The Annals of Applied Statistics
2012, Vol. 6, No. 1, 229–252
DOI: 10.1214/11-AOAS506
© Institute of Mathematical Statistics, 2012

MULTIPLE IMPUTATION FOR SHARING PRECISE GEOGRAPHIES IN PUBLIC USE DATA[1]

BY HAO WANG AND JEROME P. REITER

University of South Carolina and Duke University

When releasing data to the public, data stewards are ethically and often legally obligated to protect the confidentiality of data subjects' identities and sensitive attributes. They also strive to release data that are informative for a wide range of secondary analyses. Achieving both objectives is particularly challenging when data stewards seek to release highly resolved geographical information. We present an approach for protecting the confidentiality of data with geographic identifiers based on multiple imputation. The basic idea is to convert geography to latitude and longitude, estimate a bivariate response model conditional on attributes, and simulate new latitude and longitude values from these models. We illustrate the proposed methods using data describing causes of death in Durham, North Carolina. In the context of the application, we present a straightforward tool for generating simulated geographies and attributes based on regression trees, and we present methods for assessing disclosure risks with such simulated data.

1. Introduction. Statistical agencies, research centers and individual researchers frequently collect geographic data as an integral part of their studies. Geographic data can be highly beneficial for analyses. In studies of aging, for example, they can reveal areas where elderly people live in high densities, which is useful for policy and planning; they can illuminate how environmental factors impact the health and quality of life of elderly people; and, through contextual data, they can yield insights into the social and economic conditions and lifestyle choices of the elderly. Analysts who do not account for spatial dependencies may miss important geographic trends and differences, potentially resulting in invalid inferences.

Geographic variables also are among the most challenging data to share when making a primary data source available to other researchers and the broader public. Very fine geography, while facilitating detailed spatial analyses, enables ill-intentioned users to infer the identities of individuals in the shared file. Even modestly coarse geography can be risky in the presence of demographic or other readily available attributes, which when combined may identify individuals in the shared file. Such identifications are problematic for data collectors, who are ethically and

Received August 2010; revised August 2011.

[1] Supported by NIH Grant R21 AG032458-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Key words and phrases. Confidentiality, disclosure, dissemination, spatial, synthetic, tree.

often legally obligated to protect data subjects' confidentiality. To reduce the risks of disclosures, data collectors typically delete or aggregate geographies to high levels before sharing data. Unfortunately, deletion and aggregation sacrifice the quality of analyses that utilize finer geographic detail.

We propose to protect the confidentiality of data with fine geographic identifiers by simulating values of geographies and other identifying attributes from statistical models that capture the spatial dependencies among the variables in the collected data. These simulated values replace the collected ones when sharing data. To enable estimation of variances, the data steward generates several versions of the data sets for dissemination, resulting in multiply-imputed, partially synthetic data sets [Little (1993), Reiter (2003)]. Such data sets can protect confidentiality, since identification of units and their sensitive data can be difficult when the geographies and other quasi-identifiers 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.

The remainder of the article is as follows. In Section 2 we describe some of the shortcomings of current approaches to protecting data with geographies, and we motivate the use of multiple imputation for releasing public use data with highly resolved geographies. In Section 3 we generate multiply-imputed, partially synthetic versions of a spatially-referenced data set describing causes of death in Durham, North Carolina. As part of the application, we present an easy-to-implement data simulator based on sequential regression trees for synthesizing highly-resolved geographies or attributes. We also describe methods for assessing disclosure risks for data with synthetic geographies. These include (i) a new measure for quantifying the risks that the original geographies could be recovered from the simulated data, and (ii) a measure for assessing risks of re-identifications based on the approach of Reiter and Mitra (2009). In Section 4 we conclude with issues for implementation of the approach.

2. Motivation for using simulated geographies.

At first glance, releasing or sharing safe data seems a straightforward task: simply strip unique identifiers like names and tax identification numbers before releasing data. However, these actions alone may not suffice when other readily available variables, such as geographic or demographic data, remain on the file. These quasi-identifiers can be used to match units in the released data to other databases. When the quasi-identifiers include geographic variables, the risks of identification disclosures can be extremely high. For example, Sweeney [(2001), pages 51 and 52] showed that 97% of the records in a publicly available voter registration list for Cambridge, MA, could be identified using only birth date and 9-digit zip code. Because of the disclosive nature of geography, the U.S. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule requires that, when sharing certain health data, the released geographic units comprise at least 20,000 people [Federal Register (2000), page 82543].

Data stewards can protect confidentiality by restricting public access to the data. For example, analysts can use the data only in secure data enclaves, such as the Research Data Centers operated by the U.S. Census Bureau. Or, analysts can submit queries to remote access systems that provide statistical output without revealing the data that generated the output. While useful, restricted access strategies are only a partial solution. Analysts who do not live near a secure data enclave, or do not have the resources to relocate temporarily to be near one, are shut out from this form of access. Gaining restricted access can require months of proposal preparation and background checks; analysts cannot simply walk in to any secure data enclave and immediately start working with the data. Remote access servers limit the scope of analyses and details of output, since clever queries can reveal individual data values [Gomatam et al. (2005)]. Performing exploratory data analysis and checking model fit are difficult without access to record-level data. Hence, as recommended by two recent National Research Council panels on data confidentiality, to maintain the benefits of wide dissemination, it is necessary to supplement restricted access strategies with readily available, record-level data [National Research Council (2005, 2007)].

2.1. *Common approaches to protecting geography.*

Data stewards commonly employ several strategies for protecting confidentiality when sharing data with geographic identifiers. However, these methods can have serious impacts on the quality of the released data, as we now describe.

Data suppression. Data stewards can suppress geography or attributes from data releases. The intensity of suppression can range from not releasing entire variables, for example, stripping the file of all geographic identifiers, to not releasing small subsets of values, for example, blanking out sensitive attribute values. An example of the former is the Health and Retirement Study: the public use data do not contain any geographic information on relocations [Health and Retirement Study (2007), page 14]. Increasing the intensity of suppression generally increases data protection and decreases data quality. While intense suppression can reduce risks, it has repercussions for inferences. Wholesale deletion of geographic identifiers disables any spatial analysis. When relationships depend on the omitted geography, analysts' inferences are biased. Selective suppression of geography or attributes creates data that are missing not at random, which complicates analyses for users. When there are many records at risk, as is likely the case when the data have fine geographic identifiers, data stewards may need to suppress so many values to achieve satisfactory protection that the released data have very limited quality for spatial analysis.

Data aggregation. Data stewards can coarsen geography or other variables, for example, releasing addresses at the block or county rather than parcel level, or releasing ages in five year intervals. Aggregation reduces disclosure risks by turning unique records—which generally are most at risk—into nonunique records. For example, there may be only one person with a particular combination of demographic characteristics in a street block, but many people with those characteristics in a state. Releasing data for this person with geography at the street level might have a high disclosure risk, whereas releasing the data at the state level might not. The amount of aggregation needed to protect confidentiality depends on the nature of the data. When other identifying attributes are present, such as demographic characteristics, high-level aggregation of the geographic identifiers may be needed to achieve adequate protection. For example, there may be only one person of a certain age, sex, race and marital status—which may be available to ill-intentioned users at low cost—in a particular county, so that coarsening geographies to the county level provides no greater protection for that person than does releasing the exact address.

Aggregation preserves analyses at the level of aggregation. However, it can create ecological inference fallacies [Robinson (1950), Freedman (2004)] at lower levels of aggregation. Additionally, when geography is highly aggregated, analysts may be unable to detect important local spatial dependencies. Despite these limitations, aggregation is the most widely used solution to protect data with geographic identifiers and is routinely implemented by government agencies and other data collectors. The U.S. Census Bureau, for example, does not release geographic identifiers below aggregates of at least 100,000 people in public use files of census data. The public use files for the Health and Retirement Study aggregate geography to “a level no higher than U.S. Census Region and Division” [Health and Retirement Study (2007), page 14].

Aggregation also is frequently used to disguise values in the tails of nongeographic quasi-identifiers, especially age. The HIPAA requires that all ages above 89 be aggregated into and shared as a single category, “90 or older.”

Random noise addition. Data stewards can disguise geographic and other attribute values by adding some randomly selected amount to each confidential observed value. For geographic attributes, this involves moving an observed location to another randomly drawn location, usually within a circle of some radius r centered at the original location. The quality of inferences and the amount of protection depend crucially on r. When a large r is needed to protect confidentiality—as is likely the case when data contain readily available quasi-identifiers—inferences involving spatial relationships can be seriously degraded [Armstrong, Rushton and Zimmerman (1999), VanWey et al. (2005)]. Adding random noise to attribute values introduces measurement error, which inflates variances and attenuates regression coefficients [Fuller (1993)].

Random data swapping. Data stewards can swap data values for selected records, for example, switch values of age, race and sex for at-risk records with those for other records, to discourage users from matching, since matches may be based on incorrect data [Dalenius and Reiss (1982), Fienberg and McIntyre (2004)]. Swapping is used extensively by government agencies. It is generally presumed that swapping fractions are low—agencies do not reveal the rates to the public—because swapping at high levels destroys relationships involving the swapped and unswapped variables. Because data stewards might have to swap all geographic identifiers to ensure released records do not have their actual geographies, swapping is not effective for highly resolved geographic identifiers.

2.2. *Proposed approach*: *Simulate geographic identifiers*. The main limitation of the approaches in Section 2.1 is that they perturb the geography or other quasi-identifiers with minimal or no consideration of the relationships among the variables. Our proposed approach explicitly aims to preserve relationships among the geographic and other attributes through statistical modeling. At the same time, replacing geographic and other quasi-identifiers with imputations makes it difficult for ill-intentioned users to know the original values of those variables, which reduces the chance of disclosures.

Our approach differs from the recent proposal of Zhou, Dominici and Louis (2010), who use spatial smoothing to mask nongeographic attributes at the original locations. Releasing the original locations can result in high risks of identification disclosures when the data include fine geography. Zhou, Dominici and Louis (2010) do not intend to deal with these risks, whereas we explicitly seek to do so. We note that spatial smoothing could be used to mask attribute values after synthesis of locations.

To illustrate how our approach might work in practice, we modify the setting described by Reiter (2004a). Suppose that a statistical agency has collected data on a random sample of 10,000 heads of households in a state. The data comprise each person's street block, age, sex, income and an indicator of disease status. Suppose that combining street block, age and sex uniquely determines a large percentage of records in the sample and the population. Therefore, the agency wants to replace street block, age and sex for all people in the sample—or possibly only a fraction of the three variables, for example, only street block for some records and only age and sex for others—to disguise their identities. The agency generates values of street block, age and sex for these people by randomly simulating values from their joint distribution (see Section 2.3), conditional on their disease status and income values. This distribution is estimated with the collected data. The result is one partially synthetic data set. The agency repeats this process, say, ten times, and these ten data sets are released to the public.

To illustrate how a secondary data analyst might utilize these shared data sets, suppose that the analyst seeks to fit a logistic regression of disease status on income, age, sex and indicator variables for the person's county (obtained by aggregating the released, simulated street blocks). The analyst first estimates the regression coefficients and their variances separately in each simulated data set using standard likelihood-based estimates and standard software. Then, the analyst averages the estimated coefficients and variances across the simulated data sets. These averages are used to form 95% confidence intervals based on the simple formulas developed by Reiter (2003), described below.

The agency creates m partially synthetic data sets, $D^{(1)}, \dots, D^{(m)}$, that it shares with the public. Let Q be the secondary analyst's estimand of interest, such as a regression coefficient or population average. For $l = 1, \dots, m$, let $q^{(l)}$ and $u^{(l)}$ be respectively the estimate of Q and the estimate of the variance of $q^{(l)}$ in synthetic data set $D^{(l)}$. Secondary analysts use $\bar{q}_m = \sum_{l=1}^{m} q^{(l)}/m$ to estimate Q and $T_m = \bar{u}_m + b_m/m$ to estimate $\mathrm{var}(\bar{q}_m)$, where $b_m = \sum_{l=1}^{m} (q^{(l)} - \bar{q}_m)^2/(m-1)$ and $\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m$. For large samples, inferences for Q are obtained from the t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom is $\nu_m = (m-1)[1 + m\bar{u}_m/b_m]^2$. Details of the derivations of these methods are in Reiter (2003). Tests of significance for multicomponent null hypotheses are derived by Reiter (2005c).

Partially synthetic data sets can have positive data utility features. When data are simulated from distributions that reflect the distributions of the collected data, Reiter (2003, 2004b, 2005c) shows that analysts can obtain valid inferences (e.g., 95% confidence intervals contain the true values 95% of the time) for wide classes of estimands. These inferences are determined by combining standard likelihood-based or survey-weighted estimates; the analyst need not learn new statistical methods or software to adjust for the effects of the disclosure limitation. The released data can include simulated values in the tails of distributions, for example, there is no top-coding of ages or incomes [however, it is challenging to develop synthesis models that simultaneously protect confidentiality and preserve inferences when data are very sparse in tails; see Reiter (2005b)]. Because many quasi-identifiers including geography can be simulated, finer details of geography can be released, facilitating estimation for small areas and spatial analyses.

There is a cost to these benefits: the validity of inferences depends on the validity of the models used to generate the simulated data. The extent of this dependence is driven by the nature of the synthesis. For example, when all of age and sex are synthesized, analyses involving those variables reflect only the relationships included in the data generation models. When the models fail to reflect certain relationships accurately, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses. On the other hand, when replacing only a select fraction of age and sex and leaving many original values on the file, inferences are less sensitive to the assumptions of the simulated data models. In practice, this dependence means that data stewards should release information that helps analysts decide whether or not the simulated data are reliable for their analyses. For example, data stewards might include the data generation models (without parameter estimates) as attachments to public releases of data. Or, they might include generic statements that describe the imputation models, such as "Main effects and interactions for age, sex, income and disease status are included in the imputation models for street blocks." Analysts who desire finer detail than afforded by the imputations may have to apply for restricted access to the collected data.

When generating partially synthetic data, the data steward must choose which values to synthesize and must specify models to simulate replacements of those values. In most existing partially synthetic data sets, stewards replace all values of variables that they deem to be either (i) readily available to ill-intentioned users seeking to identify released records, or (ii) too sensitive to risk releasing exactly. However, it may be sufficient from a confidentiality perspective to replace only portions of some variables; see Little, Liu and Raghunathan (2004). The process of specifying synthesis models is typically iterative: the data steward creates synthetic data using a posited model, checks the quality of a large number of representative analyses with the synthetic data, and adjusts the models as necessary to improve quality while maintaining confidentiality protection. For examples of this process, see Drechsler and Reiter (2010) and Kinney et al. (2011).

The data steward also must determine m, that is, how many synthetic data sets to release. Generally, increasing m results in decreased standard errors in secondary analyses. However, increasing m results in greater data storage costs and possibly increased disclosure risks [Reiter and Mitra (2009)]. When small fractions of values are synthesized (e.g., around 10%), the efficiency gains from increasing m are typically modest, so that data stewards can make m modest, for example, $m = 5$, to keep risks and storage costs comparatively low. When large fractions of values are replaced, efficiency gains from increasing m can be substantial [Drechsler and Reiter (2010)]. In such cases, we recommend that data stewards select the largest m that still offers acceptable risks and storage costs.

2.3. *Synthesis models for sharing precise geographies.* Our strategy for simulating geographies involves four general steps. First, the data steward converts the geographic variables on the file to latitudes and longitudes (possibly, using UTM projection to Eastings and Northings). When the collected geographies are aggregated rather than precise locations, the data steward uses a typical value for the location of all records in that area; for example, use the latitude and longitude of the centroid of the street block. Second, the data steward estimates a model for latitudes and longitudes conditional on other variables in the data set. Third, using this model, the data steward simulates new latitudes and longitudes for every record in the file. Fourth and finally, the data steward releases multiple draws of the simulated latitudes and longitudes along with the other attributes—which also might be altered to protect confidentiality, for example, Zhou, Dominici and Louis (2010)—in the original file.

We expect that, in general, some attributes in the data will exhibit spatial dependence. When considering location as the response variable, this implies a joint distribution for latitude and longitude that depends on the attributes and is possibly multi-modal. For example, people of similar age, socio-economic status and other demographic characteristics tend to cluster in neighborhoods, and certain demographic characteristics may be highly prevalent in multiple locations but absent in others. If we ignore these features when simulating geographies—or alter geography with approaches that do not explicitly account for these associations—the spatial relationships in the data will be altered or destroyed.

To illustrate some possible response models for locations, let ϕ_i and λ_i denote the latitude and longitude, respectively, for data subject i. Let $\mathbf{x}_i$ denote the p nongeographical attributes for data subject i. One family of convenient response models is $(\lambda_i, \phi_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_i)$, where each $\boldsymbol{\mu}_i = h(\mathbf{x}_i)$ is a 2×1 vector of unknown means, $h(\mathbf{x}_i)$ is a function of the covariates, and each $\boldsymbol{\Omega}_i$ is an unknown 2×2 covariance matrix. A simple implementation is a bivariate regression model with $h(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} h_j(x_{ij})\beta_j$, where each h_j is a spline for variable j and $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}$ for all i. An alternative is a mixture model with $h(\mathbf{x}_i) = \beta_{i0} + \sum_j h_j(x_{ij})\beta_{ij}$, where $\boldsymbol{\beta}_i = (\beta_{i0}, \dots, \beta_{ip})$ and $\boldsymbol{\Omega}_i$ come from K mixture components.

In specifying a response model for locations, the data steward should include components of $\mathbf{x}$ that vary with spatial locations. The data steward also should seek a flexible model that can adapt to a potentially complex response distribution. In the application, we describe a semi-automated approach for approximating the response distribution that can be easily implemented by data stewards. We emphasize, however, that the idea of treating latitude and longitude as a response is general, and that data stewards can improve the quality of the released data by tailoring the response model to their particular problem.

To our knowledge, treating geography as a continuous response and releasing simulated draws from its distribution has not been previously implemented. However, partially synthetic data are used to protect locations in the Census Bureau's OnTheMap project [Machanavajjhala et al. (2008)]. In that project, Machanavajjhala et al. (2008) synthesize the street blocks where people live conditional on the street blocks where they work and other block-level attributes. They use multinomial regressions to simulate home-block values, constraining the possible outcome space for each individual based on where they work. Our approach differs from the OnTheMap modeling in that (i) we model more precise geography, that is, continuous versions of latitudes and longitudes, than discrete street blocks, and (ii) we do not rely on a fixed set of geographic locations, that is, where people work, to anchor the synthesis models. Furthermore, for settings with high-dimensional $\mathbf{x}_i$ and no obvious way to set constraints on the outcome space, multinomial regressions can be computationally demanding if even estimable, whereas continuous response models are readily estimated.

3. Application: Protecting a cause of death file.

We now apply the multiple imputation approach to create disclosure-protected data on a subset of North Carolina (NC) mortality records in 2002. The data include precise longitudes and latitudes of deceased individuals' residences, as well as a variety of variables related to manner of death; we consider the subset of variables in Table 1. These mortality data are in fact publicly available and so do not require disclosure protection. Nonetheless, they are ideal test data for methods that protect confidentiality of geographies since, unlike many data sets on human individuals, actual locations are available and can be revealed for comparisons. Access to the data is managed by the Children's Environmental Health Initiative at Duke University per agreement with the state of NC.

We use individuals whose place of residence was one of seven contiguous postal zones in Durham, NC. These areas are heterogeneous in terms of population density and characteristics. For simplicity, we include only individuals with race of black and white—which comprised 99% of all records in these postal zones—resulting in $n = 2{,}670$ observed cases. We also collapse the cause of death variable into two levels: death from diseases of the circulatory and respiratory system, and death from all other causes. We consider this binary variable, which we label Y, as the outcome for regression models.

In these data, Y does not exhibit strong residual spatial dependence after accounting for other variables. Therefore, for a more thorough test of the analytical validity of the synthetic data sets, we also generate a surrogate cause of death variable, $\tilde{Y}$, that exhibits spatial clustering and is dependent on several nongeographic variables. To do so, we generate outcomes as follows:

$$\tilde{Y}_i \sim \text{Bern}(\pi_i), \tag{1}$$

TABLE 1
Description of variables used in the empirical study

Variable	Range
Longitude	Recoded to go from 1–100
Latitude	Recoded to go from 1–100
Sex	Male, female
Race	White, black
Age (years)	16–99
Autopsy performed	Yes, no, missing
Autopsy findings	Yes, no missing
Marital status	5 categories
Attendant	Physician, medical examiner, coroner
Hispanic	7 categories
Education (years)	0–17 years
Hospital type	8 categories

where $\operatorname{logit}(\pi_i) = 0.02 + \text{Sex}_i + \text{Race}_i + 0.003\,\text{Age}_i + w(\mathbf{s}_i)$, $\mathbf{s}_i = (\lambda_i, \phi_i)$, and $w(\mathbf{s})$ is a mean zero Gaussian process with exponential covariance function $C(\mathbf{s}, \mathbf{s}') = \sigma_e^2 \exp(-\phi_e \|\mathbf{s} - \mathbf{s}'\|_2)$. We set the parameters of the exponential covariance function to $\sigma_e^2 = 2$ and $\phi_e = 0.06$, so that the effective range (i.e., the distance at which the spatial correlation drops to 0.05) is about $-\log(0.05)/\phi_e = 50$, which equals half of the overall range of the latitudes and longitudes in Table 1. The coefficients of the covariates are specified so that the covariates have strong effects. All results that follow use $\tilde{Y}$ in place of the actual cause of death Y; see the online supplement [Wang and Reiter (2011)] for selected results based on Y. For both $\tilde{Y}$ and Y, the results are qualitatively similar, in that the actual spatial relationships (or lack thereof) in the original data are approximately preserved in the synthetic data sets.

3.1. *Generation of synthetic data.* We examined several methods for simulating latitude and longitude, including mixtures of bivariate regressions, bivariate partition models [De'ath (2002)] using the "mvpart" function in R, Bayesian additive regression trees [Chipman, George and McCulloch (2010)], and classification and regression trees (CART) [Breiman et al. (1984)]. Among these, the CART synthesizer resulted in data sets with a desirable profile in terms of low disclosure risks and high data usefulness. Furthermore, the CART synthesizer is fastest computationally and easy to implement, as it requires minimal tuning. It scales to large data sets with many predictors and many observations. In comparison to the CART synthesizer, the Bayesian trees and mixture model synthesizers were far more computationally demanding, and the bivariate partition model synthesizer resulted in unacceptably high disclosure risks. We therefore present results only for the CART synthesizer, which we now summarize; see Reiter (2005d) for further information on CART synthesizers.

Let $\mathbf{x}$ include all nongeographic attributes in Table 1 and $\tilde{Y}$. First, we fit a regression tree of longitude on $\mathbf{x}$. Label the tree as $\mathcal{T}_\lambda$, where λ stands for longitude. Let $L_{\lambda,w}$ be the wth leaf in $\mathcal{T}_\lambda$, and let $\boldsymbol{\lambda}_{L_{\lambda,w}}$ be the $n_{L_{\lambda,w}}$ values of λ in leaf $L_{\lambda,w}$. In each $L_{\lambda,w}$, we draw $n_{L_{\lambda,w}}$ values from $\boldsymbol{\lambda}_{L_{\lambda,w}}$ using the Bayesian bootstrap [Rubin (1981)]. We then smooth the density of the bootstrapped values using a Gaussian kernel density estimator with bandwidth h_λ and support over the smallest to the largest value of $\boldsymbol{\lambda}_{L_{\lambda,w}}$. To get a synthetic longitude for the ith unit, we trace down $\mathcal{T}_\lambda$ based on the unit's values of $\mathbf{x}_i$, and we sample randomly from the estimated mixture density in that unit's leaf. The result is a set of synthetic longitudes, $\tilde{\boldsymbol{\lambda}}^{(l)}$.

Next, we fit the regression tree of latitude on $\mathbf{x}$ and the true λ; label the tree as $\mathcal{T}_\phi$, where ϕ stands for latitude. To locate the ith person's leaf in $\mathcal{T}_\phi$, we use $\tilde{\lambda}_i^{(l)}$ in place of λ_i. For units with combinations of $(\mathbf{x}_i, \tilde{\lambda}_i^{(l)})$ that do not belong to one of the leaves of $\mathcal{T}_\phi$, we search up the tree until we find a node that contains the combination, and treat that node as if it were the unit's leaf. Once each unit's leaf is located, values of $\phi_i^{(l)}$ are generated using the Bayesian bootstrap and kernel density procedure with bandwidth h_ϕ. The result is a set of synthetic latitudes, $\tilde{\boldsymbol{\phi}}^{(l)}$, and, therefore, synthetic locations $\tilde{\mathbf{s}}^{(l)} = (\tilde{\boldsymbol{\lambda}}^{(l)}, \tilde{\boldsymbol{\phi}}^{(l)})$.

We repeat the process of generating $\tilde{\mathbf{s}}^{(l)}$ independently m times, resulting in the collection of partially synthetic data sets, $D^{(l)} = \{\mathbf{x}, \tilde{\mathbf{s}}^{(l)}\}$ where $l = 1, \dots, m$. With no further synthesis of $\mathbf{x}$, these m data sets would be released to the public.

We also performed the synthesis by generating latitude first and longitude second. As reported in the online supplement [Wang and Reiter (2011)], this ordering results in slightly decreased disclosure risks and slightly worse data utility. We recommend that data stewards try both orderings and choose the one that results in the more desirable risk-utility profile. For general discussions on the order of synthesis, see Reiter (2005d) and Caiola and Reiter (2010).

We also investigate simulating both geography and nongeographic identifiers to further improve confidentiality protection. Specifically, we simulate values of race (R) and age (A) in addition to (λ, ϕ). We choose these two variables because (i) in many applications, age and race might be considered available to ill-intentioned users and hence prominent candidates for disclosure protection, (ii) their distributions clearly depend on location in the NC mortality data, and (iii) they encompass the generic modeling challenges of a continuous and a categorical variable.

The process proceeds as follows. Simulate $(\tilde{\lambda}, \tilde{\phi})$ using the CART synthesizers as before, but excluding R and A from $\mathbf{x}$. We simulate new values of A using a CART synthesizer fit with $(\mathbf{x}, \lambda, \phi)$. Each A_i is simulated based on its $(\tilde{\lambda}_i, \tilde{\phi}_i)$. We simulate new values of R using a CART synthesizer fit with $(\mathbf{x}, \lambda, \phi, A)$. Each R_i is simulated based on its $(\tilde{\lambda}_i, \tilde{\phi}_i, \tilde{A}_i)$.

For all trees, we require the smallest node size to be at least five, and we cease splitting a leaf when the deviance of values in the leaf is less than 0.0001; see Section 4 for discussion of selecting these tuning parameters. All CART models are fit in R using the "tree" function. The bandwidth sizes are directly related to the analytical utility and disclosure risks of the synthetic data sets. Here, we investigate the risk-utility trade-offs for three bandwidths: $h_\lambda = h_\phi \in \{1, 5, 10\}$. We set the bandwidth for generating $\tilde{A}$ equal to 2. We generate $m = 5$ synthetic data sets.

3.2. *Evaluation of confidentiality protection.* For an initial evaluation of the protection engendered by simulation, we plot $(\tilde{\lambda}, \tilde{\phi})$ against (λ, ϕ) for one simulated data set when only geography is imputed with $h = 1$; see Figure 1. Clearly, $(\tilde{\lambda}_i, \tilde{\phi}_i)$ can vary greatly from (λ_i, ϕ_i). However, Figure 1 is a crude evaluation, as intruders can utilize information from the multiple synthetic data sets and possibly other information to attempt disclosures.

We now outline frameworks for evaluating disclosure risks. We begin with an approach for quantifying how much intruders can learn about actual geographies from the synthetic data.

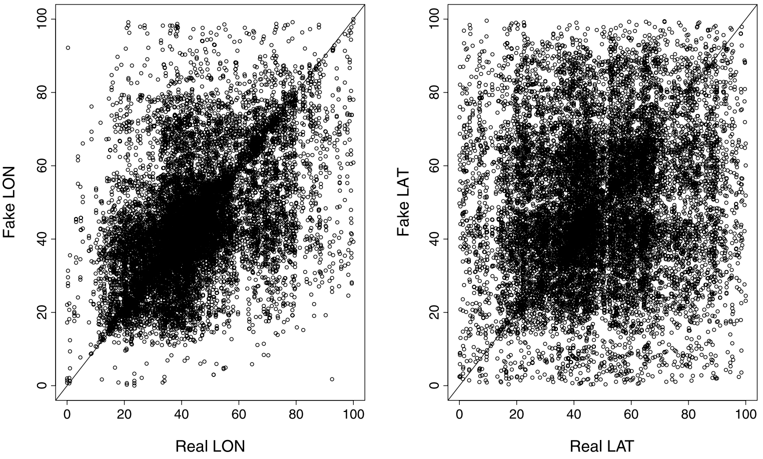

FIG. 1. *Scatter plots of synthetic longitudes (left) and latitudes (right) against real ones under the synthesis model that imputes geography only with $h = 1$.*

3.2.1. *Risk of geography disclosure.* In this section we assume that geography is the only synthesized variable, although the general ideas and approach apply to other attributes and with additional synthetic data.

Let $\tilde{\mathbf{s}}_i = (\tilde{\mathbf{s}}_i^{(1)}, \dots, \tilde{\mathbf{s}}_i^{(m)})$; let $\tilde{\mathbf{S}}$ be the collection of $\tilde{\mathbf{s}}_i$ for all n persons in the sample; and, let $\mathbf{S}$ include all n original values of $\mathbf{s}_i$. Let $\mathbf{S}_{-i}$ be all the original geography except for that of the ith person. Let M represent any meta-data released by the data steward about the synthesis models, for example, the code for the computer program that generated that synthetic data (without the original data or parameter estimates). Let I represent the intruder's prior information on persons' geography in the sample, for example, I might include $\mathbf{S}_{-i}$. Either M or I could be empty.

We posit an intruder whose goal is to estimate $\mathbf{s}_i$ for one or more target records in the database. Specifically, for any record i, the intruder seeks the posterior distribution of $\mathbf{s}_i$ given $(\mathbf{X}, \tilde{\mathbf{S}}, M, I)$. With this posterior distribution, the intruder could identify high density regions for the unknown $\mathbf{s}_i$, which, if precise enough, could be used to pinpoint the true location of the target individual. Using Bayes' rule, we have

$$P(\mathbf{s}_i|\mathbf{X}, \tilde{\mathbf{S}}, M, I) \propto P(\tilde{\mathbf{S}}|\mathbf{X}, \mathbf{s}_i, M, I) P(\mathbf{s}_i|\mathbf{X}, M, I), \tag{2}$$

where $P(\mathbf{s}_i|\mathbf{X}, M, I)$ represents the intruder's prior beliefs about $\mathbf{s}_i$.

The information in M and I play central roles in the likelihood function $P(\tilde{\mathbf{S}}|\mathbf{X}, \mathbf{s}_i, M, I)$. For example, suppose that M contains the code of the computer program used to generate the synthetic data (without original data or parameter estimates). If I includes $\mathbf{S}_{-i}$, the intruder could take guesses at $\mathbf{s}_i$ according to his or her prior distribution and, with the resulting guess of $\mathbf{S}$, determine the likelihood of $\tilde{\mathbf{S}}$. If instead I contains only a portion of the geographies or is empty, as are likely to be the cases in practice, the computation of the likelihood becomes much more complex and uncertain, since the intruder needs to guess at multiple unknown geographies. In such cases, one simple approximation of the distribution for $\mathbf{s}_i$ is the convex hull of the set $\tilde{\mathbf{s}}_i$. Given the variation in Figure 1, these regions in the mortality data could be quite large.

The intruder's prior distribution is also a key determinant of the posterior distribution of $\mathbf{s}_i$. An intruder may know the locations of all individuals in the population with certain characteristics contained in $\mathbf{x}_i$, and the prior distribution could be uniform over those locations. An intruder who knows $\mathbf{x}$ and $\mathbf{S}_{-i}$ could estimate a model from these data to predict $\mathbf{s}_i$, and use that as a prior distribution. An intruder with no external information might use a uniform distribution on the map of possible locations. Unfortunately, it is nearly impossible for the data steward to know the information possessed by the intruder. Hence, it is prudent for the data steward to consider disclosure risks under a variety of assumptions about the intruders' knowledge—including very extensive prior knowledge, which represents possible worst case scenarios—as we now demonstrate.

Using the CART synthesizer, we consider two scenarios for the NC mortality data: a high-risk scenario in which the intruders know everything except for one target's $\mathbf{s}_i$, that is, $\mathbf{X}$ and $\mathbf{S}_{-i}$, and a low-risk scenario in which the intruder does not know any records' geographies. We assume that M includes everything about the trees except the individual geographies in the nodes, that is, the data steward releases the splitting rules for each tree and the kernel bandwidths. For the risky scenario, we assume the intruder's prior distribution is uniform on a grid over a small area containing the target's true latitude and longitude, and estimate equation (2) using importance sampling; see the online supplement [Wang and Reiter (2011)] for details. Because the small area contains the true value, this prior distribution represents strong intruder prior knowledge. We note that other specifications for the prior distribution could change the value of the risk measure.

To summarize how much the CART synthesis protects geographies, we create two risk metrics. Let $(\phi_{i,t}, \lambda_{i,t})$ be a draw from $P(\mathbf{s}_i|\mathbf{X}, \tilde{\mathbf{S}}, M, I)$. The metrics are

$$R_1 = \left[\int \{(\phi_i - \phi_{i,t})^2 + (\lambda_i - \lambda_{i,t})^2\} P(\phi_{i,t}, \lambda_{i,t}|\mathbf{X}, \tilde{\mathbf{S}}, M, I)\, d\phi_{i,t}\, d\lambda_{i,t}\right]^{1/2},$$

$$R_2 = \text{number of actual cases in circle centered at } (\phi_i, \lambda_i) \qquad \text{with radius } R_1.$$

Here, R_1 measures the average Euclidean distance between the intruder's guess of geography and the actual geography. Larger values of R_1 (up to a max of $100\sqrt{2}$) indicate larger uncertainty in predicting $\mathbf{s}_i$, so that intruders' predictions are more likely to be further away from the true geography; thus, larger values of R_1 indicate smaller disclosure risks. Larger values of R_2 indicate that many actual locations (up to a max of $n = 2{,}670$) are reasonable guesses at $\mathbf{s}_i$, thus smaller disclosure risks.

Table 2 displays summary statistics for R_1 and R_2 for all $n = 2{,}670$ records in the database. For the low-risk scenario, the medians of R_1 for all three bandwidth

TABLE 2

Summary of geography disclosure risks for the low risk and high risk scenarios for different bandwidths h when synthesizing only geography. For each risk measure, α_0 *is the minimum,* α_{25} *is the first quartile, and* α_{50} *is the median*

		$h = 1$			$h = 5$			$h = 10$		
Scenario	**Risk**	α_0	α_{25}	α_{50}	α_0	α_{25}	α_{50}	α_0	α_{25}	α_{50}
Low	R_1	4.2	15.6	21.6	3.6	15.4	20.8	3.8	16.0	21.4
	R_2	36	384	680	27	373	640	43	393	674
High	R_1	0.0	4.3	9.7	2.1	9.7	14.2	3.4	13.1	17.8
	R_2	0	34	159	4	149	327	13	253	474

values are around 21 distance units, and the medians of R_2 are around 670 units, indicating that most $\mathbf{s}_i$ are estimated with sizable uncertainty. In this scenario, each person's R_1-radius circle contains at least 27 other cases. Interestingly, for this scenario, increasing the bandwidth does not substantially increase the uncertainty in $\mathbf{s}_i$. For the high-risk scenario, the intruder can estimate $\mathbf{s}_i$ with better accuracy than in the low-risk scenario. Here, both R_1 and R_2 decrease with h. In fact, when $h = 1$, there are individuals in the data who are alone in their R_1-radius circles. The boxplot of Figure 2 in the online supplement [Wang and Reiter (2011)] provides additional information about the distributions of R_1 and R_2, including those under different scenarios when generating latitude first and longitude second.

3.2.2. *Risk of identification.* The approach in Section 3.2.1 can be used to estimate posterior distributions of any attribute, of which location is one example. Often, however, data stewards want to assess the risks that individuals in the released data can be re-identified. To quantify these risks, we now compute probabilities of identification [Duncan and Lambert (1989), Fienberg, Makov and Sanil (1997), Reiter (2005a)] by adapting the approach of Drechsler and Reiter (2008) and Reiter and Mitra (2009) for synthetic geographies.

In this approach, the data steward mimics the behavior of an intruder who possesses the true values of the quasi-identifiers, including geographies, for selected target records (or even the entire database). To illustrate, suppose the intruder has a vector of information, $\mathbf{t}$, on a particular target unit in the population which may or may not correspond to a unit in the m released synthetic data sets, $D = \{D^{(1)}, \dots, D^{(m)}\}$. Let t_0 be the unique identifier (e.g., the individual's name) of the target, and let d_{j0} be the (not released) unique identifier for record j in D, where $j = 1, \dots, n$. The intruder's goal is to match unit j in D to the target when $d_{j0} = t_0$, and not to match when $d_{j0} \neq t_0$ for any $j \in D$.

Let J be a random variable that equals j when $d_{j0} = t_0$ for $j \in D$ and equals $n + 1$ when $d_{j0} = t_0$ for some $j \notin D$. The intruder thus seeks to calculate the $P(J = j|\mathbf{t}, D, M)$ for $j = 1, \dots, n + 1$. He or she then would decide whether or not any of the identification probabilities for $j = 1, \dots, n$ are large enough to declare an identification. Let $\mathbf{T}$ be all original values of the variables that were synthesized. Because the intruder does not know the actual values in $\mathbf{T}$, he or she should integrate over its possible values when computing the match probabilities. Hence, for each record in D we compute

$$P(J = j|\mathbf{t}, D, M) = \int \Pr(J = j|\mathbf{t}, D, \mathbf{T}, M, I) \Pr(\mathbf{T}|\mathbf{t}, D, M, I)\, dT.$$

This integral can be approximated using Monte Carlo approaches; details are in the online supplement. Once again, the data steward must make assumptions about I, the information the intruder knows about the targets.

Data stewards can summarize the risks for the entire data set using functions of these match probabilities [Reiter (2005a)]. Let c_j be the number of records in the data set with the highest match probability for the target $\mathbf{t}_j$. Let $g_j = 1$ if the true match is among the c_j units, and $g_j = 0$ otherwise. The expected match risk equals $\sum_j (1/c_j)g_j/n$. The true match risk equals $\sum_j k_j/n$, where $k_j = 1$ when $c_j g_j = 1$, and $k_j = 0$ otherwise. The false match risk equals $\sum_j f_j(1 - g_j)/\sum_j f_j$, where $f_j = 1$ when $c_j = 1$ and $f_j = 0$ otherwise. Effective disclosure limitation techniques have low expected and true match risks, and high false match risks.

Using the mortality data, we consider three scenarios with different information in M. In the first M contains everything, that is, details of the CART models, the splitting rules and the real data values in each leaf and internal node. Essentially, M is a data simulator that enables analysts to generate new synthetic data sets using the same process as the data steward. In the second M contains descriptions of the CART models, but not the specific splitting rules nor the real data values in each leaf and internal node. Essentially, this is akin to releasing the code used to simulate data without providing any parameter values for it. In the third M is empty, that is, the data steward says nothing about how the data were collected.

For all scenarios, we suppose that intruders have a file containing the true values of sex, race, marital status, age and geography for all $n = 2{,}670$ units in the data set, and that they seek to match records in D to this file. We also suppose that the intruder knows which records were in the sample, so that $P(J = 2671|\mathbf{t}, D, M) = 0$. We compute each target's probability independently of other targets' probabilities and match with replacement.

Table 3 summarizes risk measures in one set of $m = 5$ synthetic data sets for each bandwidth and scenario. Three general trends are evident; these persist in two additional runs of the simulation as well. First, the synthesis of age and race dramatically decreases disclosure risks. Indeed, we suspect that many data stewards would consider the numbers of true matches unacceptably high for synthesizing geography only and perhaps acceptable for synthesizing geography, age and race. Second, releasing additional information in M increases the disclosure risks. This trend is particularly pronounced when synthesizing only geography, and less so

TABLE 3

Summary of risk measures under different scenarios when synthesizing only geography and when synthesizing geography, age and race. Here, E is expected match risk, T is true match risk, and F is false match risk. Results based on one simulation run per scenario

	$h=1$			$h=5$			$h=10$		
Information in M	E	T	F	E	T	F	E	T	F
				Synthesizing geography only					
Empty	0.21	0.15	0.76	0.19	0.12	0.78	0.18	0.11	0.80
Code, no parameters	0.21	0.19	0.78	0.19	0.16	0.80	0.18	0.15	0.82
Everything	0.34	0.32	0.48	–	–	–	–	–	–
				Synthesizing geography, age and race					
Empty	0.010	0.008	0.98	0.008	0.007	0.99	0.007	0.006	0.99
Code, no parameters	0.009	0.009	0.99	0.008	0.008	0.99	0.005	0.005	0.99
Everything	0.034	0.034	0.94	–	–	–	–	–	–

when synthesizing geography, age and race. For the latter synthesis strategy, the incremental risk of releasing the synthesis code without parameters over releasing nothing is modest, suggesting that it is worth releasing M to improve analysts' understanding of the disclosure limitation applied to the data. Third, the risks tend to increase as the bandwidth for geography synthesis decreases. This is because larger h implies larger variances in the synthetic locations.

3.3. *Evaluation of analytical validity.* As with disclosure risks, the extent to which synthetic data sets can support analytically valid inferences depends on the properties of the synthesizer. In this section we examine the quality of synthetic data inferences for several estimands in the NC mortality data set. Based on the huge reductions in disclosure risks, we only consider scenarios with (λ, ϕ, R, A) synthesized. The online supplement [Wang and Reiter (2011)] provides corresponding results with only (λ, ϕ) synthesized.

Table 4 summarizes a repeated sampling experiment involving descriptive estimands at the zip code level. For each of 100 simulation runs, we create $m = 5$ synthetic data sets using the observed mortality data (with $\tilde{Y}$) and the CART synthesizers with $h \in (1, 5, 10)$. For the percentage-related estimands, the mean square error (MSE) is typically less than 3%, and for age-related estimands, the MSE is typically less than 2.5 years. The MSEs for age-related estimands are generally smaller than the other MSEs because age does not vary spatially as much as the other variables do; hence, the synthesis process for age is comparatively robust to imperfect modeling of the relationship between geographies and the attributes. The MSEs tend to increase as h increases, although the changes for the most part are only 3% or smaller. Overall, the results suggest that the synthetic data do a rea-

TABLE 4

Summary of simulation results for descriptive estimands when imputing both geography and nongeography. Q stands for the population values; ME and MSE stand for the median and mean square error of $\bar{q}_5$ across the 100 simulations

			$h=1$		$h=5$		$h=10$	
Estimand	**ZIP**	Q	**ME**	**MSE**	**ME**	**MSE**	**ME**	**MSE**
% black	Z1	61.1	59.6	2.4	56.2	4.9	55.7	5.6
	Z2	41.1	43.4	1.6	41.8	1.7	40.8	1.0
	Z3	32.6	33.7	1.7	34.0	2.0	35.5	3.0
	Z4	13.5	13.4	0.9	13.6	0.9	13.7	0.9
	Z5	46.2	44.8	1.9	43.8	2.7	42.8	3.6
	Z6	12.9	15.2	2.5	16.4	3.7	16.3	3.5
	Z7	51.3	52.6	1.9	53.9	2.9	55.5	4.8
% with educ. > 14.5	Z1	19.9	18.9	1.8	20.0	1.3	21.8	2.2
	Z2	9.6	7.5	2.3	7.4	2.3	8.1	1.6
	Z3	10.9	9.9	1.4	10.6	0.9	13.0	2.2
	Z4	28.3	29.3	1.6	27.8	1.1	28.1	0.8
	Z5	36.6	37.5	1.6	38.9	2.8	38.2	1.9
	Z6	25.8	26.5	1.6	26.7	1.8	27.3	2.0
	Z7	30.9	32.8	2.6	32.0	1.9	31.8	1.7
Avg. age	Z1	65.8	67.6	1.9	68.7	2.9	69.1	3.4
	Z2	66.2	68.4	2.3	68.7	2.5	68.6	2.4
	Z3	71.4	70.9	0.6	70.5	1.0	70.1	1.3
	Z4	72.5	71.6	0.9	71.3	1.2	71.2	1.3
	Z5	71.1	70.0	1.2	69.8	1.3	69.8	1.3
	Z6	72.1	71.3	1.0	71.4	0.8	71.5	0.7
	Z7	69.4	69.4	0.5	69.5	0.5	69.7	0.6

sonable job of preserving the aggregated spatial relationships in the data for these variables.

We next evaluate inferences from two regression models. The first is a standard logistic regression of $\tilde{Y}$ on main effects for sex, age and race. The second is a Bayesian spatial logistic regression of $\tilde{Y}$ on main effects for sex, age and race that uses an exponential covariance function for spatial random effects, as in (1). To aid in the evaluation of the synthetic data sets, we randomly choose 2,470 people as a training set to fit the models and the remaining 200 people as a testing set to evaluate the predictive performance. Because the sample size of this training set is large for fitting hierarchical spatial random-effects models, we use Gaussian predictive process models [Banerjee et al. (2008)] to reduce computational burden. To do so, we select 100 knots by randomly choosing a subset of the locations in the training set. We assign flat prior distributions on regression coefficients $\boldsymbol{\beta}$, an inverse Gamma $(2, 1)$ prior for σ_e^2 and a uniform prior on $(0.01, 1)$ for ϕ_e. The same training sample, testing sample and knots are used for all analyses, that

TABLE 5

Summary results for spatial and nonspatial logistic regressions. Results include point and variance estimates for regression coefficients, and misclassification rates (MR)

	Real data		$h = 1$		$h = 5$		$h = 10$	
	Q	$\sqrt{T}$	$\bar{q}_5$	$\sqrt{T_5}$	$\bar{q}_5$	$\sqrt{T_5}$	$\bar{q}_5$	$\sqrt{T_5}$
				Nonspatial GLM				
Intercept	−0.85	0.18	−0.76	0.21	−0.71	0.20	−0.67	0.19
Sex	0.60	0.08	0.61	0.09	0.61	0.09	0.61	0.08
Race	0.59	0.09	0.48	0.18	0.48	0.13	0.42	0.11
Age × 100	0.52	0.24	0.43	0.27	0.36	0.28	0.34	0.26
	MR		**MR**		**MR**		**MR**	
In-sample	0.42		0.42		0.42		0.42	
Out-of-sample	0.46		0.47		0.47		0.47	
				Spatial GLM				
Intercept	−1.15	0.43	−0.91	0.37	−0.83	0.44	−0.97	0.33
Sex	0.74	0.10	0.64	0.09	0.67	0.09	0.68	0.10
Race	0.82	0.12	0.62	0.23	0.61	0.15	0.56	0.13
Age × 100	0.68	0.25	0.65	0.31	0.48	0.28	0.53	0.28
σ_e^2	1.83	0.96	1.82	1.20	1.31	0.73	1.13	0.53
ϕ_e	0.05	0.02	0.06	0.02	0.06	0.01	0.06	0.01
	MR		**MR**		**MR**		**MR**	
In-sample	0.22		0.26		0.25		0.27	
Out-of-sample	0.32		0.35		0.31		0.32	

is, we do not perform a repeated sampling experiment because of computational burden of estimating the spatial regression model. All models are estimated using the "spGLM" function in R.

Table 5 summarizes the original and synthetic data inferences and predictions. For standard logistic regression, we estimate the coefficients using the methods of Reiter (2003). Misclassification rates are based on predicting $\tilde{Y}_i = 1$ when $p_i = 1/(1 + e^{-\mathbf{x}_i'\bar{\boldsymbol{\beta}}_5}) > 0.5$ and predicting $\tilde{Y}_i = 0$ otherwise, where $\bar{\boldsymbol{\beta}}_5$ is the vector of synthetic point estimates for the coefficients. For the Bayesian spatial logistic regression, we mix the posterior samples of the coefficients from each of the five synthetic data sets, and report the posterior mean and variance of the mixed samples. Misclassification rates are based on predicting $\tilde{Y}_i = 1$ when the posterior mean of p_i across the five synthetic data sets exceeds 0.5 and predicting $\tilde{Y}_i = 0$ otherwise. For both models, we compute the in-sample misclassification rates as the proportions of misclassified cases conditioned on the training set, and the out-of-sample misclassification rates as the proportions of misclassified cases conditioned on the test set. All out-of-sample predictions for the Bayesian spatial logistic regression are carried out using the "spPredict" function in R.

For the logistic regression, Table 5 indicates that synthetic point estimates are generally close to those for the observed data, although there is attenuation in the coefficients for the synthesized variables. This attenuation increases with h. Both in-sample and out-of-sample misclassification rates for the synthetic data are similar to those for the observed data.

For the spatial regression, Table 5 indicates that the synthetic point estimates are generally close to the observed data estimates, again with increasing attenuation as h gets large. The spatial random effects parameters σ_e^2 and ϕ_e in the synthetic data are similar to those from the observed data when $h = 1$, but σ_e^2 declines toward zero as h gets large. This indicates that large values of h can weaken the spatial associations in the synthetic data.

It is also informative to compare the misclassification rates for the spatial logistic regression in the synthetic data with the rates for the nonspatial logistic regression in the observed data. In particular, both in-sample and out-of-sample misclassification rates are significantly lower in spatial logistic regression for the synthetic data than those in nonspatial logistic regression for the observed data. This suggests that, when spatial dependencies are strong, releasing simulated geographies enables better predictions than suppressing geography, even when race and sex are also simulated.

The online supplement [Wang and Reiter (2011)] reports the results of the descriptive analyses and the spatial regressions based on synthetic data sets generated from the actual cause of death Y, which does not exhibit strong spatial dependence. The results for the descriptive estimands are similar to, and even slightly better than, those from Table 4. For the spatial regressions, the synthetic data sets appropriately reflect the lack of spatial dependence in Y. As a final illustration of the usefulness of the synthetic data sets, Figure 2 displays maps of location by race for the actual data and for three synthetic data sets ($m = 1$) based on a CART synthesizer with $h \in (1, 5, 10)$. Across all values of h, the synthetic data sets preserve the spatial distribution of race reasonably well.

3.4. *Comparison against random noise addition.* When considering the merits of synthetic data approaches, another relevant comparison is against other disclosure limitation procedures rather than against the original data, which cannot be made publicly available. We now compare the synthetic data sets with only geography simulated against adding random noise to geography, that is, moving an observed location to another randomly drawn location. To make results comparable, we perturb each $\mathbf{s}_i$ by drawing a random value $\mathbf{s}_i^*$ from a bivariate normal distribution with a mean equal to $\mathbf{s}_i$ and a diagonal covariance matrix with standard deviations equally set to be the corresponding $R_{1,i}/\sqrt{2}$. Here, $R_{1,i}$ is computed assuming that, in the high-risk scenario, only geography is synthesized and that $h = 1$. In this way, the synthetic and noise-infused data sets have roughly the same

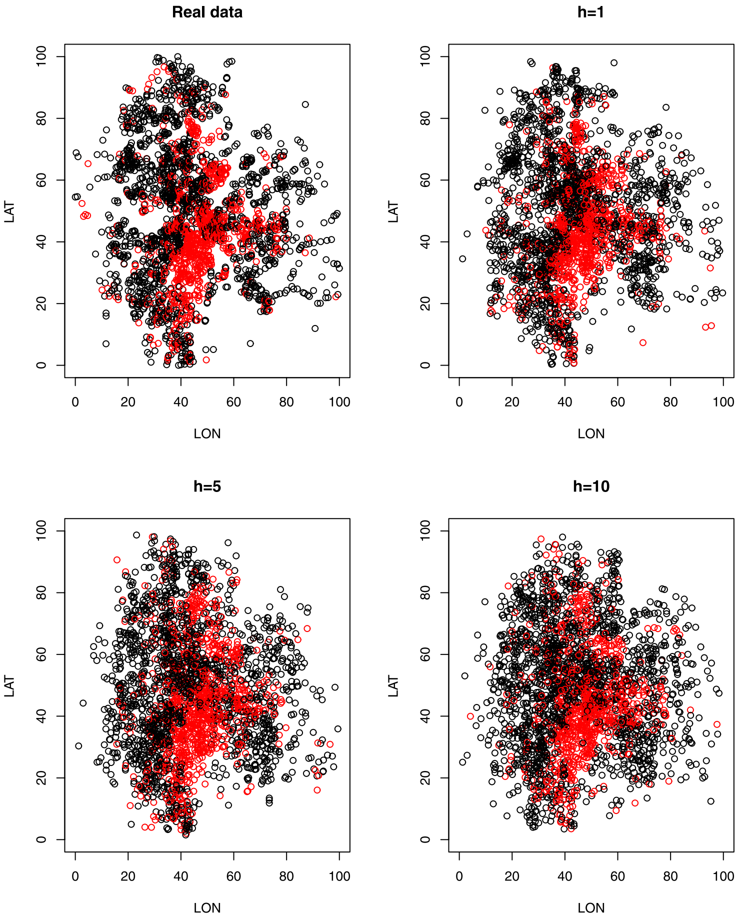

FIG. 2. *Plots of the observed data* (*upper left*) *and synthetic data sets for three levels of h. Red dots indicate locations of black people*, *and black dots indicate locations of white people. All values of* (λ, ϕ, R, A) *are synthesized.*

R_1 risks, because $\|\mathbf{s}_i^* - \mathbf{s}_i\|_2^2 \sim \frac{1}{2} R_{1,i}^2 \chi_2^2$, and, hence, $E(\|\mathbf{s}_i^* - \mathbf{s}_i\|_2^2) = R_{1,i}^2$. For comparisons, we repeat the analyses from Tables 4 and 5.

For the repeated sampling experiment, we add random noise to each location independently 100 times, thus creating 100 noise-infused data sets. For the noise infusion, four of the fourteen percentage-related estimands in Table 4 have MSE $> 3\%$. In contrast, when synthesizing geography only with $h = 1$, none of the percentage-related estimands have MSE $> 3\%$; these results are reported in the online supplement [Wang and Reiter (2011)]. For the age-related estimands, the MSEs are similar for synthetic and noise-infused data sets. Thus, for comparable levels of disclosure risks, adding random noise reduces the quality of inferences for the descriptive estimands relative to synthetic data.

For the regression analyses, we estimate the Bayesian spatial regression with one data set generated by adding random noise to geography only. The in-sample and out-of-sample misclassification rates are 0.32 and 0.38, respectively, for this noise-infused data set. We observed similar misclassification rates when repeating this analysis three more times. These misclassification rates are substantially larger than those for the corresponding synthetic data sets reported in the supplement (as well as those in Table 5), again suggesting that, for comparable risk levels, random noise does not preserve spatial relationships as well as synthetic data.

4. Concluding remarks. Although synthesizing geographies via modeling, such as the CART approach here, can preserve some spatial analyses, it does not preserve all of them. For example, two records close in space in the original data will not necessarily be close in space in the synthetic data, because their locations are independently generated from the response distribution. Additionally, simulated geographies may not preserve analyses when used to link the synthetic data with other data containing geography, since the simulated locations are conditionally independent of the variables in the linked data set that are not included in the synthesis model. Evaluating the impacts of synthetic geographies on linked analysis is a future extension of this research.

When synthesizing the nongeographic quasi-identifiers, we controlled for location as predictors in the model. An alternative approach is to simulate from hierarchical spatial models for point-referenced data, or perhaps from area-level models by aggregating locations [Banerjee, Gelfand and Carlin (2004)]. With large data sets, fitting spatial random effects models can be computationally challenging, although this can be overcome using approximations from the spatial statistics literature. Another strategy is to mask attribute data using spatial smoothing techniques [Zhou, Dominici and Louis (2010)]. We note that applying either of these approaches alone, that is, without simulating geography, leaves the original fine geography on the file, which may be too high of a disclosure risk. Evaluating the potential gains in disclosure risk and data usefulness of such strategies over the simple CART synthesizer for attributes utilized here is an area open for further theoretical and empirical investigation.

To implement the CART synthesizer, data stewards need to select the tuning parameters of the trees, that is, the minimum number of observations per leaf and the splitting criteria. These parameters control the size of the tree: increasing them results in smaller trees, and decreasing them results in larger trees. Based on our experience, we recommend that data stewards begin by setting the minimum deviance in the splitting criteria to a small number, like 0.0001 or even smaller, and requiring at least five records per leaf. These are typical default values for many applications and software routines for regression trees. The data steward then evaluates the disclosure risk and data utility associated with the synthetic data sets. If the risks are too high, the data steward can re-tune the parameters for the variables that are not sufficiently altered by the synthesis to grow smaller trees for those variables [Reiter (2005d)]. We did not prune the leaves further, as experiments with further pruning worsened the quality of the synthetic data sets without substantially improving the confidentiality protection. Growing larger trees can increase the quality of the synthetic data sets. However, it increases the time to run the synthesizer. Further, it can increase disclosure risks, for example, using trees with one observation per leaf reproduces the original data.

The CART synthesizer has appealing features: it handles continuous, categorical and mixed data; captures nonlinear relationships and complex interactions automatically; and runs quickly on large data sets. However, CART synthesizers can run into computational difficulties when categorical variables have many (e.g., >20) levels. Additionally, when some levels have low incidence rates in the data, the CART synthesizer can have difficulty preserving relationships involving those levels [Reiter (2005d)].

For simulation purposes, we illustrated the CART synthesizer using only $n =$ 2,670 records. This facilitated estimation of the spatial regressions with each of the resulting synthetic data sets. In extended investigations, we found that the CART synthesis process readily scaled for tens of thousands of mortality records. Other applications using CART synthesizers for nongeographic attributes [Reiter (2009), Drechsler (2011)] indicate that it can be applied in surveys of dimensions typical of many government surveys. When data stewards need to synthesize locations for a very large number, for example, millions, of records, a computationally convenient strategy is to partition the data into geographical strata of manageable size (tens of thousands of records), and simulate latitudes and longitudes (and attributes) by running the synthesizer independently within each stratum.

SUPPLEMENTARY MATERIAL

Computational details and further results (DOI: 10.1214/11-AOAS506SUPP; .pdf). Computational details for geography disclosure and identification risks in Sections 3.2.1 and 3.2.2; further analytical validity results; and results based on genuine cause of death.

REFERENCES

ARMSTRONG, M. P., RUSHTON, G. and ZIMMERMAN, D. L. (1999). Geographically masking health data to preserve confidentiality. *Stat. Med.* **18** 495–525.

BANERJEE, S., GELFAND, A. E. and CARLIN, B. P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.

BANERJEE, S., GELFAND, A. E., FINLEY, A. O. and SANG, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 825–848. MR2523906

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

CAIOLA, G. and REITER, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Trans. Data Priv.* **3** 27–42. MR2725418

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

DALENIUS, T. and REISS, S. P. (1982). Data-swapping: A technique for disclosure control. *J. Statist. Plann. Inference* **6** 73–85. MR0653248

DE'ATH, G. (2002). Multivariate regression trees: A new technique for modeling species environment relationships. *Ecology* **83** 1105–1117.

DRECHSLER, J. (2011). New data dissemination approaches in old Europe—Synthetic datasets for a German establishment survey. *J. Appl. Stat.* To appear.

DRECHSLER, J. and REITER, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases* (*LNCS* 5262) (J. Domingo-Ferrer and Y. Saygin, eds.) 227–238. Springer, New York.

DRECHSLER, J. and REITER, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *J. Amer. Statist. Assoc.* **105** 1347–1357. MR2796555

DUNCAN, G. T. and LAMBERT, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7** 207–217.

FEDERAL REGISTER (2000). Standards for privacy of individually identifiable health information—Final privacy rule. 45 C. F. R. Parts 160 and 164, Dept. Health and Human Services, Office of the Secretary, Washington, DC.

FIENBERG, S. E., MAKOV, U. E. and SANIL, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13** 75–89.

FIENBERG, S. E. and MCINTYRE, S. E. (2004). Data swapping: Variations on a theme by Dalenius and Reese. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and V. Torra, eds.) 14–29. Springer, New York.

FREEDMAN, D. A. (2004). The ecological fallacy. In *Encyclopedia of Social Science Research Methods* (M. Lewis-Beck, A. Bryman and T. F. Liao, eds.) **1** 293. Sage, Thousand Oaks, CA.

FULLER, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9** 383–406.

GOMATAM, S., KARR, A. F., REITER, J. P. and SANIL, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.* **20** 163–177. MR2183447

HEALTH AND RETIREMENT STUDY (2007). Data Description and Usage (2006 Core, Early, Version 2.0). Available at http://hrsonline.isr.umich.edu/meta/2006/core/desc/h06dd.pdf.

KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S. and ABOWD, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. Technical report, Center for Economic Studies Working Paper CES-WP-11-04, Census Bureau, Washington, DC.

LITTLE, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9** 407–426.

LITTLE, R. J. A., LIU, F. and RAGHUNATHAN, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (A. Gelman and X. L. Meng, eds.) 141–152. Wiley, New York.

MACHANAVAJJHALA, A., KIFER, D., ABOWD, J., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets practice on the map. In *IEEE* 24*th International Conference on Data Engineering* 277–286.

NATIONAL RESEARCH COUNCIL (2005). Expanding access to research data: Reconciling risks and opportunities. Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2007). Putting people on the map: Protecting confidentiality with linked social-spatial data. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data, Committee on the Human Dimensions of Global Change, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

REITER, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29** 181–189.

REITER, J. P. (2004a). New approaches to data dissemination: A glimpse into the future (?). *Chance* **17** 11–15. MR2061931

REITER, J. P. (2004b). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30** 235–242.

REITER, J. P. (2005a). Estimating identification risks in microdata. *J. Amer. Statist. Assoc.* **100** 1103–1113.

REITER, J. P. (2005b). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168** 185–205. MR2113234

REITER, J. P. (2005c). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *J. Statist. Plann. Inference* **131** 365–377. MR2139378

REITER, J. P. (2005d). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21** 441–462.

REITER, J. P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review* **77** 179–195.

REITER, J. P. and MITRA, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality* **1** 99–110.

ROBINSON, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review* **15** 351–357.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

SWEENEY, L. A. (2001). Computational disclosure control: A primer on data privacy protection. Ph.D. thesis, MIT, Cambridge, MA.

VANWEY, L. K., RINDFUSS, R. R., GUTTMAN, M. P., ENTWISLE, B. and BALK, D. L. (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proc. Natl. Acad. Sci. USA* **102** 15337–15342.

WANG, H. and REITER, J. (2011). Supplement to "Multiple imputation for sharing precise geographies in public use data." DOI:10.1214/11-AOAS506SUPP.

ZHOU, Y., DOMINICI, F. and LOUIS, T. A. (2010). A smoothing approach for masking spatial data. *Ann. Appl. Stat.* **4** 1451–1475. MR2758336

DEPARTMENT OF STATISTICS
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, SOUTH CAROLINA 29208
USA
E-MAIL: haowang@sc.edu

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: jerry@stat.duke.edu