STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 341–368

GENERATING SYNTHETIC MICRODATA TO ESTIMATE SMALL AREA STATISTICS IN THE AMERICAN COMMUNITY SURVEY

Joseph W. Sakshaug[1], **Trivellore E. Raghunathan**[2]

ABSTRACT

Small area estimates provide a critical source of information used to study local populations. Statistical agencies regularly collect data from small areas but are prevented from releasing detailed geographical identifiers in public-use data sets due to disclosure concerns. Alternative data dissemination methods used in practice include releasing summary/aggregate tables, suppressing detailed geographic information in public-use data sets, and accessing restricted data via Research Data Centers. This research examines an alternative method for disseminating microdata that contains more geographical details than are currently being released in public-use data files. Specifically, the method replaces the observed survey values with imputed, or synthetic, values simulated from a hierarchical Bayesian model. Confidentiality protection is enhanced because no actual values are released. The method is demonstrated using restricted data from the 2005-2009 American Community Survey. The analytic validity of the synthetic data is assessed by comparing small area estimates obtained from the synthetic data with those obtained from the observed data.

Key words: counties, microdata, multiple imputation, data confidentiality.

1. Introduction

Demand for small area estimates is growing rapidly among a variety of stakeholders who use these data to advance the study of issues affecting local communities and the lives of their residents (Tranmer et al., 2005). Statistical agencies regularly collect data from small geographic areas and are therefore in a unique position to meet some of this demand. However, they are often prevented from releasing microdata for such areas because releasing detailed geographical identifiers for small areas may increase the risk of respondent re-identification

[1] Department of Statistical Methods, Institute for Employment Research, Germany. Program in Survey Methodology, University of Michigan, USA. E-mail: joesaks@umich.edu.

[2] Department of Biostatistics, University of Michigan, USA. E-mail: teraghu@umich.edu.

and inadvertent disclosure of confidential information (Mackie and Bradburn, 2000).

In order to minimize the risk of disclosure, statistical agencies commonly adopt one or more of the following data dissemination methods: 1) release summary tables that contain aggregate data for specific geographic areas (e.g., counties, census tracts, block groups); 2) suppress geographical details in public-use microdata sets for areas that fail to meet a predefined population threshold (e.g., 100,000) and; 3) release the unmasked confidential data set to data users via a secure data enclave or Research Data Center (RDC). Although these approaches are useful in many situations, each has limitations that preclude its ability to meet the growing demand for small area data that is being fuelled by researchers, analysts, policy-makers, and community planners.

For example, summary tables are useful tools for describing basic profiles of housing- and/or person-level characteristics for a wide variety of geographical areas, but their utility is limited to addressing complex scientific hypotheses that require customizable analytic approaches that are not feasible using existing aggregate data products. Releasing public-use microdata mitigates this issue by enabling users to perform customized analyses that go beyond the capabilities of published summary tables, but the suppression of identifiers for the smallest geographic areas limits their use for studying small area phenomenon. Releasing restricted microdata via a Research Data Center overcomes the limitations of the previous two by permitting users access to the full unmasked microdata, including all small area identifiers. In order to access data within an RDC, one must submit a research proposal, apply for special sworn status, pay a data usage fee, and travel to the nearest RDC facility. Unfortunately, these requirements are too restrictive for many analysts.

1.1. Synthetic data for small geographic areas

This article investigates a fourth approach that may permit statistical agencies to release more detailed geographical information in public-use data sets without compromising on data confidentiality. The approach extends the idea, originally proposed by Rubin (1993), of replacing the observed data values with multiply-imputed, or synthetic, values. The general idea is to treat the unobserved portion of the population as missing data to be multiply imputed using a predictive model fitted using the observed data. A random sample of arbitrary size is then drawn from each synthetic population which comprises the public-use data sets. Valid inferences are obtained by analyzing each synthetic data set separately and combining the point estimates and standard errors using combining rules developed by Raghunathan, Reiter, and Rubin (2003).

The synthetic data literature focuses on preserving statistics about the entire sample, but preserving small area statistics is usually ignored. Statistics about small areas can be extremely valuable to data users, but detailed geospatial information is almost always suppressed in public-use survey data. Research on model-based small area estimation has led to a greater understanding of how small area data can be summarized by statistical models (Platek et al., 1987; Rao, 2003), and such models could potentially be used for simulating small area microdata.

1.2. Fully synthetic versus partially synthetic data

There are two general synthetic data approaches: full synthesis and partial synthesis. Under a fully synthetic design all survey variables are synthesized and no real data is released. This approach provides the highest level of privacy and confidentiality protection (Drechsler, Bender, and Raessler, 2008), but the analytic validity of inferences drawn from the synthetic data may be poor if important relationships are omitted or mis-specified in the imputation model. Partial synthesis involves synthesizing a subset of variables or records that are pre-identified as being the most vulnerable to disclosure (Little, 1993; Kennickell, 1997; Liu and Little, 2002; Reiter, 2003). If implemented properly, this approach yields high analytic validity as inferences are less sensitive to misspecification of the imputation model. However, because the observed sample units and the majority of their data values are released to the public, it does not provide the same level of disclosure protection as full synthesis (Drechsler et al., 2008).

At the present time, the creation of partially synthetic data files is the most common application of synthetic data in large databases (Abowd, Stinson, and Benedetto, 2006; Rodriguez, 2007; Kinney et al., 2011). There are worthwhile reasons why fully synthetic data may be more appropriate for small area applications. Perhaps, the most important reason is that complete 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. Maintaining high levels privacy protection should take 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 cases that are especially susceptible to re-identification. A secondary benefit of fully synthetic data is that arbitrarily large sample sizes may be drawn from the synthetic populations, facilitating analysis for data users who would otherwise be forced to exclude areas with insufficient sample sizes, or apply complex indirect estimation procedures to compensate for the lack of sampled cases.

1.3. Organization of article

This article investigates an extension to Rubin's synthetic data method for the purpose of creating fully synthetic, public-use microdata sets for small geographic areas. A hierarchical Bayesian model is used that accounts for multiple levels of geography and "borrows strength" across related areas. A sequential multivariate regression procedure is used to approximate the joint distribution of the observed data, which is then used to simulate synthetic values from the posterior predictive distribution (Raghunathan et al., 2001). How statistical agencies may generate fully synthetic data for small geographic areas is demonstrated using a subset of restricted data from the American Community Survey. Synthetic data is generated for several commonly used household- and person-level variables and their analytic validity is assessed by comparing inferences obtained from the synthetic data with those obtained from the actual data. The disclosure risk properties of the synthetic data methodology are not assessed here and are left to future work. Limitations of the app roach and possible extensions are discussed in the final section.

2. Review of fully synthetic data

2.1. Creation of fully synthetic data sets

The general framework for creating and analyzing fully synthetic data sets is described in Raghunathan et al. (2003) and Reiter (2005). Suppose a sample of size n is drawn from a finite population $\Omega = (X, Y)$ of size N, with $X = (X_i; i = 1,2, \dots, N)$ representing design, geographical, or other auxiliary information available for all N units in the population, and $Y = (Y_i; i = 1,2, \dots, N)$ representing the survey variables of interest. It is assumed that there is no confidentiality concern over releasing information about X and synthesis of these auxiliary variables is not needed, but the method can be extended to synthesize these variables if necessary. Let $Y_{obs} = (Y_i; i = 1,2, \dots, n)$ be the observed portion of Y corresponding to sampled units and $Y_{nobs} = (Y_i; i = n + 1, n + 2, \dots, N)$ be the unobserved portion of Y corresponding to the nonsampled units. The observed data set is $D = (X, Y_{obs})$. For simplicity, assume there are no item missing data in the observed data, but methods exist for handling this situation (Reiter, 2004).

Fully synthetic data sets are constructed in two steps. First, M synthetic populations $P^{(l)} = \{(X, Y^{(l)}); l = 1,2, \dots, M\}$ are generated by taking independent draws from the Bayesian posterior predictive distribution of $f(Y_{nobs}|X, Y_{obs})$ conditional on the observed data D. Alternatively, one can generate synthetic values of Y for all N units to ensure that no observed values of Y are released. The number of synthetic populations M is determined based on the desired accuracy for synthetic data inferences and the risk of disclosing confidential information. A modest number of fully synthetic data sets (e.g., 5 or 10) are usually sufficient to ensure valid inferences (Raghunathan et al., 2003). In the second step, a random sample of size n_{syn} is drawn from each of the $l = 1,2, \dots, M$ synthetic data populations, $D^{(l)} = \left(x_i, y_i^{(l)}, i = 1,2, \dots, n_{syn}\right)$. The corresponding M synthetic samples $D_{syn} = \left(D^{(l)}; l = 1,2, \dots, M\right)$ comprise the public-use data sets, which are released to, and analyzed by, data users. In practice, the first step of generating complete synthetic populations is unnecessary and we only need to generate values of Y for units in the synthetic samples. The complete synthetic population setup is useful for theoretical development of combining rules.

2.2. Obtaining inferences from fully synthetic data sets

From the 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. Suppose the analyst is interested in obtaining a point estimate q and an associated measure of uncertainty v of Q from a set of synthetic samples D_{syn} drawn from the synthetic populations $P_{syn} = \left(P^{(l)}; l = 1,2,\dots,M\right)$ under simple random sampling. The values of q and v computed on the M synthetic data sets are denoted by $\left(q^{(l)}, v^{(l)}, l = 1,2,\dots,M\right)$.

Consistent with the theory of multiple imputation for item missing data (Rubin, 1987; Little and Rubin, 2002), combining inferences about $Q = Q(X, Y)$ from a set of synthetic samples D_{syn} is achieved by approximating the posterior distribution of Q conditional on D_{syn}. The suggested approach, outlined by Raghunathan et al. (2003), is to treat $\left(q^{(l)}, v^{(l)}; l = 1,2,\dots,M\right)$ as sufficient summaries of the synthetic data sets D_{syn} and approximate the posterior density $f\left(Q|D_{syn}\right)$ using a normal distribution with the posterior mean Q computed as the average of the estimates,

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

and the approximate posterior variance is computed as,

$$T_M = (1 + M^{-1}) b_M - v_m \tag{2}$$

where$\bar{v}_M = \sum_{l=1}^{M} v^{(l)} / M$ is the overall mean of the estimated variances across all synthetic data sets ("within variance") and $b_M = \sum_{l=1}^{M} \left(q^{(l)} - \bar{q}_M\right)^2 / (M-1)$ is the variance of $q^{(l)}$ across all synthetic data sets ("between variance").

Under certain regulatory conditions specified in Raghunathan et al. (2003), $\bar{q}_M$ is an unbiased estimator of Q and $b_M - v_m$ is an unbiased estimator of the variance of Q. The $\frac{1}{M} b_M$ adjusts for using only a finite number of synthetic data sets. It should be noted that the subtraction of the within imputation variance in T_M is due to the additional step of sampling units from the synthetic populations. Because of this extra sampling step, the between imputation variance contains the true between and nearly twice the amount of within variance needed to obtain an unbiased estimate of T.

When n, n_{syn}, and M are large, inferences for scalar Q can be based on normal distributions. For moderate M, inferences can be based on t-distributions with degrees of freedom $\gamma_M = (M-1)(1-r_m^{-1})^2$, where $r_m = (1+M^{-1})b_m/\bar{v}_M$, so that a $(1-\alpha)\%$ interval for Q is $\bar{q}_M \pm t_{\gamma_M}(\alpha/2)\sqrt{T_M}$ as described in Raghunathan and Rubin (2000). Extensions for multivariate Q are described in Reiter and Raghunathan (2007).

A limitation of the variance estimator T_M is that it can produce negative variance estimates. Negative values of T_M can generally be avoided by increasing Mor n_{syn}. Numerical routines can be used to calculate the integrals involved in the construction of T_M, yielding more precise variance estimates (Raghunathan et al., 2003). A simpler variance approximation that is always positive is shown in Reiter (2002).

3. Creation of synthetic data sets for small geographic areas

Hierarchical models have been used in several applications of small area estimation (Fay and Herriot, 1979; Malec et al., 1997). See Rao (2003) for a comprehensive review of design-based, empirical Bayes, and fully Bayesian approaches for small area estimation. Hierarchical models have also been used for multiple imputation of missing data in multilevel data structures (Reiter, Raghunathan, and Kinney, 2006; Yucel, 2008).

The approach considered here involves three stages. In the first stage, the joint density of the variables to be synthesized is approximated by fitting sequential regression models based on the observed data within each small area. In the second stage, the sampling distribution of the unknown regression parameters (estimated in the first stage) is approximated and the between-area variation is modelled using auxiliary information. In the third stage, the unknown regression parameters are simulated and used to draw synthetic microdata values from the posterior predictive distribution.

Two levels of geography are considered. For illustration, consider "small areas" as counties nested within states. In illustrating the approach, the models are kept relatively simple from a computational perspective to make the modelling practical. Despite the simplified presentation, the framework can be extended to handle more sophisticated modelling approaches.

3.1. Stage 1: Approximation of joint density via sequential regression

Suppose that a simple random sample of size n is drawn from a finite population of size N. Assuming units were sampled from each county, let n_{cs} and N_{cs} denote the respective sample and population sizes for county $c = (1,2,\ldots,C_s)$ nested 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. Here only the survey variables $Y_{cs,p}$ are synthesized, but it is straightforward to synthesize the auxiliary variables $X_{cs,j}$ as well.

A desirable property of the synthetic data is that the multivariate relationships among the observed variables are maintained in the synthetic data, i.e., the joint distribution of variables given the auxiliary information $f\left(Y_{cs,1}, Y_{cs,2}, \dots, Y_{cs,P} | X_{cs,j}\right)$ is preserved. 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 (Raghunathan et al., 2001). That is, the joint density $f\left(Y_{cs,1}, Y_{cs,2}, \dots, Y_{cs,P} | X_{cs,j}\right)$ can be factored into the following conditional densities: $f\left(Y_{cs,1} | X_{cs,j}\right)$, $f\left(Y_{cs,2} | Y_{cs,1}, X_{cs,j}\right)$,…,$f\left(Y_{cs,P} | Y_{cs,1}, \dots, Y_{cs,P-1}, X_{cs,j}\right)$. In practice, a sequence of generalized linear models are fit based on the observed county-level data where the variable to be synthesized comprises the outcome variable that is regressed on any auxiliary variables or previously fitted variables, e.g., $Y_{ics,1} = (X_{ics})\beta_{cs,1} + \varepsilon_{ics}$, $Y_{ics,2} = \left(X_{ics}, Y_{ics,1}\right)\beta_{cs,2} + \varepsilon_{ics}$,…,$Y_{ics,P} = \left(X_{ics}, Y_{ics,1}, Y_{ics,2}, \dots, Y_{ics,P-1}\right)\beta_{cs,P} + \varepsilon_{ics}$. 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 generalized linear models and that each variable has been appropriately transformed to satisfy modelling assumptions. After fitting each conditional density, the vector of regression parameter estimates $\hat{\beta}_{cs,p}$, the corresponding covariance matrix $\hat{V}_{cs,p}$, and the residual variance $\hat{\sigma}^2_{cs,p}$ are extracted from each of the P regression models and incorporated into the hierarchical model described below. $p = (1,2, \dots, P)$is used to index the set of parameters associated with the p^{th} synthetic variable of interest and the p^{th} regression model from which the direct estimates are obtained.

3.2. Stage 2: Sampling distribution and between-area model

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

$$\hat{\beta}_{cs,p} \sim MVN\left(\beta_{cs,p}, \hat{V}_{cs,p}\right) \tag{3}$$

where$\beta_{cs,p}$ is the $(J + p) \times 1$ matrix of unknown regression parameters and $\hat{V}_{cs,p}$ is the corresponding $(J + p) \times (J + p)$ estimated covariance matrix obtained from Stage 1. The unknown county-level regression parameters $\beta_{cs,p}$ are assumed to follow a multivariate normal distribution,

$$\beta_{cs,p} \sim MVN\left(\beta_p Z_s, \Sigma_p\right) \tag{4}$$

where$Z_s = \left(Z_{s,k}; k = 1,2,\dots,K\right)$ is a $K \times 1$ matrix of state-level covariates, β_p is a $(J+p) \times K$ matrix of unknown regression parameters, and Σ_p 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 $\beta_p$and Σ_p, but for computational simplicity it is assumed that β_p and Σ_p are fixed at their respective maximum likelihood estimates, a common assumption in hierarchical models for small area estimation (Fay and Herriot, 1979; Datta, Fay, and Ghosh, 1991; Rao, 1999). Details for obtaining the maximum likelihood estimates using the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) are provided in Appendix 1.

Based on standard theory of the normal hierarchical model (Lindley and Smith, 1972), the unknown regression parameters $\beta_{cs,p}$ can be drawn from the following posterior distribution,

$$\tilde{\beta}_{cs,p} \sim MVN\left[\left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_p^{-1}\right)^{-1}\left(\hat{V}_{cs,p}^{-1}\hat{\beta}_{cs,p} + \hat{\Sigma}_p^{-1}\hat{\beta}_p Z_s\right), \left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_p^{-1}\right)^{-1}\right] \qquad (5)$$

where$\tilde{\beta}_{cs,p}$ is a simulated vector of values for the unknown regression parameters $\beta_{cs,p}$.

3.3. Stage 3: Simulating from the posterior predictive distribution

The ultimate objective is to generate synthetic populations for each small area using an appropriate posterior predictive distribution. Simulating a synthetic variable $\tilde{Y}_{cs} = \left(\tilde{Y}_{lcs,p}; l = 1,2,\dots,N_{cs}; p = 1,2,\dots,P\right)$ for observed variable Y_{cs} for synthetic population unit $l = (1,2,\dots,N_{cs})$ is achieved by drawing, in sequential fashion, from the following posterior predictive distributions $f\left(\tilde{Y}_{cs,1}|X_{cs},\tilde{\beta}_{cs,1}\right)$, $f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1},X_{cs},\tilde{\beta}_{cs,1}\right)$, …, $f\left(\tilde{Y}_{cs,P}|\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,P-1},X_{cs},\tilde{\beta}_{cs,1}\right)$. For example, if the first variable to be synthesized $Y_{cs,1}$ is normally distributed then $\tilde{Y}_{cs,1}$ can be drawn from a normal distribution with location and scale parameters $X_{cs}\tilde{\beta}_{cs,1}$and $\sigma_{cs,1}^2$, respectively, where $\sigma_{cs,1}^2$ may be drawn from an appropriate posterior predictive distribution, or fixed at its maximum likelihood estimate $\hat{\sigma}_{cs,1}^2$ (obtainable from Stage 1). Generating a second (normally distributed) synthetic variable $\tilde{Y}_{cs,2}$ from the posterior predictive distribution $f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1},X_{cs},\tilde{\beta}_{cs,2}\right)$ is achieved by drawing $\tilde{Y}_{cs,2}$ from $N\left[\left(X_{cs},\tilde{Y}_{cs,1}\right)\tilde{\beta}_{cs,2},\sigma_{cs,2}^2\right]$, and so on up to $\tilde{Y}_{cs,P}{\sim}N\left[\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,P-1}\right)\tilde{\beta}_{cs,P},\sigma_{cs,P}^2\right]$. Alternatively, if the variable under synthesis $Y_{cs,p}$ is binary, then $\tilde{Y}_{cs,p}$ is drawn from a binomial distribution $Bin\left[1,\hat{p}\left\{\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,P}\right\}\right]$, where $\hat{p}\left\{\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,P}\right\}$ is the predicted probability computed from the inverse-logit of $\left\{\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,P}\right\}$. For polytomous variables, the same procedure is used to obtain posterior probabilities for each categorical response, which are then used to generate the synthetic values from a multinomial distribution. The iterative simulation process continues until all synthetic variables $\left(\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,P}\right)$ are generated. The procedure is repeated M times to create multiple populations of synthetic variables $\left(\tilde{Y}_{cs,1}^{(l)}, \tilde{Y}_{cs,2}^{(l)}, \ldots, \tilde{Y}_{cs,P}^{(l)}; l = 1,2, \ldots, M\right)$. In addition, the entire cycle may be repeated several times to minimize ordering effects (Raghunathan et al., 2001).

The complete synthetic populations may be disseminated to data users, or simple random samples 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 small areas. Inferences for a variety of estimands can be obtained using the combining rules in Section 2.2.

4. American Community Survey (2005-2009)

The proposed methodology is applied to a subset of restricted county-level microdata from the 2005-2009 American Community Survey (ACS), obtained from the Michigan Census Research Data Center. The ACS is an ongoing national survey that provides yearly estimates on a variety of topics, including income and benefits, health insurance coverage, disabilities, family and relationships, and others. The ACS collects information on persons living in housing units and group quarters facilities in all 3,141 counties in the United States. Data collection is conducted using a mixed-mode design. First, questionnaires are mailed to all sampled household addresses obtained from a Master Address File. Approximately six weeks after the questionnaire is mailed the Census Bureau attempts to conduct telephone interviews with all households that do not respond by mail. Following the telephone operation, a random sample is taken from the list of addresses where interviews have not been obtained and these addresses are visited by a field representative. Full details of the ACS methodology can be found in the technical documentation (U.S. Census Bureau, 2009).

Unlike the ACS public-use microdata files, the restricted data contain identifiers for all counties in the United States. For this application, we restrict the data to occupied housing units in the Northeast region. The Northeast region consists of 217 counties, all of which included households that completed ACS interviews. We use 5 years of restricted data to facilitate the disclosure review process and allow for the publication of estimates for all counties; the latter is not permitted with fewer years. Seven household- and seven person-level variables were selected for this analysis. The variables, shown in Table 1, were chosen by statisticians at the U.S. Census Bureau specifically for this project due to their common use among data users. Some variables (e.g., household tenure status, education, race) contained numerous categories. Ideally, each category would be preserved in the synthetic data; however, the decision was made to keep the number of categories at a minimum while maximizing the number of variables used in this small demonstration project. Thus, the few polytomous variables were recoded to reduce their number of categories. Transformations were applied to the continuous variables to meet normality assumptions during the model fitting and the synthetic data generation stages. After the synthesis was completed, the variables were transformed back to their original scales. The Census Bureau applies single imputation to missing ACS values in the restricted and public-use data files. We treat these imputations as actual observations in this application.

Table 1. List of ACS Variables Used in Synthetic Data Application. Variables Shown in the Order of Synthesis

Variable	Type	Range/Categories	Transformation
Household variables			
Household size	count	1 - 20	--
Sampling weight	continuous	1 - 201	log
Total bedrooms	count	0 - 5	--
Electricity bill/mo.	continuous	1 - 687	cube root
Total rooms (excl. bedrooms)	count	1 - 7	--
Income	continuous	0 – 3,999,996	cube root
Tenure	polytomous	recoded; mortgage/loan, own free and clear, rent	--
Person variables			
Sampling weight	continuous	1 - 341	log
Gender	binary	male, female	--
Education	polytomous	recoded; < 12 years, 12 years, 13-15 years, 16+ years	--
Hispanic ethnicity	binary	yes, no	--
Age	continuous	0 - 115	--
Race	polytomous	recoded; white, black, other	--
Living in poverty	binary	yes, no	--

Ten fully synthetic household- and person-level data sets were generated for each county. To ensure that each synthetic data set contained ample numbers of households and persons within each county, synthetic samples were created to be approximately equivalent to 20% of the total number of households based on the decennial census count. This yielded a total synthetic sample size of 3,963,715 households and 10,192,987 persons in the Northeast region.

The first survey variable to be synthesized was household size. Creating a household size variable facilitates the subsequent generation of synthetic person-level data. Household size was simulated using a Bayesian Poisson-Gamma model conditional on the observed household size variable with unknown hyperparameters fixed at their marginal maximum likelihood estimates obtained using the Newton-Raphson algorithm (see Appendix 2 for details). All subsequent variables were synthesized using the hierarchical modelling approach described in Section 3. State-level covariates $Z_S$that were incorporated into the hierarchical model included population size (2005 estimate: log-transformed) and the number of metropolitan and micropolitan areas. These covariates were obtained from the Census Bureau website.

For numerical variables (continuous, count), design-based estimates of regression parameters were obtained by fitting normal linear models within each county and synthetic values were drawn from the Gaussian posterior predictive distribution. For binary variables, logistic regression models were used to obtain the design-based parameter estimates and synthetic values were drawn from the binomial posterior predictive distribution. Logistic regression was also applied to polytomous variables after breaking them up into a series of conditional binary variables, estimating the propensity of a case belonging to a particular category versus all other categories, and using those propensities to predict case membership. We considered using multinomial regression for polytomous variables, but preliminary testing yielded convergence and stability problems for many counties. Therefore the decision was made to use the modified logistic regression approach. To increase the stability of the estimated regression coefficients, a minimum sample size rule of $10 \cdot p$ was applied within each county. If the target county did not meet this sample size threshold then nearby counties were pooled together until the criterion was met.

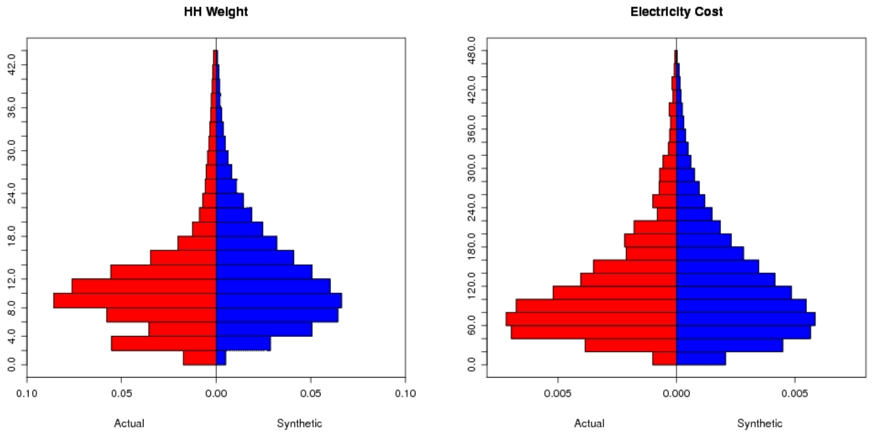
The household variables were synthesized first, followed by the person variables. After the synthetic household data sets had been created, they were converted to person-level data sets based on values of the synthetic household size variable. Taylor series linearization (Binder, 1993) was used to adjust the variances of the design-based regression estimates for the additional homogeneity due to persons clustered within households. To reduce the ordering effect induced by synthesizing the variables in a prescribed order, we repeat the entire synthetic data process 4 additional times, each time conditioning on the full set of synthetic variables generated from the previous implementations. Finally, it should be noted that the person-level variables were synthesized independently of the household-level variables. Although multiple imputation theory dictates that one should condition on all available information (Rubin, 1987), we found in preliminary runs that cycling between household- and person-level synthesis by aggregating person-level variables up to the household-level did not yield satisfactory inferences, possibly due to the non-standard distributions that the aggregation procedure produced. After applying several transformation procedures to the aggregated person-level variables, which did not significantly improve the imputations, we decided to keep the household and person levels separate for this demonstration project.

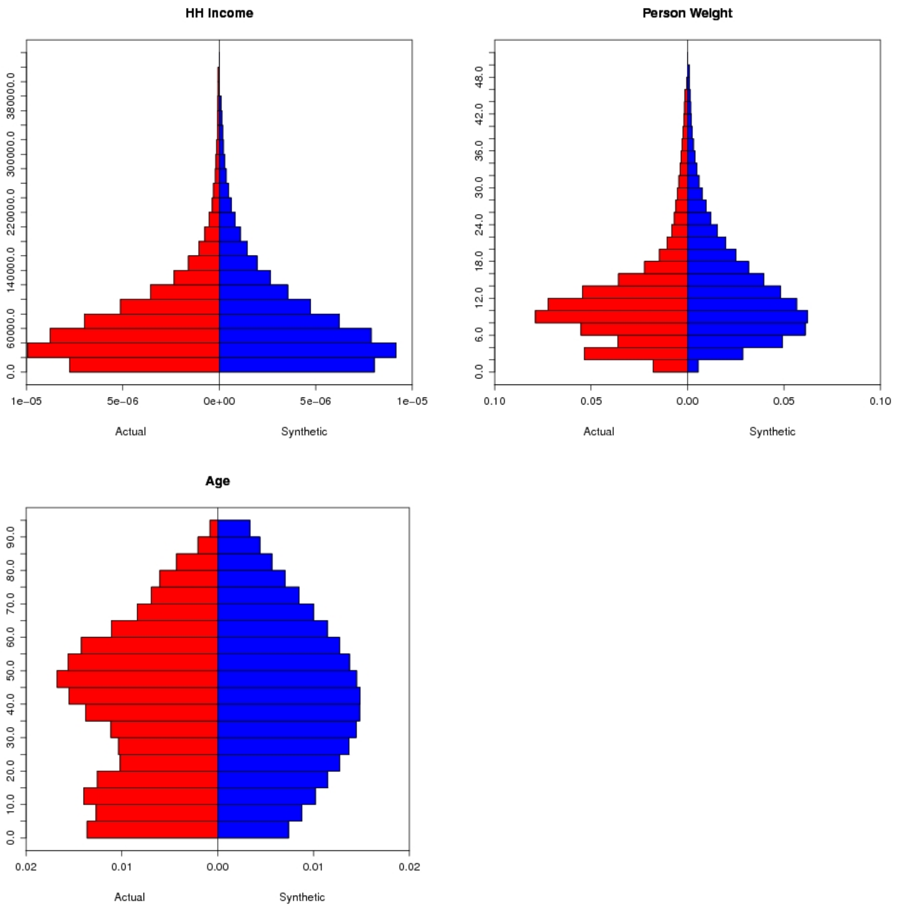
All results were reviewed and approved by the U.S. Census Bureau's Disclosure Review Board.

4.1. Validity of univariate estimates

Figure1 contains back-to-back histograms depicting the overall distributions for each continuous household- and person-level variable. The actual distributions are shown in the left panel and the synthetic distributions in the right panel. All variables are presented in their original scale. Visual comparisons show that for some variables, the synthetic data distribution corresponds to the actual data distribution reasonably well, but for others, the correspondence is poorer. Although the bulk of the distributions are generally maintained in the synthetic data, not every peak and valley is preserved. Those variables which do not follow a smooth parametric form tend to be most susceptible to a lack of correspondence. For example, the shape of the age distribution is bimodal denoting the highest frequency of people between the ages of 0-20 and 45-55. The synthetic age values, which are simulated from a normal distribution, fail to reflect the underlying bimodality. To a lesser degree, the sampling weight variables exhibit some bimodality at the left-most portion of their distributions, which is also not accounted for by the synthetic data. More sophisticated techniques, such as mixture modelling or nonparametric imputation may do a better job of preserving these non-standard distributional forms.

Figure 1. Back-to-Back Histograms of Actual (Left) and Synthetic (Right) Distributions for Continuous ACS Household- and Person-Level Variables in the Northeast Region.





While it is useful to compare synthetic and actual variable distributions for purposes of evaluation, data users are most interested in the validity of the small estimates obtained from the synthetic data. Table 2shows summary measures of univariate county-level estimands obtained from the synthetic and actual data. The first column contains the original set of ACS variables as well as recoded binary variables indicating overall income percentiles (50th, 75th and 90th) and specific subgroups (income x tenure; poverty x race/ethnicity). The second column shows the average county mean obtained from the synthetic and actual data, across all 217 counties. The third and fourth columns show the average standard deviation and standard error of the county means. The last column contains the intercept and slope values obtained from regressing the actual county means against the corresponding synthetic means. Intercept values close to zero and slope values close to one indicate strong correspondence between the synthetic and actual data estimates.

The synthetic data estimates, based on the original ACS variables, correspond roughly to the actual estimates, on average; out of the 9 household- and 12 person-level estimands, 5 and 10 of them yield synthetic point estimates that lie within two standard errors of the actual estimates, respectively, on average. The largest deviations occur for the tenure variable where the percentage of housing units being rented is overestimated by about two percentage points, on average, and the percentages of housing units owned free and clear and being financed through a mortgage or loan, are both underestimated in the synthetic data by about one and three percentage points, respectively, on average. These deviations are evident from examination of scatter plots of synthetic and actual county-level estimates (not shown, but available upon request). Similar over- and under-estimation effects appear in estimates of the other polytomous variables (education, race), but to a lesser extent. The cause of these effects is likely driven by two joint factors. The overestimation is likely due to the pooling of nearby counties to facilitate model fit for target counties that contained insufficient numbers of rented housing units; the rarest of the three membership categories. For the affected counties, the act of pooling at the estimation stage yields a higher rate of rented housing units in the synthetic data, which is closer to the population average. The underestimation in the other tenure estimates is driven by the fact that rental status was the first tenure category to be simulated, followed by ownership (conditional on not being rented) and mortgage/loan status (conditional on not being rented or owned). A consequence of this step-by-step conditional simulation approach is that the higher rates of rented housing units generated for the areas with inadequate samples sizes are offset by lower rates of ownership and mortgage/loan status for these smaller areas.

Aside from the positive/negative deviations among the polytomous estimates, the other estimates, based on continuous and binary ACS variables, appear to be reasonably valid as indicated by the diagnostic measures in Table 2. Many of the estimands yield intercept and slope values for the linear regression of actual county means against the synthetic means that are close to zero and one, respectively, indicating good correspondence between the actual and synthetic estimates. However, some of the continuous variables including electricity bill amount, household income, and, especially, age, yield larger deviations from the ideal intercept and slope values. The largest deviation occurs for the age estimates, which are likely due to the aforementioned bimodality of the age distribution that is reflected poorly in the synthetic data. The resulting synthetic county-level age estimates tend to be biased upward, particularly, for the counties with the highest average ages.

The validity of the percentile and subgroup estimates is mixed. The percentage of households with incomes exceeding the 50th percentile in the synthetic data corresponds closely to the actual percentages, on average. However, the estimates based on the 75th and 90th percentiles are higher in the synthetic data by about 1.5-2.0 percentage points, on average. Scatterplots of the county-level percentile means (not shown, but available upon request) indicate that the correspondence between synthetic and actual means becomes poorer as the percentile increases. Almost all of the income and poverty subgroup means lie within 1-2 standard errors of their corresponding actual means, on average. However, a positive and negative bias can be seen for synthetic estimates of mean income among mortgaged and rented housing units from scatterplots (not shown, but available upon request); a result that is likely due to the aforementioned under- and over-estimation of these tenure variables in the synthetic data, respectively.

A few remarks can be made about the uncertainty of the synthetic estimates. Based on multiple imputation theory, we would expect the synthetic standard deviations to be approximately the same and the standard errors to be larger than the actual standard deviations and standard errors, respectively, on average. This expectation is confirmed for some, but not all estimates. In most cases, the synthetic data standard deviations are close to their actual data counterparts. A particular exception is age, which yields larger standard deviations in the synthetic data, on average, due to the aforementioned bimodal age distribution, which is smoothed over in the synthetic data causing more age values to lie further away from the mean. On average, about half of the synthetic standard errors is equal to or greater than the corresponding actual standard errors. Estimates of income tend to have smaller standard errors in the synthetic data, on average, as a result of outlying observations being less preserved in the synthetic data. Moreover, the underestimated variances could be caused by misspecification of the imputation model and/or poor choice of transformation for preserving the tail-end of the distribution in the synthetic data, a problem which has been highlighted in earlier research on the estimation of imputed totals in skewed populations (Rubin, 1983). Another possible source of variation not accounted for in the synthetic data is due to the fact that the hyperparameters were fixed at their maximum likelihood estimates (see Section 3.2), rather than being randomly drawn from an *a priori* distribution.

Table 2. Summary Measures of Actual and Synthetic County Means

	Avg. Mean		**Avg. Standard Deviation**		**Avg. Standard Error of Mean**		**Regression of Actual Means on Synthetic Means**	
	Actual	Synthetic	Actual	Synthetic	Actual	Synthetic	Intercept	Slope
Household variables								
Household size	2.12	2.12	1.46	1.45	0.02	0.01	0.02	0.99
Sampling weight	9.99	10.20	7.21	7.04	0.11	0.11	0.01	0.98
Total bedrooms	2.88	2.82	0.96	1.09	0.02	0.01	0.15	0.97
Electricity bill/mo.	118.89	119.37	78.72	78.33	1.25	1.10	9.90	0.91
Total rooms	3.23	3.18	1.19	1.28	0.02	0.02	0.09	0.99
Income	67983.9	67382.4	68481.3	54081.9	1067.3	692.6	4681.7	0.94
Tenure (%)								
Mortgage/loan	49.00	47.03	49.38	49.30	0.82	0.74	0.04	0.95
Own free & clear	31.12	30.37	45.53	44.97	0.77	0.72	0.05	0.85
Rent	19.88	22.60	38.86	41.00	0.63	0.63	-0.05	1.09

Table 2. Summary Measures of Actual and Synthetic County Means (cont.)

	Avg. Mean		**Avg. Standard Deviation**		**Avg. Standard Error of Mean**		**Regression of Actual Means on Synthetic Means**	
	Actual	Synthetic	Actual	Synthetic	Actual	Synthetic	Intercept	Slope
Recoded variables								
Income > 50th pctile,%	44.65	44.56	48.24	48.19	0.80	0.56	0.01	0.97
Income > 75th pctile,%	19.34	21.49	37.34	38.69	0.59	0.43	-0.00	0.91
Income > 90th pctile,%	6.78	8.38	22.96	24.58	0.35	0.24	0.56	0.74
Income (Mortgage=1)	84667.0	86992.6	69019.2	58960.1	1536.0	1195.3	5460.0	0.91
Income (Own=1)	61076.6	60456.9	76053.1	45083.6	2132.8	1232.7	1717.0	0.98
Income (Rent=1)	38844.5	36921.9	37759.4	32527.3	1436.0	1166.5	3480.0	0.99
Person variables								
Sampling weight	10.27	10.67	7.59	8.02	0.08	0.14	-0.09	0.97
Gender (%)	48.63	48.63	49.97	49.97	0.53	0.44	0.04	0.91
Education (%)								
< 12 years	31.48	31.67	46.31	46.31	0.49	0.39	0.09	0.71
12 years	28.34	27.74	44.40	44.06	0.48	0.57	0.01	0.97
13-15 years	20.33	20.25	40.11	40.04	0.43	0.50	0.01	0.96
16+ years	19.85	20.35	38.72	39.14	0.40	0.51	-0.01	1.00
Hispanic (%)	3.85	4.23	15.72	16.99	0.14	0.26	-0.00	1.00
Age	40.89	41.16	22.98	30.34	0.25	0.27	22.02	0.46
Race (%)								
White	92.21	91.34	22.17	24.08	0.20	0.36	0.01	1.00
Black	3.55	4.01	14.54	16.26	0.13	0.26	-0.01	1.00
Other	4.24	4.65	14.54	18.61	0.16	0.27	-0.00	1.00
Poverty (%)	8.65	9.04	27.54	28.13	0.30	0.53	-0.00	1.00
Recoded variables								
Poverty (White=1; %)	7.93	8.19	26.41	26.84	0.30	0.51	-0.00	1.00
Poverty (Black=1; %)	20.48	21.30	36.86	37.03	4.62	3.52	-0.01	1.01
Poverty (Other=1; %)	16.62	17.84	35.37	36.07	2.96	4.38	0.01	0.87
Poverty (Hispanic=1; %)	19.92	21.11	37.08	37.96	3.52	5.54	-0.01	0.98

4.2. Validity of multivariate estimates

The next set of analyses examine the analytic validity of synthetic multivariate estimates obtained from multiple regression models. Table 3 shows average coefficient estimates (and their standard errors) for two regression models fit within each county. The first model fits a household-level linear regression of income (cube root) on the remaining ACS household covariates, and the second model fits a person-level logistic regression of poverty status on the remaining person covariates. Both models yield coefficient estimates based on the synthetic data that closely resemble those based on the actual data. Nearly all of the synthetic data coefficient estimates lie within one standard error of their corresponding actual data estimates, on average. Scatterplots of the synthetic and actual county regression coefficients (not shown, but available upon request) show that the synthetic data county estimates are in agreement with the actual county estimates as the points lie about the 45 degree line. However, there are clear biases associated with some coefficients, particularly, those associated with tenure variables that have already been shown to be affected by biases in the synthetic data. The standard errors of the synthetic data estimates appear to be on par, and in some cases, twice as large as those of the actual data estimates. In summary, the multivariate relationships examined here appear to be reasonably valid in the synthetic data. This is a reassuring result given that these relationships were explicitly accounted for in the synthetic data generation models.

Table 3. Summary Measures of Actual and Synthetic Linear and Logistic County Regression Coefficients

	Avg. Beta Coefficient		**Avg. Standard Error of Beta Coefficient**	
Linear regression of household income (cube root) on household-level covariates	Actual	Synthetic	Actual	Synthetic
Intercept	24.34	24.26	1.11	1.09
Household size	1.52	1.44	0.14	0.14
Sampling weight	-0.04	-0.05	0.24	0.26
Total bedrooms	1.15	1.23	0.19	0.18
Electricity bill/mo.	0.99	1.04	0.18	0.17
Total rooms	1.25	1.26	0.14	0.13
Tenure				
Mortgage/loan	Ref	Ref	Ref	Ref
Own free & clear	-3.47	-3.05	0.37	0.34
Rent	-6.01	-6.84	0.44	0.47
	Avg. Beta Coefficient		**Avg. Standard Error of Beta Coefficient**	
Logistic regression of poverty status on person-level covariates	Actual	Synthetic	Actual	Synthetic
Intercept	-2.39	-2.32	0.16	0.24
Sampling weight	0.25	0.25	0.07	0.10
Gender: Male	-0.33	-0.34	0.08	0.08
Education				
<12 years	Ref	Ref	Ref	Ref
12 years	-0.36	-0.35	0.12	0.13
13-15 years	-0.62	-0.63	0.13	0.15
16+years	-1.52	-1.59	0.18	0.30
Hispanic	0.36	0.27	0.29	0.63
Age	-0.00	0.01	0.00	0.07
Race				
White	Ref	Ref	Ref	Ref
Black	0.28	0.22	0.34	0.87
Other	0.41	0.41	0.25	0.56

5. ACS simulation

This section evaluates the repeated sampling properties of small area inferences drawn from the synthetic data based on a simulation study. In this simulation, we use public-use ACS microdata for the Northeast region for years 2005-2007. The smallest geographical unit in the public-use microdata is a Public-Use Microdata Area (PUMA). PUMAs are defined as areas which contain at least 100,000 persons. In many cases, PUMAs overlap exactly with counties with the exception of very large counties, which are split into multiple PUMAs, and very small counties, which are combined with nearby counties to form a single PUMA. There are 405 PUMAs located in the Northeast region. For this simulation study, the ACS data is treated as a population from which subsamples are drawn. 500 stratified random subsamples are drawn from each PUMA with replacement. Each subsample accounts for approximately 30% of the total sample in each PUMA. Each ACS subsample is used as the basis for constructing a synthetic population from which 100 synthetic samples are drawn. This resulted in a total of 50,000 synthetic data sets.

Two types of inferences can be obtained from the synthetic data: conditional and unconditional. Conditional synthetic inferences are obtained from synthetic samples that are based on a single observed sample drawn from the population. This is the situation that most commonly occurs in practice, where a survey is carried out on a single population-based sample and the synthetic data is generated conditional on that sample. Unconditional inferences are obtained from synthetic samples that are based on multiple, or repeated, population-based samples. Obtaining unconditional inferences is not feasible in practice but is possible in the simulation study considered here.

To obtain conditional inferences, 500 sets of 10 synthetic samples are randomly selected (with replacement) from each of the 100 synthetic samples generated conditional on each of the 500 ACS subsamples. For each set of 10 synthetic samples, a synthetic estimate and associated 95% confidence interval are obtained for each variable in each PUMA using the combining rules of Section 2.2. To obtain unconditional inferences, 100 sets of 10 synthetic samples are randomly selected with replacement across each of the 100 ACS subsamples and point estimates and associated confidence intervals are again obtained using the relevant combining rules.

We use two evaluative measures to assess the validity of the synthetic data estimates. The first one is confidence interval coverage (CIC). For conditional inference, CIC is defined as the proportion of times that the synthetic data confidence interval, computed at the 0.05 level,$\left[L_{\hat{q}_M,syn}, U_{\hat{q}_M,syn}\right]$ contains the actual estimate $\hat{y}_{act}$:

$$Q_{CIC} = I\left(\hat{y}_{act} \in \left[L_{\hat{q}_M,syn}, U_{\hat{q}_M,syn}\right]\right)$$

where $I(\cdot)$ is an indicator function. $Q_{CIC} = 1 \text{if} L_{\hat{q}_M,syn} \leq \hat{y}_{act} \leq U_{\hat{q}_M,syn}$ and $Q_A = 0$ otherwise.

For unconditional inference, the only difference is that the CIC is calculated as the proportion of times that the synthetic data confidence interval contains the “true” population value Y_{pop}, i.e., $L_{\hat{q}_M,syn} \leq Y_{pop} \leq U_{\hat{q}_M,syn}$.

The second evaluative measure is referred to as the confidence interval overlap (CIO; Karr et al., 2006). CIO is defined as the average relative overlap between the synthetic and actual data confidence intervals. For every estimate the average overlap is calculated as,

$$Q_{CIO} = \frac{1}{2}\left(\frac{U_{over}-L_{over}}{U_{act}-L_{act}} + \frac{U_{over}-L_{over}}{U_{syn}-L_{syn}}\right),$$

where U_{act} and L_{act} denote the upper and the lower bound of the confidence interval for the actual estimate $\hat{y}_{act}$, U_{syn} and L_{syn} denote the upper and the lower bound of the confidence interval for the synthetic data estimate $\hat{q}_M$, and U_{over} and L_{over} denote the upper and lower bound of the overlap of the confidence intervals from the original and synthetic data for the estimate of interest. Q_{CIO}can take on any value between 0 and 1. A value of 0 means that there is no overlap between the two intervals and a value of 1 means that the synthetic interval completely covers the actual interval. Calculating the confidence interval overlap is only possible for conditional inferences. This measure yields a more accurate assessment of data utility in the sense that it accounts for the significance level of the estimate. That is, estimates with low significance might still have a high confidence interval overlap and therefore a high data utility even if their point estimates differ considerably from each other.

5.1. Validity of univariate estimates

Table 4 shows the average confidence interval coverage (CIC) and confidence interval overlap (CIO) across all PUMAs for univariate household-level estimands. The conditional CIC is high for non-recoded estimates ranging from 0.86-0.99. The income by tenure subgroup estimates also yield relatively high conditional CIC values (range: 0.89-0.97). The CIC values for income percentile estimates do not fare as well as they tend to decline monotonically as the percentiles increase. The same general trend is observed for the conditional CIO values, which closely resemble the CIC values. Regarding the unconditional inferences, the CIC values tend to be slightly higher than the corresponding values obtained from the conditional evaluation. The actual CIC values, obtained from the actual ACS subsamples, tend to be very close to the synthetic CIC values, if not slightly higher, except for the aforementioned percentile estimates which demonstrate weaker coverage for the most extreme percentiles.

Table 4. Simulation-Based Confidence Interval Results for PUMA Means

	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
Household variables				
Household size	0.99	0.97	0.98	0.98
Sampling weight	0.95	0.99	0.99	0.98
Bedrooms	0.89	0.87	0.93	0.98
Electricity cost/mo.	0.86	0.87	0.91	0.98
Rooms	0.97	0.93	0.98	0.98
Household income	0.90	0.91	0.94	0.98
Tenure				
Own free & clear	0.93	0.92	0.96	0.98
Rent	0.94	0.96	0.96	0.98

Table 4. Simulation-Based Confidence Interval Results for PUMA Means (cont.)

Recoded variables				
Income > 50thpctile	0.89	0.92	0.94	0.98
Income > 75thpctile	0.71	0.71	0.80	0.98
Income > 90thpctile	0.52	0.60	0.62	0.97
Income (Mortgage=1)	0.89	0.88	0.94	0.97
Income (Own=1)	0.91	0.98	0.96	0.96
Income (Rent=1)	0.97	0.93	0.99	0.96

5.2. Validity of multivariate estimates

Multivariate simulation results are shown in Table 5. This table shows average CIC and CIO values for regression coefficient estimates obtained within each PUMA from a linear regression of income (cube root) on household-level covariates. The conditional CIC and CIO values are high and range from 0.93-0.99 and 0.90-0.98, respectively, indicating good analytic validity for these multivariate statistics. The unconditional CIC values range from 0.85-0.92, which are slightly below the actual CIC values obtained from the observed data (0.98). The lowest unconditional CIC values (0.85 and 0.87) are associated with the household tenure categories. Given that the analytic model being evaluated here is one of the same models used during the synthetic data generation process, it is not surprising that the analytic validity of the estimates is generally high. Overall, we believe this result is reassuring and underscores the importance of ensuring that the models used during the imputation process sufficiently overlap with the analytic models of interest.

Table 5. Simulation-Based Confidence Interval Results for PUMA Regression Coefficients

Linear regression of income (cube root) on	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
Intercept	0.98	0.97	0.92	0.98
Household size	0.98	0.95	0.91	0.98
Sampling weight	0.99	0.97	0.92	0.98
Total bedrooms	0.98	0.98	0.91	0.98
Electricity bill/mo.	0.99	0.97	0.91	0.98
Total rooms	0.98	0.97	0.92	0.98
Tenure				
Mortgage/loan	Ref	Ref	Ref	Ref
Own free & clear	0.95	0.90	0.87	0.98
Rent	0.93	0.96	0.85	0.98

6. Conclusions

Data users are increasingly interested in producing small area estimates, but statistical agencies are prevented from releasing these data due to disclosure concerns. In this article, a synthetic data methodology for generating and disseminating public-use microdata for small geographic areas was evaluated using restricted data from the U.S. Census Bureau. Compared with current practices of disseminating detailed geographical data, the synthetic data framework offers data users the flexibility of performing their own customizable geographic analyses using data that can presumably be released to the public without restriction.

The empirical evaluations show that the synthetic data generated from a Bayesian hierarchical model yields generally valid univariate and multivariate county-level estimates and repeated sampling properties. However, limitations of the method were apparent when simulating synthetic data for non-standard distributions and for polytomous variables when sample size limitations required pooling of nearby counties. Such limitations can potentially be overcome with more sophisticated modelling approaches, such as nonparametric imputation or mixture modelling, which was beyond the scope of this demonstration project. In addition, the "empirical" Bayesian approach considered here by fixing the hyperparameters at their maximum likelihood estimates may have underestimated the uncertainty of the synthetic data estimates, resulting in smaller standard errors and narrower confidence intervals. Although some underestimation of uncertainty might be welcomed in fully-synthetic data applications where standard errors are expected to be much higher relative to the observed standard errors, a more principled approach that accounts for all sources of variation might be viewed more favourably by sceptical data users.

Several extensions of this work are currently being considered. The preservation of skewed and non-standard distributions is an important issue that will need to be addressed prior to pubic release of synthetic small area microdata. Parametric modelling approaches are inherently limited in real-world applications where many of the most commonly used variables do not follow a smooth distributional form. The use of transformations to achieve normality is one possible solution; however, such transformations are not always effective for some types of distributions (e.g., bimodal). One must also consider the possibility that the same transformation might not work in all small areas. In this application, a single transformation was applied across all counties based on the overall distribution. Incorporating a tuning parameter in the hierarchical modelling approach that accounts for distributional differences across small areas might yield higher quality synthetic data and small area estimates with greater analytic validity. Another possible extension of this work is complex sample surveys. Although the ACS does not employ a complex sample design, most large-scale surveys do, and studies have shown that ignoring important design features during the imputation process can have drastic effects on the validity of the resulting estimates (Reiter, Raghunathan, and Kinney, 2006). Finally, the disclosure risk properties associated with fully synthetic data need to be studied in greater depth. Although we argue that fully synthetic data greatly enhances data confidentiality and prevents respondent re-identification because no observed data is released to the public, the extent to which confidentiality is protected needs to be systematically and empirically assessed.

Despite the potential for future improvements, the methodology examined here shows some promise and could be implemented by large-scale survey projects, such as the American Community Survey, to release more geographically-relevant data to the public. Such efforts could potentially help meet the growing demand for small area microdata, which is expected to grow among a variety of data users across many disciplines.

Acknowledgements

An earlier version of this paper was presented at the 2011 Joint Statistical Meetings in Miami, FL and was selected for a Student Paper Award from the Government Statistics Section/Social Statistics Section/Survey Research Methods Section of the American Statistical Association. This research was supported by dissertation grants from the National Science Foundation (SES-0918942), the U.S. Census Bureau (YA132309SE0354), and the Centers for Disease Control and Prevention (1-R36-SH-000016-01). The authors are very grateful to Maggie Levenstein, Clint Carter, Arne Reznick, and Lynn Riggs for their assistance and support with accessing the restricted data from the Michigan Census Research Data Center.

REFERENCES

ABOWD, J. M., STINSON, M., BENEDETTO, G., (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. http://www.census.gov/sipp/SSAfinal.pdf.

BINDER, D. A., (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review,* 51, 279–292.

DATTA, G. S., FAY, R. E., GHOSH, M., (1991). Hierarchical and Empirical Bayes Analysis in Small-Area Estimation. *Proceedings of the Annual Research Conference,* U. S. Bureau of the Census, 63–78.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B., (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1–38.

DRECHSLER, J., BENDER, S., RÄSSLER, S., (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. *Transactions on Data Privacy,* 105–130.

FAY, R. E., HERRIOT, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association,* 74, 269–277.

KARR, A. F., KOHNEN, C. N., OGANIAN, A., REITER, J. P., SANIL, A. P., (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *The American Statistician,* 60, 224–232.

KENNICKELL, A. B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques*. W. Alvey and B. Jamerson (eds.) Washington D. C.: National Academy Press, 248–267.

KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S., ABOWD, J. M., (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review,* 79, 362–384.

LINDLEY, D. V., SMITH, A. F. M., (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B,* 34, 1–41.

LITTLE, R. J. A., (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics,* 9, 407–426.

LITTLE, R. J. A., RUBIN, D. B., (2002). *Statistical Analysis with Missing Data*. 2nd Edition. Wiley.

LIU, F., LITTLE, R. J. A., (2002). Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. In *ASA Proceedings of the Joint Statistical Meetings,* 2, 2133–2138.

MACKIE, C., BRADBURN, N., (2000). *Improving Access to and Confidentiality of Research Data: Report of a Workshop.* Commission on Behavioral and Social Sciences and Education, National Research Council. National Academy Press, Washington, D. C.

MALEC, D., SEDRANKS, J., MORIARITY, C. L., LECLERE, F. B., (1997). Small Area Inference for Binary Variables in the National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815–826.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E., SINGH, M. P., (1987). *Small Area Statistics*. Wiley, New York.

RAGHUNATHAN, T. E., RUBIN, D. B., (2000). Bayesian Multiple Imputation to Preserve Confidentiality in Public-Use Data Sets. *ISBA 2000 The Sixth World Meeting of the International Society for Bayesian Analysis*.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J., SOLENBERGER, P., (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology,* 27, 85–95.

RAGHUNATHAN, T. E, REITER, J. P., RUBIN, D. B., (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics,* 19, 1–16.

RAO, J. N. K., (1999). Some Recent Advances in Model-based Small Area Estimation. *Survey Methodology,* 25, 175–186.

RAO, J. N. K., (2003). *Small Area Estimation*. Wiley, New York.

REITER, J. P., (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. *Journal of Official Statistics,* 18, 531–544.

REITER, J. P., (2003). Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology,* 29, 181–188.

REITER, J. P., (2004).Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology,* 30, 235–242.

REITER, J. P., (2005). Releasing Multiply-Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society, Series A,* 168, 185–205.

REITER, J. P., RAGHUNATHAN, T. E., KINNEY, S. K., (2006). The Importance of Modeling the Survey Design in Multiple Imputation for Missing Data. *Survey Methodology,* 32, 143–150.

REITER, J. P., RAGHUNATHAN, T. E., (2007).The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association,* 102, 1462–1471.

RODRIGUEZ, R., (2007). Synthetic Data Disclosure Control for American Community Survey Group Quarters. In *ASA Proceedings of the Joint Statistical Meetings,* 1439–1450.

RUBIN, D. B., (1983). A Case-Study of the Robustness of Bayesian/Likelihood Methods of Inference: Estimating the Total in a Finite Population using Transformations to Normality. In *Scientific Inference, Data Analysis and Robustness*. G.E.P. Box, T. Leonard, and C.F. Wu (eds.) New York: Academic Press, 213–244.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley: New York.

RUBIN, D. B., (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Microdata. *Journal of Official Statistics,* 9, 461–468.

TRANMER, M., PICKLES, A., FIELDHOUSE, E., ELLIOT, M., DALE, A., BROWN, M., MARTIN, D., STEEL, D., GARDINER, C., (2005). The Case for Small Area Microdata. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168, 29–49.

U.S. CENSUS BUREAU, (2009). American Community Survey: Design and Methodology. http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf

YUCEL, R. M., (2008). Multiple Imputation Inference for Multivariate Multilevel Continuous Data with Ignorable Non-Response. *Philosophical Transactions of the Royal Society A,* 366, 2389–2403.

APPENDICES

Appendix 1. EM algorithm for estimating Bayesian hyperparameters

The EM algorithm is used to estimate the unknown population parameters $\beta_p$and $\Sigma_p$from the following setup,

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

$$\beta_{cs,p} \sim MVN\left(\beta_p Z_s, \Sigma_p\right)$$

where$p = (1,2,\dots,P)$ is used to index the set of parameters associated with the p^{th} synthetic variable of interest and the p^{th} regression model from which the direct estimates $\hat{\beta}_{cs}$ and $\hat{V}_{cs}$ were obtained in Step 1.

The E step consists of solving the following expectations,

$$\beta^*_{cs,p} = E\left(\beta_{cs,p}\right) = \left[\left(\hat{\mathrm{V}}^{-1}_{cs,p} + \Sigma_p^{-1}\right)^{-1}\left(\hat{\mathrm{V}}^{-1}_{cs,p}\hat{\beta}_{cs} + \Sigma_p^{-1}\beta_p Z_s\right)\right]$$

$$\left[\beta_{cs,p}\left(\beta_{cs,p}\right)^T\right]^* = E\left[\beta_{cs,p}\beta^T_{cs,p}\right] = \left(\hat{\mathrm{V}}^{-1}_{cs,p} + \Sigma_p^{-1}\right)^{-1} + \beta^*_{cs,p}\left(\beta^*_{cs,p}\right)^T$$

Once these expectations are computed they are then incorporated into the maximization (M-step) of the unknown hyperparameters β_p and$\hat{\Sigma}_p$ using the following equations,

$\hat{\beta}_p = \beta^*_{+s,p} Z_s \left(Z_s Z_s^T\right)^{-1}$, where $\beta^*_{+s} = \left(\sum_{c=1}^{C_s} \beta^*_{cs}\right)/C_s$, and

$$\hat{\Sigma}_p = \frac{\left[\sum_{s=1}^{S} \left[\sum_{c=1}^{C_s}\left(\beta^*_{cs,p} - \hat{\beta}_p Z_s\right)\left(\beta^*_{cs,p} - \hat{\beta}_p Z_s\right)^T\right] \Big/ C_s\right]}{S}$$

After convergence the maximum likelihood estimates are incorporated into the posterior distribution of $\beta_{cs,p}$ shown in equation [5].

Appendix 2. Creation of synthetic household size variable

Let Z_{hcs} be the number of people in household $h = (1,2,\dots,n_{cs})$ in county $c = (1,2,\dots,C_s)$ within state $s = (1,2,\dots,S)$. Assume that $Z_{hcs} \sim Poisson(\lambda_{cs})$and $\lambda_{cs} \sim Gamma(\alpha_s,\beta_s)$. Conditional on the data and $(\alpha_s,\beta_s; s = 1,2,\dots,S)$ it is straightforward to simulate values of Z_{hcs}.

First, obtain the marginal maximum likelihood estimates of $(\alpha_s,\beta_s; s = 1,2,\dots,S)$ through Newton-Raphson for each state independently. Also, obtain the covariance matrix $\hat{V}_s = Cov(\hat{\alpha}_s,\hat{\beta}_s)$ by inverting the observed Fisher Information matrix. The marginal likelihood is given by,

$$\int \left\{ \prod_{c=1}^{C_s} e^{-\beta_s\lambda_{cs}} \lambda_{cs}^{\alpha_s-1} \left(\prod_{h=1}^{n_{cs}} e^{-\lambda_{cs}}\, \lambda_{cs}^{Z_{hcs}} \right) / \Gamma(\alpha_s) d\lambda_{cs} \right\}$$

$$= \prod_{c=1}^{C_s} \int e^{-(\beta_s+n_{cs})\lambda_{cs}}\, \lambda_{cs}^{Z_{+cs}+\alpha_s-1} / \Gamma(\alpha_s)\beta_s^{\alpha_s}\, d\lambda_{cs}$$

$$= \prod_{c=1}^{C_s} \{\Gamma(Z_{+cs}+\alpha_s)\}(\beta_s+n_{cs})^{-(Z_{+cs}+\alpha_s)} / \Gamma(\alpha_s)\, \beta_s^{\alpha_s}$$

where $Z_{+cs} = \sum_{h=1}^{n_{cs}} Z_{hcs}$. Taking the logarithms, the quantity to be maximized with respect to α_s and b_s via the Newton-Raphson is,

$$L = \sum_{c=1}^{C_s} \{log\Gamma(Z_{+cs}+\alpha_s) - (Z_{+cs}+\alpha_s)log(\beta_s+n_{cs})\} - C_s log\Gamma(\alpha_s) + C_s\alpha_s log(\beta_s)$$

The first and second derivatives of this function are,

$$\frac{\partial L}{\partial \alpha_s} = \sum_{c=1}^{C_s} \{\psi(Z_{+cs}+\alpha_s) - log(\beta_s+n_s)\} - C_s\psi(\alpha_s) + C_s log(\beta_s)$$

$$\frac{\partial L}{\partial \beta_s} = -\sum_{c=1}^{C_s} \{(Z_{+cs}+\alpha_s)/(\beta_s+n_s)\} + C_s\alpha_s/\beta_s$$

$$\frac{\partial^2 L}{\partial \alpha_s^2} = \sum_{c=1}^{C_s} \psi'(Z_{+cs}+\alpha_s) - C_s\psi'(\alpha_s)$$

$$\frac{\partial^2 L}{\partial \beta_s^2} = \sum_{c=1}^{C_s} \{(Z_{+cs}+\alpha_s)/(\beta_s+n_s)^2\} - \alpha_s C_s/\beta_s^2$$

$$\frac{\partial^2 L}{\partial \beta_s \partial \alpha_s} = -\sum_{c=1}^{C_s} 1/(\beta_s+n_s) + C_s/\beta_s$$

The logarithm of the gamma function, its first and second derivatives can be accurately approximated as follows,

$$log\Gamma(z) = -log\sum_{i=1}^{26} c_i z^i$$

$$\psi(z) = \frac{\partial}{\partial z} log\Gamma(z) = -\frac{\sum_{i=1}^{26} ic_i z^{i-1}}{\sum_{i=1}^{26} c_i z^i}$$

$$\psi'(z) = \left(\frac{\sum_{i=1}^{26} ic_i z^{i-1}}{\sum_{i=1}^{26} c_i z^i}\right)^2 - \frac{\sum_{i=1}^{26} i(i-1)c_i z^{i-2}}{\sum_{i=1}^{26} c_i z^i}$$

The constants c_i can be found in Abramowitz and Stegun (1965). The Newton-Raphson method is applied iteratively to obtain maximum likelihood estimates of $\alpha_s$and β_s,

$$\begin{pmatrix} \alpha_{s,n+1} \\ \beta_{s,n+1} \end{pmatrix} = \begin{bmatrix} \partial^2 L / \partial\alpha_{s,n}^2 & \partial^2 L / \partial\alpha_{s,n}\partial\beta_{s,n} \\ \partial^2 L / \partial\beta_{s,n}\partial\alpha_{s,n} & \partial^2 L / \partial\beta_{s,n}^2 \end{bmatrix}^{-1} \begin{pmatrix} \partial L / \partial\alpha_{s,n} \\ \partial L / \partial\beta_{s,n} \end{pmatrix}$$

The logarithm of the estimates for α_s and β_s are then assumed to follow the hierarchical model,

$$\begin{pmatrix} log\,\hat{\alpha}_s \\ log\,\hat{\beta}_s \end{pmatrix} \sim N\left[\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix}, \begin{bmatrix} 1/\hat{\alpha}_s & 0 \\ 0 & 1/\hat{\beta}_s \end{bmatrix} \hat{V}_s \begin{bmatrix} 1/\hat{\alpha}_s & 0 \\ 0 & 1/\hat{\beta}_s \end{bmatrix}\right] = N\left[\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix}, \hat{\Sigma}_s\right]$$

$$\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix} \sim N\left[\begin{pmatrix} \theta \\ \phi \end{pmatrix}, \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{22} & \Omega_{22} \end{bmatrix}\right] = N\left[\begin{pmatrix} \theta \\ \phi \end{pmatrix}, \Omega\right]$$

The Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is used to obtain maximum likelihood estimates of (θ, ϕ, Ω). The E step is carried out by solving the following expectation equations,

$$\begin{pmatrix} log\,\alpha_s^* \\ log\,\beta_s^* \end{pmatrix} = E\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix} = \left[\left(\hat{\Sigma}_s^{-1} + \Omega^{-1}\right)^{-1}\left(\hat{\Sigma}_s^{-1}\begin{pmatrix} log\,\hat{\alpha}_s \\ log\,\hat{\beta}_s \end{pmatrix} + \Omega^{-1}\begin{pmatrix} \theta \\ \phi \end{pmatrix}\right)\right]$$

$$\left[\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix}\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix}^T\right]^* = E\left[\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix}\begin{pmatrix} log\,\alpha_s \\ log\,\beta_s \end{pmatrix}^T\right]$$

$$= \left(\hat{\Sigma}_s^{-1} + \Omega^{-1}\right)^{-1} + \begin{pmatrix} log\,\alpha_s^* \\ log\,\beta_s^* \end{pmatrix}\begin{pmatrix} log\,\alpha_s^* \\ log\,\beta_s^* \end{pmatrix}^T$$

and the *M* step is performed by solving the following maximization equations,

$$\begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix} = \frac{\left[\sum_{s=1}^{S}\begin{pmatrix}log\ \alpha_s^*\\ log\ \beta_s^*\end{pmatrix}\right]}{S}$$

$$\hat{\Omega} = \begin{bmatrix}\hat{\Omega}_{11} & \hat{\Omega}_{12}\\ \hat{\Omega}_{22} & \hat{\Omega}_{22}\end{bmatrix} = \frac{\left[\sum_{s=1}^{S}\left(\begin{pmatrix}log\ \alpha_s^*\\ log\ \beta_s^*\end{pmatrix} - \begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix}\right)\left(\begin{pmatrix}log\ \alpha_s^*\\ log\ \beta_s^*\end{pmatrix} - \begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix}\right)^T\right]}{S}$$

It is then straightforward using this setup to synthesize the number of members in each household by treating the parameter estimates of (θ, ϕ, Ω) as known and retracing back to simulate values of Z_{hcs} using the following 3 steps:

Step 1: Simulate Gamma parameters α_s and β_s from the bivariate normal distribution, $\begin{pmatrix}\tilde{\alpha}_s\\ \tilde{\beta}_s\end{pmatrix} \sim exp\left[N\left[\left(\hat{\Sigma}_s^{-1} + \hat{\Omega}^{-1}\right)^{-1}\left(\hat{\Sigma}_s^{-1}\begin{pmatrix}log\ \hat{\alpha}_s\\ log\ \hat{\beta}_s\end{pmatrix} + \Omega^{-1}\begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix}\right), \left(\hat{\Sigma}_s^{-1} + \hat{\Omega}^{-1}\right)^{-1}\right]\right]$,

Step 2: Simulate Poisson parameter λ_{cs} from the Gamma distribution given the county population size, number of households, and simulated parameters obtained from Step 1,
$\tilde{\lambda}_{cs} \sim Gamma\left(Z_{+cs} + \tilde{\alpha}_s, \tilde{\beta}_s + \ n_{cs}\right)$,

Step 3: Simulate household size $Z_{hcs}$from the Poisson distribution,
$\tilde{Z}_{hcs} \sim Poisson\left(\tilde{\lambda}_{cs}\right)$.