^s, X, M) \approx (1/L) \sum_{j=1}^L \left(\prod_{i=1}^G \prod_{b=1}^B (p_i^{(lb)})^{\tilde{c}_i^{(lb)}} \right) \left(\frac{\lambda_s^{(jb)} / \lambda_{s_t}^{(jb)}}{\sum_{h=1}^L \lambda_s^{(hb)} / \lambda_{s_t}^{(hb)}} \right). \quad (15)$$

We repeat this computation for $l = 1, \dots, m$ times, plugging the m results into (13). Finally, to approximate ρ_t^s , we compute (13) for each s and multiply each resulting value by $P(s_t = s | X, S_{-t}, M)$, and we normalize the collection of G results (hence, computation of c is never required). As a note on computation, the terms in (14) for which b does not match the attributes of record t cancel when normalizing, so that one can replace the expression in (14) with $\prod_{i=1}^G (\lambda_i^{(b)})^{\tilde{c}_i^{(lb)}}$.

Although the importance sampling uses the actual values of S to make proposals for $\boldsymbol{\lambda}$, any S^* could be used. Hence, intruders are able to utilize these approximations as well.

Setting $A = S_{-t}$ simplifies computation immensely, in that when computing ρ_t^s we have to impute new values only for s_t . In contrast, to compute ρ_t^s when $A = S_{-t}$, the intruder needs to

impute possible values for all unknown geographies. This introduces a potentially large number of computations. One case of particular interest is when A is empty, representing no intruder knowledge. To avoid imputing all of S , one rough approximation is to use each of the m sets of $\tilde{S}_{-t}^{(l)}$ as a draw of S_{-t} , and average the m resulting values of ρ_t^s .

3.1.2. Summary Measures—After obtaining the posterior probabilities, agencies need to summarize these probabilities to evaluate individual and file level disclosure risks. We now present four such risk measures. Each is based on the assumption that the intruder uses the cell s with maximum ρ_t^s as the best guess for s_t .

The first measure assesses the risks that intruders learn true grid cells given the synthetic data; hence, it is an attribute disclosure risk measure. For all $t = 1, \dots, n$ individuals in the file, let $r_t = 1$ if the maximum posterior probability for record t happens to be on the true s_t (with no ties), and let $r_t = 0$ otherwise. That is, for all t , let $r_t = 1_{\{arg \max_s (\rho_t^s) = s_t\}}$. A file level risk measure is the percentage of records with $r_t = 1$, i.e.,

$$R_{all} = \sum_{t=1}^n r_t / n. \quad (16)$$

Intuitively, smaller values of R_{all} are preferable to larger values for confidentiality protection.

As noted by many experts in disclosure estimation (e.g., [27]), agencies pay special attention to risks for records with unique combinations of variables in the sample (although arguably uniqueness in the population is more relevant). Singletons are more likely to be identified, since matches to external data are guaranteed to be correct (assuming no errors in matching). With this issue in mind, we introduce a measure that focuses on individuals with unique combinations of (i, b) . Formally, for all $t = 1, \dots, n$, let $c_{(t)}$ be the count of individuals in D matching the grid cell and attribute pattern of t . Let $a_t = 1$ if $c_{(t)} = 1$, and let $a_t = 0$ if $c_{(t)} > 1$. The second risk measure is the percentage of records with unique patterns that the intruder correctly locates,

$$R_{unq} = \frac{\left(\sum_{t=1}^n a_t r_t \right)}{\left(\sum_{t=1}^n a_t \right)}. \quad (17)$$

Both R_{all} and R_{unq} do not distinguish between intruders whose best guess is close (but not equal) to the actual grid cell and whose best guess is far from the actual grid cell. To distinguish these, we present a third risk measure based on distances. For each t , let d_t be the distance between s_t and the grid cell with the maximum probability, so that

$$d_t = \|s_t - arg \max_s (\rho_t^s)\|. \quad (18)$$

The agency can assess the distributions of d_t over all t to determine if, for example, distances tend not to be concentrated at small values. We compute d_t as the Euclidean distance between s_t and the centroid of the grid cell with maximum probability. Thus, when $r_t = 1$, d_t is bounded by half of the diagonal of a grid cell.

While R_{all} , R_{unq} , and the distribution of d_t summarize risks that intruders learn true geographies, they are not readily interpretable as measures of identification disclosure risk. In particular, in some grid cells many records have the same attribute pattern b , so that intruders cannot distinguish between them. For an extreme example, consider using only one grid cell for the entire area. Here, $r_t = 1$ for all t , since *a priori* everyone is guaranteed to be in the cell. Thus, $R_{all} = R_{unq} = 1$, and all d_t are equal. However, since coordinates are sampled randomly within the single cell, releasing $\hat{S}$ introduces zero risks that individual records will be identified (assuming the study area is already known to the intruder).

To quantify identification disclosure risk, we use a measure similar to one presented in [21]. For each t , we compute $z_t = r_t/c_{(t)}$. This corresponds to the probability that an intruder guesses correctly when randomly choosing a match from among the $c_{(t)}$ qualifying records with the same attribute pattern and grid cell as record t . A file level risk measure is

$$R_{id} = \sum_{i=1}^n z_t / n. \quad (19)$$

Thus, R_{id} can be considered the expected number of correct identifications when randomly choosing a match. Implicit in this interpretation is the assumption that the intruder knows that record t is among the n records collected in the original survey. Agencies can relax this assumption by replacing $c_{(t)}$ with the number of individuals in the population (not D) with the same attribute pattern and grid cell as record t . When this population count is unknown, as is often the case, the agency must estimate it, for example using log-linear models or other approaches [27, 28, 29].

3.2. Inferences with partially synthetic data

The inferential methods for partially synthetic data depend on the nature of the analysis, as we now describe. Let Q be a scalar estimand of interest, such as a population mean or regression coefficient. Suppose that, if given D , the analyst would use normal distributions for inference, $(Q - q) \sim N(0, u)$. Here, q is a point estimator of Q such as an unbiased estimator or posterior mean, and u is the associated variance. For each $l = 1, \dots, m$, let $q^{(l)}$ and $u^{(l)}$ be the estimates of q and u computed with $D^{(l)}$. For inferences about Q the analyst needs the following quantities:

$$\bar{q}_m = \sum_{l=1}^m q^{(l)} / m$$

$$b_m = \sum_{l=1}^m \left(q^{(l)} - \bar{q}_m \right)^2 / (m - 1)$$

$$\bar{u}_m = \sum_{l=1}^m u^{(l)} / m.$$

The analyst uses $\bar{q}_m$ as a point estimate of Q with associated variance $T_m = b_m/m + m$. Inferences are based on t -distributions, $(Q - \bar{q}_m) \sim t_{\nu}(0, T_m)$ with $\nu_m = (m - 1)(1 + m/b_m)^2$ degrees of freedom.

With spatial data, analysts often estimate spatial regression models with Bayesian methods [17, 30]. Here, the combining rules in [24] may not apply, particularly when the posterior distributions of parameters are not well-approximated by normal distributions. Instead, analysts can use the approach described by Zhou and Reiter [31]. Specifically, the analyst fits the Bayesian model in each synthetic data set, coming up with m sets of posterior samples of the parameters of interest. The analyst then mixes all m sets of draws to estimate the posterior distribution.

4. Illustrative Application

We now apply the methods of Section 2 and Section 3 to create and evaluate partially synthetic locations for a subset of North Carolina (NC) mortality records from 2002. Similar data were used by [6]. The data include precise longitudes and latitudes of deceased individuals' residences and several variables related to manner of death. As explained by Wang and Reiter [6], these data are publicly available and so do not require disclosure protection; however, because the data include point-referenced locations that can be revealed for comparisons, they represent an ideal testbed for methods that protect confidentiality of geographies. These data also enable us to illustrate how one can apply areal spatial models to protect confidentiality even with point-referenced data.

We selected a subset of individuals with residences in seven contiguous zip codes in Durham, NC, which is an area of approximately 20 by 20 miles. To mimic the types of variables discussed in the motivating example of Section 2, we include as attributes individuals' sex (male or female), race (black or white), age (<60, 60-75, 75-85, or >85 years), education (less than high school, high school, some college, more than 4 years of college) and an indicator if the cause of death was caused by cancer or some failure of the immune system versus all other causes. The final sample includes $n = 6294$ individuals. We include only two races (accounting for 97% of the observations in the full data) for convenience of illustration, as the spatial distribution of these two races is clearly evident in the data; see Figure 1(a). Similar exploratory maps indicate that variables other than race are more-or-less randomly distributed across the area.

With this set of variables, analysts might treat the indicator for reason of death, which we label as Y , as an outcome variable to be predicted from the other variables. Here, the analyst may want to control for spatial dependencies in the prediction model, for example to improve predictive power or to account for heterogeneity not captured by the predictors. However, since Y does not follow strong spatial patterns—see Figure 1(b)—it is not a particularly useful outcome for testing how well the synthetic data preserve spatial

dependencies. We therefore created a surrogate outcome, $\tilde{Y}$, for which location matters. In particular, for $t = 1, \dots, n$, we set

$$\tilde{Y}_t \sim \text{Bern}(\pi_t), \quad \text{logit}(\pi_t) = X_t \beta + w(s_t). \quad (20)$$

Here, X_t includes main effects for sex, race, age, and education. Each $\beta_k \sim N(0, .3)$, and $w(s_t)$ is a mean-zero Gaussian process [17] with exponential covariance function such that, for any two locations $s = s'$, $\text{Cov}(s, s') = \sigma^2 \exp(-\phi \|s - s'\|)$ with $\phi = .6$ and $\sigma^2 = 1$. As evident in Figure 1(c), this results in a stronger spatial pattern than Y . We use $D = (\tilde{Y}, X)$ for all subsequent analyses.

4.1. Generation of the synthetic data sets

We generate synthetic locations using the approach in Section 2 with all $B = 128$ attribute patterns formed from the variables in $D = (\tilde{Y}, X)$. For illustration, we use square grids with three sets of sizes: 10×10 , 20×20 , and 30×30 . The coordinates were rescaled to fall in $[0, 10] \times [0, 10]$, respecting the original proportion of the horizontal and vertical ranges.

Exploratory data analysis suggests that including only main effects in (1) – (9) is reasonable for these data. Regardless of grid size, we use the prior distributions in (5) – (9) with $\nu_\mu = \nu_{a_k} = 5$ for all attributes k ; with $(a_\theta, b_\theta) = (a_{\phi_k}, b_{\phi_k}) = (.1, .1)$ for all k ; and with $(a_\psi, b_\psi) = ((.1)(.7^2)\bar{n}, .1)$, where $\bar{n}$ is the average number of neighbors per grid cell. These values of (a_ψ, b_ψ) are suggested by [17, equation 5.48] to account for the difference in the dimensions of the spatial and non-spatial effects. We obtained similar results using $(a_\theta, b_\theta) = (a_{\phi_k}, b_{\phi_k}) = (2, 1)$ for all k with $(a_\psi, b_\psi) = (2, 1/[(.7^2)\bar{n}])$.

We run the MCMC for 10001 iterations, tossing the first 1000 as burn-in. We assess the convergence of the chain by analyzing the trace plots of the main effects of the attributes and the posterior mean surfaces of $\lambda^{(b)}$. Posterior intervals for all coefficients across the three grid sizes indicate that race and education are the strongest predictors of the Poisson rates; see the results in the online supplement. Figure 2 displays the posterior mean surfaces of λ for the 20×20 grid for two attribute patterns. The first corresponds to white women over age 85 with education less than high school and $\tilde{Y} = 0$; this has the highest frequency among the combinations. The second pattern corresponds to black men less than age 60 with more than four years of college and $\tilde{Y} = 1$; this pattern is far less frequent (only eight individuals).

After the burn-in, we sample $m = 10$ synthetic locations following the approach in Section 2, using a systematic sample of every one thousandth draw.

4.2. Evaluating the utility of the synthetic data sets

We evaluate the utility of the synthetic data by comparing various analyses on the original data (with $\tilde{Y}$) and synthetic data sets. These analyses include estimates of demographic characteristics by zip code, posterior inference from a spatial regression involving $\tilde{Y}$ on the remaining variables, and maps of synthetic locations by various demographic categories.

Figure 3 displays the proportions of black people in each of the seven zip codes for the three grid sizes. Here, we determine the 95% confidence intervals using the methods in [24], described in Section 3.2. With the 30×30 grid and to a slightly lesser extent the 20×20 grid, the confidence intervals from the synthetic data largely overlap with those based on D . The intervals for the 10×10 grid are not as high quality. Figure 4 displays analogous results for the proportion of cases with $\tilde{Y} = 1$. Once again, the intervals based on $\tilde{D}$ and D largely overlap, with generally increasing quality as the grid becomes thinner. For the three different grid sizes, the fraction of total variance (T_m) due to variability in the $m = 10$ synthetic point estimates ($b_m/10$) is typically around 15% (sd=6%). We note that, in both figures, the posterior means for the 30×30 grid occasionally are further from the original proportions than those from other grid sizes; this is due primarily to chance.

To evaluate finer spatial information in the synthetic locations, we next estimate the spatial logistic regression in (20) based on $\tilde{D}$. As vague prior distributions, we use

$$\beta \sim N(0, 100I) \quad (21)$$

$$\phi \sim \text{Uniform}(0.6, 2.9) \quad (22)$$

$$1/\sigma^2 \sim \text{Gamma}(2, 1) \quad (23)$$

The bounds of the prior distribution for the spatial decay parameter ϕ are defined based on the effective range for distances equal to 10% and 50% of the maximum distance between two points in the data ($\bar{d} = 10.5$). Using the relation between the range and ϕ with the exponential covariance function, we obtain the bounds as approximately $3/(\bar{d})$ and $3/(.1 \bar{d})$ [17].

We estimate the posterior distribution of all parameters using the spGLM function from the spBayes in R. Since n is relatively large, we fit the model using a predictive process [32, 33] with 100 knots to obtain posterior samples of the parameters. We estimate a separate MCMC chain for each synthetic data set, running each for 100000 iterations, and combine the resulting posterior draws to obtain synthetic data inferences. We also estimate the posterior distributions using D for comparison. The variability among the synthetic point estimates contributes on average 2.5% (sd=2%) to the total posterior variances of the parameters. We computed these proportions using b_m/m over the posterior variances.

Figure 5 displays the posterior mean and 95% central credible intervals for the coefficients and spatial covariance parameters. The credible intervals for the coefficients based on $\tilde{D}$ are very similar to those based on D . Additionally, the posterior distributions of ϕ based on $\tilde{D}$ are similar to those based on D across all grid sizes. The posterior mean of σ^2 when using the 10×10 grid is noticeably lower than the posterior mean when using D . This gap decreases as we increase the number of grid cells. We believe that the results with 20×20 and 30×30 are close enough to those based on D that many analysts would be comfortable interpreting this spatial regression based on $\tilde{Y}$.

Finally, Figures 6 and 7 display the distribution of race and $\tilde{Y}$ in the original and four randomly chosen synthetic data sets for the 20×20 grid. Results for the 30×30 grid are similar. The maps confirm the trends from the previous analyses: spatial patterns in these variables are maintained in the synthetic data sets. More detailed evidence of this is evident in Figure 8 and Figure 9, which display the original and synthetic points for the two patterns from Figure 2: white women over age 85 with education less than high school and $\tilde{Y} = 0$; and black men less than age 60 with more than four years of college and $\tilde{Y} = 1$. Overall, the spatial patterns are approximately preserved, as the points are spread around similar areas. Nonetheless, the synthetic locations still can differ from the original ones, as is necessary to reduce disclosure risks.

4.3. Evaluating the risk of the synthetic data sets

For all risk analyses, we adopt the “worst case” scenario assumption outlined in Section 3.1, so that for each t we set $A = S_{-t}$. Table 1 displays the values of $(R_{all}, R_{unq}, R_{id})$ for the generated $\tilde{D}$ at each grid size. Across all three grids, no more than 9% of all locations have correctly identified grid cells. As evident from R_{unq} , records with unique combinations of (i, b) have slightly higher risks of location disclosure. However, the values of R_{unq} indicate that roughly 90% of cases with unique patterns are not correctly located. The intruder has no way of determining which among these cases are correctly located. The values of R_{all} and R_{unq} are larger for the 10×10 grid than the 20×20 grid. As G gets smaller, the area of individual grid cells increases, so that the intruder has greater chances of guessing the true cells correctly.

Across all scenarios, the expected number of correct identifications is no more than 7.5%. As expected, R_{id} is largest for the 30×30 grid. The values of R_{id} for the 10×10 and 20×20 grids are nearly identical, suggesting that most of the identification disclosure risk with these grids comes from cases that have unique combinations of (i, b) across all grids. These R_{id} values are smaller than R_{id} for the 30×30 grid since larger areas contain fewer unique combinations of (i, b) .

Figure 10 displays histograms of d_t for all n individuals. In the 10×10 grid, a distance greater than 2.12 means that s_t is not one of the neighbors of the intruder’s best guess. This threshold is 1.06 and .71 for the 20×20 and 30×30 grids, respectively. As evident in Figure 10, in each scenario most d_t exceed these critical distances. Thus, the intruder’s guesses tend to be far from the true locations.

5. Concluding Remarks

The results from the synthesis of the Durham, NC, mortality data illustrate how tuning the grid size effectively trades disclosure risk for data quality. Given that the risk computations presumed an intruder with very, and perhaps unrealistically, detailed knowledge, we suspect that many agencies would be comfortable with the risks of releasing the synthetic data generated via the 20×20 or 30×30 grids. These releases were superior in data quality compared to the synthetic data generated via the 10×10 grid. Of course, agencies can and should evaluate the quality of additional representative statistical analyses when comparing

the risk-utility profiles of any proposed release, as well as consider (when sensible) multiple grids of differing sizes.

After generating synthetic locations, the agency still may deem the disclosure risks too large for some records. As suggested by a reviewer, *post hoc* agencies can smooth the synthetic location probabilities for each risky case over additional grid cells, and re-draw synthetic locations for those cases. Alternatively, agencies can use the location probabilities from coarser synthesis models for the risky cases. When the number of risky cases is small, either of these two approaches should not seriously degrade the quality of the synthetic data. Such *post hoc* changes in the synthesis model should be reflected (at least approximately) in the likelihood $P(\tilde{S}_t = s, X, A, M)$ when re-computing the risk measures.

We simulated locations only for Durham, NC, which comprised 6294 cases. Extending to much larger data sets, for example the entire state of NC, demands more efficient computational algorithms than those described and used here. One convenient strategy is to partition the data into geographical regions effectively modeled with manageable grid size, and simulate grid cells by running the synthesizer independently within each region. This has the added benefit of exactly preserving spatial analyses that use the regions as the finest level of geography. One also could tailor the grid size in each region to improve risk-utility profiles, for example using a coarser grid in regions where disclosure risks appear to be high and a finer grid in regions where risks appear to be low.

The synthetic data reflect only the relationships between geographies and attributes that are encoded in the synthesis models. Thus, non-geographic attributes not in the models may have distorted synthetic spatial distributions. Similar problems arise when adding non-geographic attributes from another database to the original file by means of matching the actual locations from the other database to the synthetic locations. The synthesized geographies are conditionally independent of the appended attributes (given the attributes in the original file), which may not mimic reality. On the other hand, synthetic geographies can potentially allow analysts to discern associations between the attributes in the original file and contextual variables affiliated with geographies (at the grid cell or coarser levels), such as the number of parks or grocery stores near a particular location, even if the contextual variables derive from external information. For example, suppose a disease occurs most often in a particular set of grid cells marked by an unusual feature, such as heavy polluters. When disease status is included in the model, the synthetic data should appropriately impute that set of locations for people with the disease, and hence connect the disease incidence to the locations of the polluters.

These issues suggest that agencies include in the synthesis model as many attributes that vary spatially (as indicated by exploratory data analyses and prior scientific knowledge) as possible while ensuring acceptable disclosure risks. The over-arching goal of model selection is to reflect the spatial relationships in the collected sample faithfully as opposed to, for example, minimizing out-of-sample prediction errors. When the number of potentially relevant variables is large enough to make model estimation unwieldy, the agency can exclude attributes that do not make important contributions to the predicted Poisson rates.

The agency can evaluate the synthetic geography models by comparing inferences made with synthetic locations to those made with observed locations, using analyses representative of those anticipated to be of interest to users [34]. The agency can release such evaluations to the public so that analysts can assess what types of analyses are not supported by the synthetic geographies. Another possibility is for agencies to provide feedback to analysts about the quality of the synthetic data inferences for specific estimands; see [35] and [36] for proposals to build automated systems that offer such feedback.

Disease mapping models disregard within-grid cell heterogeneity in spatial intensity surfaces. With fine enough grids, in many data settings such heterogeneity may be modest enough to be swamped by the inherent variability in the synthesis process. When preservation of finer spatial structure is desired, one could use log Gaussian Cox processes [37] and associated computational strategies for fitting them [38, 39]. We leave comparative evaluation of point pattern models and disease mapping models on dimensions of risk, utility, and computational expediency for future research.

Acknowledgments

The authors thank Sudipto Banerjee and Andrew Finley for the helpful comments and suggestions about fitting the spatial regression models.

Contract/grant sponsor: This research was supported by grants from the US National Institute of Health (1R21AG032458) and the US National Science Foundation (SES-11-31897).

A. List of Notations

- $D = (S, X)$ denotes the original data on n individuals, where $S = (s_1, \dots, s_n)$ includes each individual's location and X is the $n \times p$ matrix of each individual's non-spatial attributes.
- $k = 1, \dots, p$ index the number of non-spatial covariates in X .
- d_k denotes the number of levels in X_k .
- $b = 1, \dots, B$ index each distinct attribute pattern in X , where $B \leq \prod_{k=1}^p d_k$
- $x_k^{(b)}$ denotes the value of X_k in b , for each (b, k) .
- $i = 1, \dots, G$ index each grid cell on the study area.
- $c_i^{(b)}$ denotes the number of observations in cell i with attributes b .
- $Z_k^{(b)}$ is a $d_k \times 1$ vector comprising a one at position $x_k^{(b)}$ and zeros elsewhere, for $k = 1, \dots, p$.
- $\lambda_i^{(b)}$ denotes the intensity mean surface of the Poisson model in cell i with attributes b .
- From equation (1):
 - μ is the overall intercept

- each $\alpha_k = (\alpha_{k1}, \dots, \alpha_{kd_k})$ is a $d_k \times 1$ vector of main effects for attribute k
- φ_i is a grid-specific spatial effect
- each $\varphi_{ik} = (\varphi_{ik1}, \dots, \varphi_{ikd_k})$ is a $d_k \times 1$ vector of grid-specific spatial effects for attribute k
- $\epsilon_i^{(b)}$ is an error term
- From equation (2), for $i = 1, \dots, G$:
 - θ_{-i} includes the values of θ_j for all $j \neq i$
 - $\bar{\theta}_i$ is the average of the n_i values of θ_j for cells j that are neighbors of cell i
 - σ_θ^2 is the prior variance hyperparameter for all i
- From equation (3), for ikj : $i = 1, \dots, G$, $k = 1, \dots, p$, $j = 2, \dots, d_k$:
 - $\theta_{-i,kj}$ includes the values of $\varphi_{l,kj}$ for all $l \neq i$
 - $\bar{\phi}_{ikj}$ is the average of the n_i values of $\varphi_{l,kj}$ for cells l that are neighbors of cell i
 - $\sigma_{\phi_{kj}}^2$ is the prior variance hyperparameter for all φ_{ikj}
- From equations (4)-(9):
 - σ_ϵ^2 is the prior variance hyperparameter for all $\epsilon_i^{(b)}$
 - ν_μ is the prior variance hyperparameter for μ
 - ν_{α_k} is the prior variance for each α_{kj}
 - a_θ and b_θ are the prior hyperparameters for σ_θ^2
 - a_{φ_k} and b_{φ_k} are the prior hyperparameters for each $\sigma_{\phi_{kj}}^2$
 - a_ϵ and b_ϵ are the prior hyperparameters for σ_ϵ^2
- $l = 1, \dots, m$ index each draw of the multiple sets of synthetic locations, where m is the number of released synthetic data sets.
- $p_i^{(lb)}$ denotes the normalized probability surface in cell i with attributes b from the draw l .
- $\tilde{D}^{(l)} = (\tilde{S}^{(l)}, X)$ denotes the l -th partially synthetic data set, where $\tilde{S}^{(l)} = (\tilde{s}_1^{(l)}, \dots, \tilde{s}_n^{(l)})$ is one set of synthetic locations.
- M denotes information released about the synthetic data model.

- A denotes any auxiliary information the intruder may possess about the geographies of the records on the file.
- s_t denotes the location of some record t that the intruder seeks to determine.
- $\{\rho_t^s\}$ denotes the posterior probability distribution of the location of record t with $s = 1, \dots, G$, given the information available to the intruder.
- S_{-t} includes the values of $s_{t'}$ for all $t' \neq t$.
- $S_t^s = (s_t = s, S_{-t})$ denotes a new set of locations where s_t is replaced by s .
- $\tilde{c}_i^{(lb)}$ denotes the number of synthetic points in cell i with attributes b from set l .
- L denotes the total number of posterior samples of λ obtained from the MCMC.
- r_t denotes the indicator variable if the maximum posterior probability for record t happens to be on the true s_t .
- R_{all} denotes the proportion of records with $r_t = 1$.
- $c_{(t)}$ denotes the count of individuals in D matching the grid cell and attribute pattern of record t .
- a_t denotes the indicator if record t is the only observation with its attribute level on its grid cell.
- R_{unq} denotes the percentage of records with unique patterns that the intruder correctly locates.
- d_t denotes the distance between s_t and the grid cell with maximum probability.
- z_t denotes the probability that an intruder guesses correctly when randomly choosing a match from among the records with the same attribute pattern and grid cell as record t .
- R_{id} denotes the expected number of correct identifications when randomly choosing a match.
- From Section 3.2:
 - Q denotes a scalar estimand of interest when making inference with partially synthetic data
 - Q is estimated by a point estimator q with variance u
 - $\bar{q}_m$ is the mean of the point estimators from the multiple synthetic data sets
 - b_m and ν_m are terms used to calculate the variance of $\bar{q}_m$, denoted by T_m
 - ν_m denotes the degrees of freedom of the approximated distribution of the estimator $\bar{q}_m$
- From equation (20):

- $\tilde{Y}_t$ denotes the surrogate outcome generated from the spatial logistic model
- β denotes the k fixed coefficients for the main effects included in the model, where $k = 1, \dots, p$
- $w(s_t)$ is a mean-zero Gaussian process for the record t
- σ^2 denotes the partial sill and ϕ denotes the decay parameter at the exponential covariance function
- $\bar{n}$ denotes the average number of neighbors per grid cell.

B. Summary of the posterior distributions of the parameters in $\log \lambda$

We summarize on Table 2 the results of some coefficients from the model for $\log \lambda$ on Equation (1) for the three different grid sizes. The table includes the posterior mean and 95% HPD credible intervals for the intercept μ and the main effects coefficients described by Equations (5)–(6) on Section 2. The posterior analyses were made based on the 10001 MCMC iterations, tossing the first 1000 as burn-in. From the credible intervals, we can see that race and education are the strongest predictors of the Poisson rates.

Table 2

Posterior mean and 95% **HPD** intervals for the coefficients on the expression for $\log \lambda$, for the different grid sizes

	size 10			size 20			size 30		
	mean	LB	UB	mean	LB	UB	mean	LB	UB
μ	-10.94	-11.25	-10.66	-12.51	-12.87	-12.12	-13.44	-13.80	-13.08
$\alpha_{\bar{Y}}$	-0.01	-0.15	0.13	-0.10	-0.30	0.10	-0.12	-0.33	0.09
α_{sex}	0.02	-0.09	0.13	-0.02	-0.18	0.14	-0.03	-0.19	0.13
α_{race}	-1.79	-2.12	-1.48	-2.13	-2.61	-1.69	-2.35	-2.88	-1.85
$\alpha_{\text{age } 2}$	0.09	-0.11	0.32	0.11	-0.14	0.36	0.11	-0.13	0.37
$\alpha_{\text{age } 3}$	0.20	-0.03	0.40	0.18	-0.10	0.45	0.16	-0.13	0.45
$\alpha_{\text{age } 4}$	-0.05	-0.27	0.19	-0.13	-0.44	0.17	-0.18	-0.49	0.16
$\alpha_{\text{edu } 2}$	-0.08	-0.24	0.08	-0.09	-0.28	0.11	-0.10	-0.30	0.10
$\alpha_{\text{edu } 3}$	-0.46	-0.68	-0.25	-0.54	-0.81	-0.27	-0.57	-0.85	-0.29
$\alpha_{\text{edu } 4}$	-2.04	-2.48	-1.61	-2.09	-2.60	-1.63	-2.21	-2.74	-1.69

References

1. National Research Council. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. The National Academies Press; Washington DC: 2007.
2. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*. Mar; 1999 18(5):497–525. [PubMed: 10209808]
3. VanWey LK, Rindfuss RR, Guttman MP, Entwisle B, Balk DL. Confidentiality and spatially explicit data: concerns and challenges. *Proceedings of the National Academy of Sciences*. 2005; 102:15 337–15 342.

4. Zayatz L. Disclosure avoidance practices and research at the U. S. Census Bureau: an update. *Journal of Official Statistics*. 2007; 23:253–255.
5. Young C, Martin D, Skinner CJ. Geographically intelligent disclosure control for flexible aggregation of census data. *International Journal of Geographical Information Science*. 2009; 23:457–482.
6. Wang H, Reiter JP. Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics*. 2012; 6(1):229–252. [PubMed: 23990852]
7. Machanavajjhala, A.; Kifer, D.; Abowd, J.; Gehrke, J.; Vilhuber, L. Privacy: Theory meets practice on the map; IEEE 24th International Conference on Data Engineering; 2008;
8. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*. 1987; 43(3):671–681. [PubMed: 3663823]
9. Besag J, York J, Molli A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; 43:1–20.
10. Clayton, DG.; Bernardinelli, L. Bayesian methods for mapping disease risk. In: Elliott, P.; Cuzick, J.; English, D.; Stern, R., editors. *Geographical and Environmental Epidemiology: Methods for Small Area Studies*. Oxford University Press; 1992. p. 205-220.
11. Wakefield J. Disease mapping and spatial regression with count data. *Biostatistics*. 2007; 8(2): 158–183. [PubMed: 16809429]
12. Willenborg, L.; de Waal, T. *Elements of Statistical Disclosure Control*. Springer-Verlag; New York: 2001.
13. National Research Council. Panel on Data Access for Research Purposes. Committee on National Statistics. Division of Behavioral and Social Sciences and Education. *Expanding Access to Research Data: Reconciling Risks and Opportunities*. The National Academies Press; Washington DC: 2005.
14. An D, Little R, McNally JW. A multiple imputation approach to disclosure limitation for high-age individuals in longitudinal studies. *Statistics in Medicine*. 2010; 29:1769–1778. [PubMed: 20552576]
15. Zhou Y, Dominici F, Louis TA. A smoothing approach for masking spatial data. *Annals of Applied Statistics*. 2010; 4(3):1451–1475.
16. Kaufman CG, Sain SR. Bayesian functional ANOVA modeling using gaussian process prior distributions. *Bayesian Analysis*. 2010; 5(1):123–149.
17. Banerjee, S.; Gelfand, AE.; Carlin, BP. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC; 2004.
18. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 1992; 41(2):337–348.
19. Duncan GT, Lambert D. The risk of disclosure for microdata. *Journal of Business and Economic Statistics*. 1989; 7:207–217.
20. Fienberg SE, Makov UE, Sanil AP. A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics*. 1997; 13:75–89.
21. Reiter JP. Estimating identification risks in microdata. *Journal of the American Statistical Association*. 2005; 100:1103–1113.
22. Drechsler, J.; Reiter, JP. Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In: Domingo-Ferrer, J.; Saygin, Y., editors. *Privacy in Statistical Databases (LNCS 5262)*. Springer-Verlag; New York: 2008. p. 227-238.
23. Reiter JP, Mitra R. Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*. 2009; 1:99–110.
24. Reiter JP. Inference for partially synthetic, public use microdata sets. *Survey Methodology*. Dec; 2003 29(2):181–188.
25. Skinner C. Statistical disclosure risk: Separating potential and harm. *International Statistical Review*. 2012; 80:349–368.
26. Robert, CP.; Casella, G. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc.; Secaucus, NJ, USA: 2005.

27. Skinner CJ, Shlomo N. Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*. 2008; 103:989–1001.
28. Forster J, Webb E. Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series C*. 2007; 56:551–557.
29. Manrique-Vallier D, Reiter JP. Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*. 2012; 107(500):1385–1394. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.2012.710508>.
30. Gelfand AE, Holder M, Latimer A, Lewis PO, Rebelo AG, Silander JA, Wu S. Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*. 2006; 1(1):41–92.
31. Zhou X, Reiter JP. A note on Bayesian inference after multiple imputation. *The American Statistician*. 2010; 64(2):159–163. doi:10.1198/tast.2010.09109.
32. Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70(4): 825–848. doi:10.1111/j.1467-9868.2008.00663.x. URL <http://dx.doi.org/10.1111/j.1467-9868.2008.00663.x>.
33. Finley AO, Sang H, Banerjee S, Gelfand AE. Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*. 2009; 53(8):2873–2884. doi:10.1016/j.csda.2008.09.008. URL <http://www.sciencedirect.com/science/article/pii/S0167947308004386>. [PubMed: 20016667]
34. Reiter JP, Drechsler J. Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*. 2010; 20:405–422.
35. Reiter JP, Oganian A, Karr AF. Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis*. 2009; 53:1475–1482.
36. McClure DR, Reiter JP. Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*. 2012; 4(1)
37. Møller J, Syversveen AR, Waagepetersen RP. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*. 1998; 25(3):451–482. doi:10.1111/1467-9469.00115. URL <http://dx.doi.org/10.1111/1467-9469.00115>.
38. Brix A, Diggle PJ. Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2001; 63(4):823–841. doi: 10.1111/1467-9868.00315. URL <http://dx.doi.org/10.1111/1467-9868.00315>.
39. Rodrigues A, Diggle PJ. Bayesian estimation and prediction for inhomogeneous spatiotemporal log-Gaussian Cox processes using low-rank models, with application to criminal surveillance. *Journal of the American Statistical Association*. 2012; 107(497):93–101. doi: 10.1080/01621459.2011.644496. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.2011.644496>.

Figure 1.
Plots of original locations labeled by different attributes

Figure 2.

Posterior mean surface of λ for the 20×20 grid for white women over age 85 with education less than high school and $\bar{Y} = 0$ (left), and for black men less than age 60 with more than four years of college and $\bar{Y} = 1$ (right).

Figure 3.

Comparison of point estimates and 95% confidence intervals for proportion of black people per zip code, estimated with the original and synthetic data for the three grid sizes.

Figure 4.

Comparison of point estimates and 95% confidence intervals for the proportion of $\tilde{Y} = 1$ per zip code, estimated with the original and synthetic data for the three grid sizes

Figure 5.

Comparison of posterior means and 95% credible intervals from the spatial regression, estimated with the original and the synthetic data sets for the three grid sizes.

Figure 6.
Plot of the original and synthetic locations labeled by race

Figure 7.
Plot of the original and synthetic locations labeled by cause of death $\tilde{Y}$

Figure 8.

Plots of the original and synthetic locations for white women over age 85 with education less than high school and $\tilde{Y}=0$.

Figure 9.

Plots of the original and synthetic locations for black men less than age 60 with more than four years of college and $\tilde{Y} = 1$.

Figure 10.

Histogram of d_t , for the different grid sizes. The highlighted area represents the proportion of observations with $r_t = 1$.

Table 1

Summary of risk measures for the three grid sizes.

Grid size	10 × 10	20 × 20	30 × 30
R_{all}	.082	.060	.088
R_{unq}	.134	.080	.112
R_{id}	.045	.044	.074