Synthetic Data for Small Area Estimation

Joseph W. Sakshaug and Trivellore E. Raghunathan

University of Michigan, Ann Arbor MI 48104, USA

Abstract. Increasingly, researchers are demanding greater access to microdata for small geographic areas to compute estimates that may affect policy decisions at local levels. Statistical agencies are prevented from releasing detailed geographical identifiers in public-use data sets due to privacy and confidentiality concerns. Existing procedures allow researchers access to restricted geographical information through a limited number of Research Data Centers (RDCs), but this method of data access is not convenient for all. An alternative approach is to release fully-synthetic, public-use microdata files that contain enough geographical details to permit small area estimation. We illustrate this method by using a Bayesian Hierarchical model to create synthetic data sets from the posterior predictive distribution. We evaluate the analytic validity of the synthetic data by comparing small area estimates obtained from the synthetic data with estimates obtained from the U.S. American Community Survey.

Keywords: Synthetic Data, Small Area Estimation, Disclosure, Microdata.

1 Introduction

The demand for greater access to microdata for counties, municipalities, neighborhoods, and other small geographic areas is ever increasing [1]. Analysts require such data to answer important research questions that affect policy decisions at local levels. Statistical agencies regularly collect data from small areas, but are prevented from releasing detailed geographic identifiers due to the risk of disclosing respondent identities and their sensitive attributes.

Existing data dissemination practices for small geographic areas include: 1) releasing summary tables containing aggregate-level data only; 2) suppressing geographical details in public-use microdata files for areas that do not meet a predefined population threshold (e.g., 100,000 persons) and; 3) permitting access to restricted geographical identifiers through a limited number of Research Data Centers (RDCs). Although useful in some situations, none of these methods is likely to satisfy the various needs of researchers, students, policy-makers, and community planners, who are fueling the demand for small area estimates.

This article investigates a fourth approach that statistical agencies may implement to release more detailed geographical information in public-use data sets. The approach builds on the statistical disclosure control method, originally proposed by Rubin [2], of creating multiple synthetic populations conditional on the observed data and releasing samples from each synthetic population which comprise the public-use

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 162–173, 2010

data files. Valid inferences on a variety of estimands are obtained by analyzing each data file separately and combining the results using methods described in [3].

The synthetic data literature focuses on preserving statistics about the entire sample, but preserving small area statistics is not addressed. Statistics about small areas can be extremely valuable to data users, but detailed geographical identifiers are almost always excluded from public-use microdata sets. Significant theoretical and practical research on model-based small area estimation has led to a greater understanding of how small area data can be summarized (and potentially simulated) by statistical models [4,5]. The majority of this research involves the use of Bayesian Hierarchical models, which are used to "borrow strength" across related areas and to increase the efficiency of the resulting small-area estimates. The use of Bayesian Hierarchical models for multi-level imputation, and, particularly, for synthetic data applications, is rare [20,21,22].

Under a fully-synthetic design all variables are synthesized and few (if any) observed data values are released. This design offers greater privacy and confidentiality protection compared to synthesizing only a subset of variables [10], but the analytic validity of inferences drawn from the synthetic data may be poor if important relationships are omitted or misspecified in the imputation model. A less extreme approach involves synthesizing a partial set of variables or records that are most vulnerable to disclosure [6,7,8,9]. If implemented properly, this approach yields high analytic validity as inferences are less sensitive to the specification of the imputation model, but it may not provide the same level of protection as fully-synthetic data because the observed sample units and the majority of their data values are released to the public [10].

At the present time, statistical agencies have only released partially synthetic data files [11,12,13]. There are worthy reasons why fully-synthetic data may be more appropriate for small area applications. The most important reason is that full synthesis can offer stronger levels of disclosure protection than partial synthesis. Data disseminators are obligated by law to prevent data disclosures and may face serious penalties if they fail to do so. Hence, maintaining high levels of privacy protection takes precedence over maintaining high levels of analytic validity. This point is particularly important for small geographic areas, which may contain sparse subpopulations and higher proportions of unique individuals who are especially susceptible to re-identification. A secondary benefit of creating fully-synthetic data sets is that an arbitrarily large sample size may be drawn from the synthetic population, facilitating analysis for data users who would otherwise have to exclude or apply complicated indirect estimation procedures to areas with sparse (or nonexistent) sample sizes. Synthetic sample sizes may be deliberately chosen to facilitate the use of direct estimation methods and standard statistical software and ease the burden of analysis for data users.

In this article, we propose an extension to existing synthetic data procedures for the purpose of creating synthetic, public-use microdata sets for small geographic areas from which valid small area inferences may be obtained. A Bayesian hierarchical model is developed that accounts for the hierarchical structure of the geographical areas and "borrows strength" across related geographic areas. A sequential multivariate regression procedure [14] is used to approximate the joint distribution of the observed data and to simulate synthetic values from the resulting posterior predictive

distribution. We demonstrate how statistical agencies may generate fully-synthetic data for small geographic areas on a subset of data from the U.S. American Community Survey. Synthetic data is generated for several commonly used household- and person-level variables and their analytic validity is evaluated by comparing small area inferences obtained from the synthetic data with those obtained from the observed data. We do not evaluate the disclosure risk properties of the proposed synthetic data approach and leave this to future work.

2 Review of Fully Synthetic Data

The general framework for creating and analyzing fully synthetic data sets is described in [3] and [15]. Suppose a sample of size *n* is drawn from a finite population $\Omega = (X, Y)$ of size *N*, with $X = (X_i;\ i = 1,2, \ldots, N)$ representing the design or geographical variables available on all *N* units in the population, and $Y = (Y_i;\ i = 1,2, \ldots, N)$ representing the survey variables of interest, observed only for the sampled units. Let $Y_{obs} = (Y_i; i = 1,2, \ldots, n)$ be the observed portion of *Y* corresponding to sampled units and $Y_{nobs} = (Y_i, i = n+1, n+2, \ldots, N)$ be the unobserved portion of *Y* corresponding to the nonsampled units. The observed data set is $D = \{X, Y_{obs}\}$. For simplicity, we assume there are no item missing data in the observed survey data set, but methods exist for handling this situation [16].

Fully synthetic data sets are constructed based on the observed data *D* in two steps. First, multiple synthetic populations are generated by simulating $\left(Y_{nobs}^{(l)}; l = 1,2, \ldots M\right)$ for the nonsampled units using independent draws from the Bayesian posterior predictive distribution, $(Y_{nobs}|X, Y_{obs})$, i.e., conditional on the observed data *D*. Alternatively, one can generate synthetic values of *Y* for all *N* units based on the posterior predictive distribution of "future" or "super" populations $\left(Y_f|D\right)$, conditional on the observed data. This procedure ensures that the synthetic populations contain no real values of *Y*, thereby avoiding the release of any observed value of *Y*. Second, a random sample (e.g., simple random sample) of size n_{syn} is drawn from each of *M* synthetic populations. These sampled units comprise the public-use data sets that are released to, and analyzed by, data users.

From these publicly-released synthetic data sets, data users can make inferences about a scalar population quantity $Q = Q(X, Y)$, such as the population mean of *Y* or the population regression coefficients of *Y* on *X*. In each synthetic data set, the user estimates *Q* with some point estimator *q* and an associated measure of uncertainty *v*. Let $\left(q^{(l)}, v^{(l)}; l = 1,2, \ldots, M\right)$ be the values of *q* and *v* computed on the *M* synthetic data sets. We assume that these quantities are estimated based on a simple random sampling design. Under assumptions described in [3], the data user can obtain valid inferences for scalar *Q* by combining the $q^{(l)}$ and $v^{(l)}$ using the following quantities:

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)}/M \tag{1}$$

$$b_M = \sum_{l=1}^{M} \left(q^{(l)} - \bar{q}_M\right)^2 / (M-1) \tag{2}$$

$$\bar{v}_M = \sum_{l=1}^{M} v^{(l)} / M \tag{3}$$

where $\bar{q}_M$ is used to estimate Q, and

$$T_M = (1 + M^{-1}) b_M - \bar{v}_M \tag{4}$$

is used to approximate the variance of $\bar{q}_M$. A disadvantage of T_M is that it can be negative. Negative values generally can be avoided by making M and n_{syn} large. A more precise variance estimator that is always positive is outlined in [3]. Inferences for scalar Q are based on a normal distribution when $T_M > 0$, n, M, and n_{syn} are large. For moderate M, inferences can be based on t-distributions [17].

3 Creation of Synthetic Data Sets for Small Area Estimation

We adopt a Bayesian approach, using a hierarchical imputation model, to generate synthetic data for small area estimation. Hierarchical models have been used in several applications of small area estimation [18,19]; see [5] for a comprehensive review of design-based, empirical Bayes, and fully Bayesian approaches for small area estimation. Hierarchical models have also been used for imputation of missing data in multilevel data structures [20,21].

Our approach involves three stages. In the first stage, incremental regression models are fit using the observed data within small areas to approximate the joint conditional density of the set of variables to be synthesized. In the second stage, the joint sampling distribution of population parameters is approximated and the between-area variation is modeled by incorporating covariates from larger geographical areas (e.g., states). In the final stage, the population parameters are simulated by taking independent draws from the posterior predictive distribution and are used to generate the synthetic microdata.

In illustrating the modeling steps, we take a pragmatic approach by keeping the models relatively simple from a computational perspective. Despite the simplified presentation, the basic framework can potentially handle more sophisticated modeling approaches on a routine basis. Limitations of our approach and alternatives are discussed in Section 5.

3.1 Stage 1: Direct Estimates

For descriptive purposes, we introduce the following notation. We define small areas as *counties*, nested within *states*, which could be nested within an even larger area (e.g., region). Specifically, suppose that a sample of size n is drawn from a finite population of size N. Let n_{cs} and N_{cs} denote the respective sample and population

sizes for county $c = 1,2,\ldots,C_s$ within state $s = 1,2,\ldots,S$. Let $Y_{cs} = \left(Y_{ics,p}; i = 1,2,\ldots,n_{cs};\ p = 1,2,\ldots,P\right)$ represent the $n_{cs} \times P$ matrix of survey variables collected from each survey respondent located in county *c* and state *s*. Let $X_{cs} = \left(X_{ics,j}; i = 1,2,\ldots,n_{cs}, n_{cs}+1,\ldots,N_{cs};\ j = 1,2,..,J\right)$ represent the $N_{cs} \times J$ matrix of auxiliary or administrative variables known for every population member in a particular county and state. Although we consider synthesis of the survey variables Y_{cs} only, it is straightforward to synthesize the auxiliary variables X_{cs} as well.

A desirable property of synthetic data is that the multivariate relationships between the observed variables are maintained in the synthetic data, i.e., the joint distribution of variables is preserved. The first task is to specify the joint conditional distribution of the observed county-level variables to be synthesized $\left(Y_{cs,1}, Y_{cs,2}, \ldots, Y_{cs,P} | X_{cs,j}\right)$, where synthetic values are drawn from a corresponding posterior predictive distribution. Specifying and simulating from the joint conditional distribution can be difficult for complex data structures involving large numbers of variables representing a variety of distributional forms. Alternatively, one can approximate the joint density as a product of conditional densities [14]. Drawing synthetic variables from the joint posterior density $\left(Y_{cs,1}, Y_{cs,2}, \ldots, Y_{cs,P} | X_{cs,j}\right)$ can be achieved by sampling from$\left(Y_{cs,1} | X_{cs,j}\right)$, $\left(Y_{cs,2} | Y_{cs,1}, X_{cs,j}\right)$,…,$\left(Y_{cs,P} | Y_{cs,1}, \ldots, Y_{cs,P-1}, X_{cs,j}\right)$. In practice, a sequence of generalized linear models is fit on the observed county-level data where the variable to be synthesized comprises the outcome variable and any auxiliary variables or previously fitted variables are used as predictors, e.g., $Y_{ics,p+1} = \left(X_{ics}, Y_{ics,1,\ldots,p}\right)\beta_{cs} + \varepsilon_{ics}$. The choice of model (e.g., Gaussian, binomial) is dependent on the type of variable to be synthesized (e.g., continuous, binary). It is assumed that any complex survey design features are incorporated into the general linear models and that each variable has been appropriately transformed, if needed, to satisfy linear regression assumptions. After fitting each conditional density, estimates of the population regression coefficients $\hat{\beta}_{cs}$, the corresponding covariance matrix $\hat{V}_{cs}$, and the residual variance $\hat{\sigma}^2_{cs}$ are obtained and incorporated into the hierarchical structure described below in Section 3.2.

3.2 Stage 2: Sampling Distribution and Between-Area Model

In the second stage of synthetic data creation, the joint sampling distribution of the design-based county-level regression estimates $\hat{\beta}_{cs}$ (obtained from each conditional density in Stage 1) is approximated by a multivariate normal distribution,

$$\hat{\beta}_{cs} \sim MVN\left(\beta_{cs}, \hat{V}_{cs}\right),$$

where β_{cs} is a $(J+p) \times 1$ matrix of population regression coefficients and $\hat{V}_{cs}$ is the $(J+p) \times (J+p)$ corresponding covariance matrix estimated from the first stage. Further, the county-level population parameters β_{cs} are assumed to follow a multivariate normal distribution,

$$\beta_{cs} \sim MVN(\beta Z_s, \Sigma),$$

where $Z_s = \left(Z_{s,k}; k = 1,2,\ldots,K\right)$ is a $K \times 1$ matrix of state-level covariates, β is a $(J+p) \times K$ matrix of population regression coefficients, and Σ is a $(J+p) \times (J+p)$ covariance matrix. State-level covariates are incorporated into the hierarchical model in order to "borrow strength" from related areas. Prior distributions may be assigned to the unknown parameters β and Σ, but for ease of presentation, we assume that β and Σ are fixed at their respective maximum likelihood estimates (MLE), a common assumption in hierarchical models for small area estimation [18,23,24].

3.3 Stage 3: Generating Synthetic Populations within Small Areas

The ultimate objective is to generate synthetic populations within a small area using an appropriate posterior distribution. To this end, one can simulate the unknown population regression parameters β_{cs} specified in the hierarchical model described in Section 3.2. Based on standard theory of the normal hierarchical model [25], the posterior predictive distribution of the population regression coefficients is,

$$\tilde{\beta}_{cs} \sim MVN\left[\left(\hat{V}_{cs}^{-1} + \hat{\Sigma}_{\text{MLE}}^{-1}\right)^{-1}\left(\hat{V}_{cs}^{-1}\hat{\beta}_{cs} + \hat{\Sigma}_{\text{MLE}}^{-1}\hat{\beta}_{MLE}Z_s\right), \left(\hat{V}_{cs}^{-1} + \hat{\Sigma}_{\text{MLE}}^{-1}\right)^{-1}\right],$$

where $\tilde{\beta}_{cs}$ is a simulated vector of values for the vector of population regression coefficients β_{cs} . Simulating a synthetic variable $\tilde{Y}_{cs}$ for observed variable Y_{cs} can then be achieved by drawing $\tilde{Y}_{cs}$ from a parametric distribution with location and scale parameters $X_{cs}\tilde{\beta}_{cs}$ and σ_{cs}^2 , respectively, where σ_{cs}^2 may be drawn from an appropriate posterior predictive distribution $(\sigma_{cs}^2|Y_{cs}, X_{cs})$, or the maximum likelihood estimate $\hat{\sigma}^2$ obtained from Section 3.1 may be used. For example, to simulate a normally distributed variable $Y_{cs,1}$ one can draw $\tilde{Y}_{cs,1}$ from the distribution $N\left(X_{cs}\tilde{\beta}_{cs}, \hat{\sigma}^2\right)$. Generating a second (normally distributed) synthetic variable $\tilde{Y}_{cs,2}$ from the posterior predictive distribution $\left(Y_{cs,2}|Y_{cs,1}, X_{cs}\right)$ is achieved by drawing $\tilde{Y}_{cs,2}$ from $N\left(X_{cs}^*\tilde{\beta}_{cs}, \hat{\sigma}^2\right)$, where $X_{cs}^* = \left(X_{cs}, \tilde{Y}_{cs,1}\right)$. If the second synthetic variable is binary, then $\tilde{Y}_{cs,2}$ is drawn from $Bin\left(1, \hat{p}\left(X_{cs}^*\tilde{\beta}_{cs}\right)\right)$, where $\hat{p}\left(X_{cs}^*\tilde{\beta}_{cs}\right)$ is the predicted probability computed from the inverse-logit of $X_{cs}^*\tilde{\beta}_{cs}$. For polytomous variables, the same procedure is adapted to obtain posterior probabilities for each categorical response and the synthetic values are sampled from a multinomial distribution. This iterative process continues until all synthetic variables $\left(\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,P}\right)$ are generated. Multiple conditioning cycles can be implemented to minimize ordering effects [14]. The entire procedure is repeated M times to create multiple replicates of synthetic variables $\left(\tilde{Y}_{cs,1}^{(l)}, \tilde{Y}_{cs,2}^{(l)}, \ldots, \tilde{Y}_{cs,P}^{(l)}; l = 1,2,\ldots,M\right)$.

The entire synthetic populations may be disseminated to data users, or a simple random sample of arbitrary size may be drawn from each population and released. Stratified random sampling may be used if different sampling fractions are to be applied within the small areas. Inferences for a variety of small-area estimands Q_{cs} and large-area estimands Q_s or Q can be obtained using the combining rules in Section 2.

4 Evaluation of Synthetic Data for Small Area Inferences

In this section, we illustrate the above procedure on a subset of public-use microdata from the U.S. American Community Survey (ACS). We generate fully-synthetic data sets for relatively small geographic areas and evaluate the analytic validity of the resulting estimates. The data consist of seven household-level variables and eight person-level variables measured on 846,832 households and 2,093,525 persons during years 2005-2007. The variables, shown in Table 1, were chosen by researchers at the U.S. Census Bureau for this project. The smallest geographic unit that is identified in the ACS microdata is a Public-Use Microdata Area (PUMA). A PUMA is a census area containing around 100,000 people. All such areas are non-overlapping and are nested within a state. We restrict our sample to the Northeast region, which contains 9 states and 405 PUMAs.

We generate $M = 10$ fully-synthetic data sets for each "small area" (i.e., PUMA). To ensure that each synthetic data set contains ample numbers of households and/or persons within PUMAs, we create synthetic samples that are larger than the observed samples in each PUMA. Specifically, we generate synthetic sample sizes that are equivalent to 20% of the total number of U.S. households located within each PUMA based on the 2000 decennial census counts. This yielded a total synthetic sample size of 4,436,085 households for the Northeast region. Conceptually, this is equivalent to drawing a stratified random sample of households from each of $M = 10$ synthetic data populations.

The first survey variable to be synthesized is household size. Creating a household size variable will facilitate the creation of synthetic person-level variables in a later step. Because no administrative or other conditioning variables X_{cs} are available for

Table 1. List of ACS variables used in the synthetic data evaluation

Variable	Range/Categories
Household variables	
Household Size	1 - 20
Sampling weight	1 - 516
Total bedrooms	0 - 5
Electricity bill/mo.	1 - 600
Total rooms (excl. bedrooms)	1 - 7
Tenure	mortgage/loan, own free and clear, rent
Income	-33,998 – 2,158,100
Person variables	
Sampling weight	1 - 814
Gender	male, female
Education	16 categories, recoded less than high school, high school, some college, and college graduate
Ethnicity	Hispanic, non-Hispanic
Age	0 - 95
Race	9 categories, recoded white, black, other
Moved in last year	yes, no
Living in poverty	yes, no

this application, household size is simulated using a Bayesian Poisson-gamma model conditional on the observed household size variable with unknown hyperparameters estimated using maximum likelihood estimation. The remaining household-level variables are synthesized using the hierarchical modeling procedure described in Section 3. The sampling weights (both household and person) are included among the set of variables to be synthesized. State-level covariates Z_s, including population size (log-transformed), number of metropolitan, and number of micropolitan areas, are incorporated into the hierarchical model.

Linear regression models are used within PUMAs to obtain design-based estimates of population parameters for all numerical variables (with the previously noted exception of household size). Synthetic values of numerical variables are sampled from a Gaussian posterior predictive distribution. For binary variables, logistic regression models are used to obtain design-based population coefficients and corresponding synthetic values are sampled from a binomial posterior predictive distribution; the same procedure is applied to polytomous variables, which are broken up into a series of binary variables. To increase the stability of the design-based population regression coefficients, we apply a minimum sample size rule of $15 \cdot p$ within each PUMA. If a PUMA did not meet this minimum threshold, then nearby PUMAs were pooled together until the criterion was met.

Once the household variables were synthesized, the synthetic household data sets were transformed to person-level data sets and the person-level variables were synthesized conditional on the household variables. Taylor series linearization [26] was used to obtain design-based regression coefficients, accounting for the clustering of persons within households. To reduce the ordering effect of synthesizing the household variables first, we performed an additional conditioning cycle where each synthetic variable is conditioned on the full set of household- and person-level variables from the previous implementation.

4.1 Univariate Inferences for Small Areas

We evaluate the analytic validity of the synthetic data by comparing PUMA estimates obtained from the synthetic data with those obtained from the observed data for all 405 PUMAs. First, we compute basic univariate estimates, namely, means (or proportions) and standard errors for each PUMA; multivariate estimates are evaluated in Section 4.2. The sampling weights (synthetic and observed) are used to obtain adjusted point estimates and standard errors.

Table 2 presents the overall mean of the (weighted) PUMA means and standard errors obtained from the synthetic and observed data. The last column contains the slope (β_1) of the observed point estimates regressed against the synthetic point estimates for all 405 PUMAs. A slope equal to (or close to) 1 indicates a strong linear correspondence between the synthetic and observed estimates. On average, the synthetic PUMA means are generally within two standard errors of the observed PUMA means and the estimated slopes are reasonably close to the desired value, i.e., $(\beta_1 = 1)$. One exception is the Age variable, which is overestimated by the synthetic data. Upon inspection, the observed age variable has a bimodal distribution, which is not ideally simulated with a Gaussian distribution; this is a limitation of the parametric Bayesian simulation framework. Nonparametric strategies are likely to produce

Table 2. Mean of synthetic and observed PUMA means/proportions and standard errors and regression slope of actual means on the synthetic means for all 405 PUMAs

	Synthetic	Observed	
Variable	Mean (SE)	Mean (SE)	Slope (β_1)
Household variables			
Household Size	2.32 (0.03)	2.32 (0.03)	0.99
Sampling weight	33.70 (0.38)	33.96 (0.50)	0.99
Total bedrooms	2.70 (0.03)	2.66 (0.03)	1.02
Electricity bill/mo	115.53 (2.14)	114.19 (2.32)	1.04
Total rooms	3.05 (0.03)	2.99 (0.03)	1.01
Tenure			
Own free & clear	0.22 (0.01)	0.21 (0.01)	0.97
Rent	0.34 (0.01)	0.35 (0.01)	1.01
Mortgage/loan	0.44 (0.01)	0.44 (0.01)	0.98
Income	76144.90 (1576.65)	73658.80 (1780.44)	0.93
Person variables			
Sampling weight	35.80 (0.51)	35.49 (0.34)	0.98
Gender: male	0.49 (0.01)	0.49 (0.01)	0.57
Education			
Less than HS	0.33 (0.01)	0.33 (0.01)	0.99
HS graduate	0.24 (0.01)	0.24 (0.01)	0.98
Some college	0.20 (0.01)	0.20 (0.01)	0.95
College graduate	0.23 (0.01)	0.23 (0.01)	1.05
Hispanic ethnicity	0.12 (0.06)	0.11 (0.01)	1.16
Age	41.33 (0.34)	38.00 (0.38)	1.07
Race			
White	0.75 (0.01)	77.0 (0.01)	1.02
Black	0.13 (0.01)	0.11 (0.01)	1.03
Other	0.12 (0.01)	0.12 (0.01)	1.05
Moved in last year	0.11 (0.01)	0.12 (0.01)	1.13
Living in poverty	0.12 (0.01)	0.12 (0.01)	1.04

more desirable results. The Gender variable yields a relatively low slope value due to small variations in observed proportions of males across PUMAs. Overall, we believe the quality of the synthetic estimates is good relative to the observed data for obtaining univariate estimates. Aggregating the synthetic data to the state- and region-levels yielded estimates with similar correspondence to the observed data (not shown), indicating that synthetic data may be useful for producing valid estimates across multiple levels of geography.

4.2 Multivariate Inferences for Small Areas

Next we evaluate the analytic validity of the synthetic data for multivariate estimates. Table 3 presents summary results of two multiple regression models fitted within each PUMA. The first model regresses household income on the remaining household-level variables, and the second model regresses a recoded binary variable indicating college attendance (some college/college degree vs. less than high school/high school graduate) against all other person-level variables. Pseudo-maximum likelihood

Table 3. Mean of synthetic and observed PUMA regression coefficients and standard errors and regression slope of actual coefficients on the synthetic coefficients for all 405 PUMAs

	Synthetic	Observed	
	Coef (SE)	Coef (SE)	Slope (β_1)
Regression coefficients of household income (cube root):	25.37 (1.24)	26.05 (1.53)	0.93
Intercept	1.64 (0.19)	1.52 (0.24)	0.90
Household Size	1.27 (0.28)	1.20 (0.35)	0.98
Total bedrooms	0.94 (0.20)	0.89 (0.26)	0.98
Electricity bill/mo. (cube root)	1.34 (0.24)	1.35 (0.29)	0.99
Total rooms (excl. bedrooms)			
Tenure			
Own free & clear	-3.99 (0.61)	-4.13 (0.79)	1.08
Rent	-5.97 (0.70)	-6.15 (0.85)	0.97
Regression coefficients of college attendance on:			
Intercept	-0.70 (0.08)	-1.07 (0.09)	1.17
Gender: male	-0.09 (0.06)	-0.07 (0.08)	0.97
Hispanic ethnicity	-0.62 (0.15)	-0.53 (0.24)	1.07
Age	0.01 (0.01)	0.02 (0.01)	1.36
Race			
Black	-0.30 (0.14)	-0.18 (0.27)	1.01
Other	0.01 (0.14)	0.03 (0.17)	1.01
Poverty	-0.73 (0.11)	-0.80 (0.17)	0.87
Moved in last year	0.57 (0.13)	0.45 (0.14)	0.82

estimation is used to incorporate the relevant sampling weights [27]. The summary measures shown in Table 3 consist of overall means of the estimated regression coefficients and corresponding standard errors obtained from each PUMA. The last column represents the slope of the observed point estimates regressed against the synthetic point estimates for all 405 PUMAs. Because these models resemble those used earlier to approximate the joint posterior distribution of county-level parameters, we should expect close correspondence between the synthetic and observed point estimates, and more efficient synthetic data estimates. Indeed, we find that, on average, the synthetic point estimates correspond well with the observed point estimates in both direction and magnitude. The synthetic point estimates lie within about two standard errors of the observed point estimates, on average, and are generally more efficient than the observed point estimates. We find similar correspondence between the synthetic and observed data estimates when the data are aggregated to higher levels of geography (e.g., states, region).

5 Conclusions

This paper addresses an important data dissemination issue facing statistical agencies, which is how to meet the growing demand for high quality, public-use microdata for small geographic areas while protecting data confidentiality and privacy of respondents. These competing aims are likely to garner even more attention in the future as

research into small area effects and societal sensitivity to privacy and confidentiality continues to grow.

We propose a fully-synthetic data approach that utilizes a hierarchical model for creation of microdata for small geographic areas. The resulting data sets could conceivably be released to the public, along with additional data products that contain finer levels of data than are currently being released. The methodology is flexible, easy to implement, and can be straightforwardly adapted to a variety of data sources representing various geographical structures and variable types.

Results of the empirical evaluation suggest that valid small area inferences can be obtained from fully-synthetic data for basic descriptive and multivariate estimands. Although PUMAs are generally not considered to be "small areas," small-scale evaluations of the proposed synthetic data method using county-level data in the Census Research Data Center has yielded similarly valid results, with a slight loss in efficiency for the smallest areas.

One issue that was not empirically addressed in this paper is the level of disclosure protection offered by the synthetic data for small areas. Although there is evidence that fully-synthetic data offers better protection against disclosure than partially-synthetic data [10], this may not be true for small geographic areas or sparse subpopulations. Further research is needed to determine whether fully-synthetic data offers adequate levels of disclosure protection to be suitable for public release in a small area context.

In the evaluation, we did not assess the validity of the synthetic data for obtaining subgroup estimates or modeling interactions. Such estimates are particularly important to researchers studying subpopulations segregated within small geographic areas. Current work is underway to build complex relationships and interactions into the synthetic data generation process. An area for future work is the development of easy-to-implement, nonparametric approaches that weaken the dependence of the synthetic data inferences on the specification of the imputation models. In addition, further evaluations of the repeated sampling properties of the resulting synthetic data are needed to assess confidence interval coverage and data utility.

Acknowledgments. This research was supported by grants from the U.S. Census Bureau (YA-132309SE0354) and the U.S. National Science Foundation (SES-0918942).

References

1. Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel, D., Gardiner, C.: The case for small area microdata. J. Roy. Stat. Soc. A 168, 29–49 (2005)
2. Rubin, D.B.: Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. J. Off. Stat. 9, 461–468 (1993)
3. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple imputation for statistical disclosure limitation. J. Off. Stat. 19, 1–16 (2003)
4. Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P.: Small area statistics. Wiley, New York (1987)
5. Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)

6. Little, R.J.A.: Statistical analysis of masked data. J. Off. Stat. 9, 407–426 (1993)
7. Kennickell, A.B.: Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In: Alvey, W., Jamerson, B. (eds.) Record Linkage Techniques 1997, pp. 248–267. National Academy Press, Washington DC (1997)
8. Liu, F., Little, R.J.A.: Selective multiple imputation of keys for statistical disclosure control in microdata. In: Proceedings of the Joint Statistical Meetings, pp. 2133–2138. American Statistical Association, Blacksburg (2002)
9. Reiter, J.P.: Inference for partially synthetic public use microdata sets. Surv. Methodol. 29, 181–188 (2003)
10. Drechsler, J., Bender, S., Raessler, S.: Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv. 1(3), 105–130 (2008)
11. Rodriguez, R.: Synthetic data disclosure control for American Community Survey group quarters. In: Proceedings of the Joint Statistical Meetings, pp. 1439–1450. American Statistical Association, Salt Lake City (2007)
12. Abowd, J.M., Stinson, M., Benedetto, G.: Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program (2006)
13. Kinney, S.K., Reiter, J.P.: Making public use, synthetic files of the Longitudinal Business Database. In: Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference Proceedings, Istanbul, Turkey (2008)
14. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv. Methodol. 27, 85–95 (2001)
15. Reiter, J.P.: Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. J. Roy. Stat. Soc. A 168, 185–205 (2005)
16. Reiter, J.P.: Simultaneous use of multiple imputation for missing data and disclosure limitation. Surv. Methodol. 30, 235–242 (2004)
17. Reiter, J.P.: Satisfying disclosure restrictions with synthetic data sets. J. Off. Stat. 18, 531–544 (2002)
18. Fay III., R.E., Herriot, R.A.: Estimates of income for small places: an application of James-Stein procedures to Census data. J. Amer. Stat. Assoc. 74(366), 269–277 (1979)
19. Malec, D., Sedransk, J., Moriarity, C.L., LeClere, F.B.: Small area inference for binary variables in the National Health Interview Survey. J. Amer. Stat. Assoc. 92(439), 815–826 (1997)
20. Yucel, R.M.: Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. Phil. Trans. R. Soc. A 366(2008), 2389–2403 (1874)
21. Reiter, J.P., Raghunathan, T.E., Kinney, S.: The importance of modeling the sampling design in multiple imputation for missing data. Surv. Methodol. 32, 143–150 (2006)
22. Yu, M.: Disclosure Risk Assessments and Control. Doctoral Dissertation, University of Michigan (2008)
23. Datta, G.S., Fay, R.E., Ghosh, M.: Hierarchical and empirical Bayes multivariate analysis in small area estimation. In: Proceedings of the Bureau of the Census 1991 Annual Research Conference, pp. 63–79. U.S. Bureau of the Census, Washington (1991)
24. Rao, J.N.K.: Some recent advances in model-based small area estimation. Surv. Methodol. 25, 175–186 (1999)
25. Lindley, D.V., Smith, A.F.M.: Bayes estimates for the linear model. J. Roy. Stat. Soc. B 34(1), 1–41 (1972)
26. Binder, D.A.: On the variances of asymptotically normal estimators from complex surveys. Int. Stat. Rev. 51, 279–292 (1983)
27. Skinner, C.J., Holt, D., Smith, T.M.F.: Analysis of complex surveys. Wiley, Chichester (1989)