Synthetic Data for Small Area Estimation

by

Joseph Walter Sakshaug

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2011

Doctoral Committee:

Professor Trivellore E. Raghunathan, Chair
Professor James M. Lepkowski
Professor Roderick J. Little
Research Professor Richard L. Valliant

Acknowledgements

I would like to thank my committee for their time and effort. In addition, I would like to thank the following persons for their assistance and support in using the Michigan Census Research Data Center: Maggie Levenstein, Lynn Riggs, Clint Carter, and Arnie Reznick. Finally, I would to thank the U.S. Census Bureau, the Centers for Disease Control and Prevention, the National Science Foundation, and the Rackham School of Graduate Studies for funding my dissertation research.

Table of Contents

List of Figures

List of Tables

Abstract

Small area estimates provide a critical source of information used by a variety of stakeholders to study human conditions and behavior at the local level. Statistical agencies regularly collect survey microdata from small geographic areas but are prevented from identifying these areas in public-use microdata sets due to disclosure concerns. Alternative data dissemination methods include releasing summary tables for small areas and accessing restricted identifiers via Research Data Centers. This dissertation proposes a new method of disseminating public-use microdata that contains more geographical details than are currently being released. The basic idea is to replace the observed survey values with imputed, or synthetic, values. Data confidentiality is enhanced because no actual values are released.

This dissertation proposes three statistical methods for generating synthetic data for small geographic areas. The first method utilizes a fully-parametric hierarchical Bayesian model that is used to generate synthetic microdata from the posterior predictive distribution. The second method consists of a nonparametric procedure for generating synthetic data for continuous non-normal distributions. The third method accounts for complex sample design features and permits the generation of synthetic data for both sampled and nonsampled small areas.

These three methods are demonstrated and evaluated using a mix of public-use and restricted microdata from the American Community Survey and National Health

Interview Survey. Each of the methods is evaluated using empirical, simulation, and cross-validation studies. The analytic validity of the methods is assessed by comparing the small area estimates obtained from the synthetic data with those obtained from the observed data.

Chapter 1

Introduction

1 Introduction

Increasingly, researchers are demanding greater access to survey microdata for small geographic areas to compute estimates that may influence policy and intervention strategies at local levels. Statistical agencies regularly collect survey and census data from small geographic areas, but are prevented from releasing detailed geographical identifiers in public-use data sets due to privacy concerns and disclosure risks.

The conflicting tradeoff between data utility and data protection has motivated statistical agencies to consider data dissemination procedures that allow researchers to access restricted geographical identifiers while keeping disclosure risks at tolerable levels. Existing data dissemination practices include: 1) releasing summary tables containing aggregate-level data for small geographic areas; 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).

Each of the current data dissemination practices has limitations that may discourage users from using the survey data. For example, summary tables are limited to existing data products and cells may still be suppressed due to insufficient cell sizes. Public-use microdata provide users with additional flexibility to perform customized analyses on-demand, but the lowest level of geography (such as Public-Use Microdata

Areas (PUMAs)) may not be sufficient for small area estimation. Accessing the raw microdata is possible in an RDC, but potential users must apply for a clearance request, travel to the nearest RDC, and pay usage fees, which may not be possible for some data users.

This dissertation proposes a fourth data dissemination approach that statistical agencies may adopt to release more detailed geographical information in public-use microdata sets. The approach builds on the method, originally proposed by Rubin (1993), of creating multiply-imputed, or synthetic, data sets that are released to the public in lieu of the observed survey data sets. The basic idea is to treat the non-sampled portion of the population as missing data to be replaced with multiply-imputed data. Samples are then drawn from the synthetic data populations and released as public-use data sets. Valid inferences are obtained by applying standard combining rules to the synthetic data (Raghunathan, Reiter, and Rubin, 2003). Data confidentiality is greatly enhanced because no observed data values are released to the public.

The synthetic data literature focuses on preserving statistics about the entire sample, but ignores the preservation of small area statistics. Statistics about small areas can be extremely valuable to data users, but detailed geographical identifiers are almost always suppressed from public-use microdata sets. Releasing synthetic data for small geographic areas may be ideally suited for releasing restricted geographical information while overcoming the limitations of other disclosure avoidance methods.

Several methodological challenges must be overcome in order to determine whether synthetic data can be a viable alternative to existing data dissemination methods. Small area inferences obtained from the synthetic data should resemble the corresponding inferences obtained from the actual data. Distributions of variables observed in small geographic areas should also be preserved, including those that do not follow standard parametric forms (e.g., Gaussian) as is often the case for many key survey variables. Finally, many survey data sets are collected under a complex sample design. The synthetic data should account for complex design features, such as stratification, clustering, and weights to ensure that valid inferences for small areas can be preserved (Reiter, Raghunathan, and Kinney, 2006).

In this dissertation, I develop methods of generating synthetic data for small area estimation that address these issues. Each chapter answers one of the following research questions:

Chapter 2: How to generate synthetic data sets that preserve inferences obtained from small geographic areas?

Chapter 3: How to generate synthetic data sets based on nonparametric methods that preserve non-standard distributional forms for continuous variables?

Chapter 4: How to generate synthetic data sets for small geographic areas that incorporates complex sample design features into the synthetic data generation process?

1.1 Background and Significance

Many statistical agencies and survey organizations disseminate data on individual units in public-use data files (i.e., microdata). Data disseminators strive to release files that are informative for a wide range of statistical analyses, yet safe from disclosures instigated by data users seeking to learn respondents' identities or sensitive attributes.

Disclosure risks have received much attention due to the proliferation of readily available commercial and non-commercial databases. Coupled with advances in statistical, computing, and data linkage techniques, the potential exists for an intruder to re-identify a de-identified survey record. Statistical agencies that fail to prevent disclosures of respondents' identities may be subject to serious legal consequences. An act of disclosure may discourage the public from participating (or providing accurate answers) in future surveys if they believe their privacy is threatened.

Data disseminators use many techniques to minimize disclosure risks. They include recoding exact values if they exceed a specific threshold (e.g., recoding 80,000 to "50,000 or more"), recoding variables into coarse categories (e.g., releasing only 5-year intervals for age), swapping the values of variables for records that are statistically similar (Dalenius and Reiss, 1982; Reiss, 1984), and adding random noise to data values. Although these methods enhance confidentiality protection to some degree, they can also distort relationships between variables in the data set and can introduce bias. They can complicate analyses for data users because specialized analytic methods may be needed to adjust for the distorted data (e.g., using measurement error models to analyze data with added noise).

An alternative disclosure limitation method is to synthesize the observed data using a probabilistic imputation model. The basic idea of releasing multiply-imputed, or synthetic, data sets in lieu of the observed data sets was initially proposed by Rubin (1993) and further developed by Raghunathan et al. (2003) and Reiter (2005). Synthetic data has been shown to have advantages over alternative statistical perturbation methods (Winkler, 2004; Reiter, 2005). Fully-synthetic data has two key advantages: it offers enhanced confidentiality protection because no observed information is released and the approach allows data users to produce valid inferences for various estimands by using complete-data statistical methods and software.

The basic idea behind synthetic data generation is to treat the unobserved portion of the population as missing data to be replaced with multiply-imputed data. The observed data is used to construct a posterior predictive distribution from which the multiply-imputed values are drawn. Multiple synthetic populations are generated and a sample is drawn from each synthetic population which comprises the public-use data files. Synthetic sample sizes can be drawn such that they exceed the observed sample size to facilitate the application of direct estimation methods during analysis.

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 $\left(q^{(l)}, v^{(l)}; l = 1,2, \dots, M\right)$ be the values of q and v computed on the M synthetic data sets. It is assumed that these quantities are estimated based on a simple random sampling design. Under assumptions described in Raghunathan et al. (2003), 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 the synthetic sample size n_{syn} large. A more precise variance estimator that is always positive is outlined in Raghunathan et al. (2003). Inferences for scalar Q are based on a normal distribution when $T_M > 0$, n, M, n_{syn} are large. For moderate M, inferences can be based on t-distributions (Reiter, 2002).

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 (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. A less extreme approach involves synthesizing a partial set of variables or records that are most vulnerable to disclosure (Little, 1993; Kennickell, 1997; Liu and Little, 2002; Reiter, 2003). If implemented properly, this approach yields high analytic validity because inferences are less sensitive to the specification of the imputation model. However, partial synthesis may not provide the same level of protection as full synthesis because the observed sample units, and the majority of their data values, are released to the public (Drechsler, Bender, and Raessler, 2008).

The existing synthetic data literature focuses on preserving statistics about the entire sample, but doesn't address the problem of preserving statistics within small geographic areas. Building synthetic data generation models that incorporate the hierarchical structure associated with each geographical area (e.g., state, county) offers a promising solution. The main goal of this strategy is to enable data users to produce valid statistics for various levels of geography using a single set of synthetic data files.

In order to achieve approval from statistical agencies, as well as data users, the hierarchical synthetic data approach must be flexible enough to overcome several practical challenges, such as preserving variable distributions that do not follow strict parametric forms. Many key survey variables are not easily simulated using parametric distributions. This is an open area of research in the synthetic data literature, and more broadly, in the multiple imputation literature (He and Raghunathan, 2006).

Another practical issue for generating hierarchical synthetic data is that the approach should be flexible enough to handle different types of surveys, such as those that were collected using an *EPSEM* design, or more sophisticated sampling designs that may include clustering, stratification, and unequal probabilities of selection. Accounting for complex design features in multiple imputations for missing data has been addressed in the literature (Reiter, Raghunathan, and Kinney, 2006), but not in the context of synthetic data for small area estimation.

1.2 Modeling Approach and Evaluation

The approach here adopts Bayesian methods, using a hierarchical imputation model, to generate synthetic data for small area estimation. There involves three stages. In the first, sequential regression models are fit using the observed data within small areas (e.g., counties) to approximate the joint density of the set of variables to be synthesized. In the second, the joint sampling distribution of the population regression parameters is approximated and the between-area variation is modeled by incorporating area-level covariates. In the third, the population parameters are simulated and synthetic data is generated by taking independent draws from the posterior predictive distribution within each small area.

The modeling approach is modified in later chapters to incorporate a nonparametric adjustment procedure to handle continuous variables that do not follow a standard distributional form (Chapter 3), and to account for complex sample design features in the synthetic data generation process and also generate valid synthetic data for non-sampled small areas (Chapter 4).

The proposed synthetic data procedures are demonstrated on and evaluated against actual survey data obtained from the American Community Survey and National Health Interview Survey. Both surveys suppress small area identifiers (e.g., county- and sub-county identifiers) in public-use data files. In this demonstration project, we use a mix of public-use microdata and restricted-use microdata to evaluate the methods using different levels of geography provided in each survey data set (i.e., Public-Use Microdata Areas for ACS public-use data; and counties for restricted-use ACS and NHIS data). The validity of univariate and multivariate small area inferences are assessed by comparing the inferences obtained from the synthetic data sets with those obtained from the actual data sets.

1.3 Benefits and Potential Impacts

This research has several potential impacts. The most important is the potential for greater access to detailed geographical information in public-use data sets while preserving data confidentiality. Existing procedures for accessing survey data for small geographic areas is somewhat cumbersome and may not be convenient for all data consumers. Synthetic data offers a more flexible alternative and since no actual data is released to the public, confidentiality protection is enhanced.

This research also addresses an important gap in the literature, which is how to preserve small area statistics in synthetic microdata sets. Small area statistics are quite valuable to researchers, policy-makers, and students, but they are difficult to obtain due to privacy concerns.

The methods proposed in this dissertation are flexible, do not require Markov Chain Monte Carlo (MCMC) algorithms (Geman and Geman, 1984; Gelfand and Smith, 1990), and can be applied to various types of survey data sets. These procedures may be adopted by statistical agencies and lead to new data products.

This research may also stimulate a shift in how public-use data files are released to the public. Some statistical agencies are already moving towards customized data extract systems. The synthetic data framework is flexible enough to handle detailed users requests and could be operationalized in an automated fashion, so that users can choose exactly which variables and complex design features to incorporate into the synthetic data generation process.

Chapter 2

Synthetic Data for Small Area Estimation in the American Community Survey

1 Introduction

Demand for small area estimates is growing heavily among a variety of stakeholders who use these data to advance the study of issues affecting 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 doing so, because releasing detailed geographical identifiers for small areas can increase the risk of respondent re-identification and inadvertent disclosure of confidential information.

In order to minimize the risk of disclosure, statistical agencies commonly adopt one of the following methods of data dissemination: 1) release summary tables that contain aggregated data for specific geographic areas (e.g., counties, census tracts, block groups); 2) suppress geographical details in public-use microdata sets for all 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 enclave or Research Data Center (RDC). Although useful in some situations each approach has limitations that preclude its ability to meet the growing demand for small area data that is being fueled 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 within a wide variety of geographical areas, but their utility is limited for addressing complex scientific hypotheses that require additional variables, interactions, or modeling approaches not obtainable from existing aggregated 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 microdata via a Research Data Center overcomes the limitations of the previous two by permitting users access to the full unmasked microdata, including detailed geographical 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 sometimes long distances to the nearest RDC facility. Unfortunately, these requirements are too restrictive for many analysts.

1.1 Synthetic Data for Small Geographic Areas

This chapter investigates a fourth approach that statistical agencies may adopt 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 not addressed. Statistics about small areas can be extremely valuable to data users, but detailed geospatial information is almost always suppressed in public-use survey data. 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 (Platek et al., 1987; Rao, 2003).

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, 2005). If implemented properly, this approach yields high analytic validity as inferences are less sensitive to misspecification of the imputation model, but 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, Bender, and Raessler, 2008).

At the present time, statistical agencies have only released partially synthetic data files (Rodriguez, 2007; Abowd, Stinson, and Benedetto, 2006; Kinney and Reiter, 2008). 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 offers 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 Chapter

This chapter proposes 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 developed 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 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 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 assessed by comparing inferences obtained from the synthetic data with those obtained from the actual data. The empirical evaluation of the disclosure risk properties of the proposed synthetic data approach are left to future work. Limitations of the approach 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, Reiter, and Rubin (2003) and Reiter (2004). 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 survey data set, but that 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, 10, or 20) 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 $P_{syn} = \left(P^{(l)}; l =\right.$

$1,2,\dots,M\big)$ 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, Reiter, and Rubin (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, Reiter, and Rubin (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 the units that comprise the synthetic samples from each multiply-imputed synthetic population. Because of this additional 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) and Reiter (2005).

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 M or n_{syn}. Numerical routines can be used to calculate the integrals involved in the construction of T_M, yielding more precise variance estimates (Raghunathan, Reiter, and Rubin, 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 (Yucel, 2008; Reiter, Raghunathan, and Kinney, 2006).

The approach involves three stages. In the first, 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, the sampling distribution of the unknown regression parameters estimated in stage 1 is approximated and the between-area variation is modeled using auxiliary information. In the third, the unknown regression parameters are simulated and used to draw synthetic microdata values from the posterior predictive distribution.

Only two levels of geography are considered. Consider “small areas” as counties nested within states in the U.S. In illustrating the modeling steps, the models are kept relatively simple from a computational perspective to make the modeling practical. Despite the simplified presentation, the framework can handle more sophisticated modeling 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 $c = (1,2, \dots, C_s)$ nested within state $s = (1,2, \dots, S)$. Let $Y_{cs} = \left(Y_{ics,p}; i = 1,2, \dots, n_{cs};\ p = 1,2, \dots, P\right)$ represent the $n_{cs} \times P$ matrix of survey variables collected from each survey respondent located in county c and state s. Let $X_{cs} = \left(X_{ics,j}; i = 1,2, \dots, n_{cs}, n_{cs} + 1, \dots, N_{cs};\ j = 1,2, .., J\right)$ 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 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 modeling 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 β_p and Σ_p, but for computational simplicity I assume that β_p and Σ_p are fixed at their respective maximum likelihood estimates (MLE), 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,

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

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 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 $f\left(\tilde{\sigma}_{cs,1}^{2}|Y_{cs,1},X_{cs},\sigma_{cs,1}^{2}\ \right)$, or fixed at the 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 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 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 small areas. Inferences for a variety of estimands can be obtained using the combining rules in Section 2.2.

4 Application: American Community Survey (Public-Use Microdata)

In this section, consider a subset of public-use microdata from the 2005-2007 U.S. American Community Survey (ACS). The ACS is an ongoing national survey that provides yearly estimates on a variety of topics, including income and benefits, health insurance, disabilities, family and relationships, among others. The ACS collects information on persons living in housing units and group quarters facilities in all 3,142 counties. 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 for all addresses that do not respond by mail. Following the telephone operation, a sample is taken from addresses which have not been interviewed and these addresses are visited by a field representative. Full details of the ACS methodology can be found elsewhere (Census Bureau, 2009).

The smallest geographic unit that is identified in the public-use ACS microdata is a Public-Use Microdata Area (PUMA). PUMAs are census areas that contain at least 100,000 persons, are nested within states or equivalent entities, cover the entirety of the United States, Puerto Rico, Guam, and the U.S. Virgin Island, are built on counties and census tracts, and are contiguous. For this application, I restrict the ACS sample to the Northeast region of 9 states and 405 PUMAs. ACS data was collected in each of these PUMAs during the 3-year study period. I also restrict the data to seven household- and seven person-level variables measured on the 599,450 households and 1,506,011 persons in the ACS Northeast region sample. The variables, shown in Table 2.0, were chosen by statisticians at the U.S. Census Bureau specifically for this project.

$M = 10$ fully synthetic data sets are generated for each "small area" or PUMA. To ensure that each synthetic data set contains ample numbers of households and/or persons within PUMAs, synthetic samples are larger than the observed sample sizes, approximately equivalent to 20% of the total number of households located in each PUMA based on the 2000 decennial census counts. This yielded a total synthetic sample size of 3,963,715 households and 10,192,987 persons in the Northeast region.

Table 2.0. 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 - 516	log
Total bedrooms	count	0 - 5	--
Electricity bill/mo.	continuous	1 - 600	cube root
Total rooms (excl. bedrooms)	count	1 - 7	--
Income	continuous	0 – 2,158,100	cube root
Tenure	polytomous	recoded; mortgage/loan, own free and clear, rent	--
Person variables			
Sampling weight	continuous	1 - 814	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 - 95	--
Race	polytomous	recoded; white, black, other	--
Living in poverty	binary	yes, no	--

The first survey variable to be synthesized was household size. Creating a household size variable facilitates the generation of synthetic person-level variables in a later step. 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 Newton-Raphson algorithm; see Appendix 2 for details). All subsequent variables were synthesized using the hierarchical modeling approach described in Section 3. State-level covariates Z_s incorporated into the hierarchical model included: population size (2005 estimate: log-transformed), number of metropolitan and micropolitan areas obtained from the Census Bureau website for year 2005.

For numerical variables (continuous, count), design-based estimates of regression parameters were obtained by fitting normal linear models within each PUMA 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. The same approach was applied to polytomous variables after breaking them up into a series of binary variables. To ensure the stability of the design-based regression estimates, a minimum PUMA sample size rule of $15 \cdot p$ was applied within each PUMA. If a PUMA did not meet this sample size threshold, then nearby PUMAs were pooled together until the criterion was met.

After the household variables were synthesized, the synthetic household data sets were converted to person-level data sets and the person-level variables were synthesized unconditional to the household-level variables. 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. Finally, 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. All estimates are based on unweighted data.

4.1 Validity of Univariate Estimates

Figures 2.1 and 2.2 contain back-to-back histograms depicting the overall distributions of each household- and person-level variable, respectively. The actual distribution is shown in red and the synthetic distribution in blue. All variables are presented on the untransformed scale. The results are mixed. For some variables, the synthetic data distribution resembles the actual data distribution reasonably well, but for others, the correspondence is poor. The continuous variables, shown in the top row of each figure, exhibit the most discordance. Although the bulk of the actual 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. (This is expected because the parametric model is dominating the result. A more careful modeling approach, such as a mixture model, would generate synthetic data that is better matched distributionally.) For example, the age distribution shown in Figure 2.2 has an approximately bimodal shape which is poorly reflected in the synthetic data. A mixture model or nonparametric imputation procedure might do a better job of preserving non-standard distributional forms than the parametric procedure we consider here.

Figure 2.1. Back-to-Back Histograms of Actual (Red) and Synthetic (Blue) Distributions for ACS Household-Level Variables in the Northeast Region.

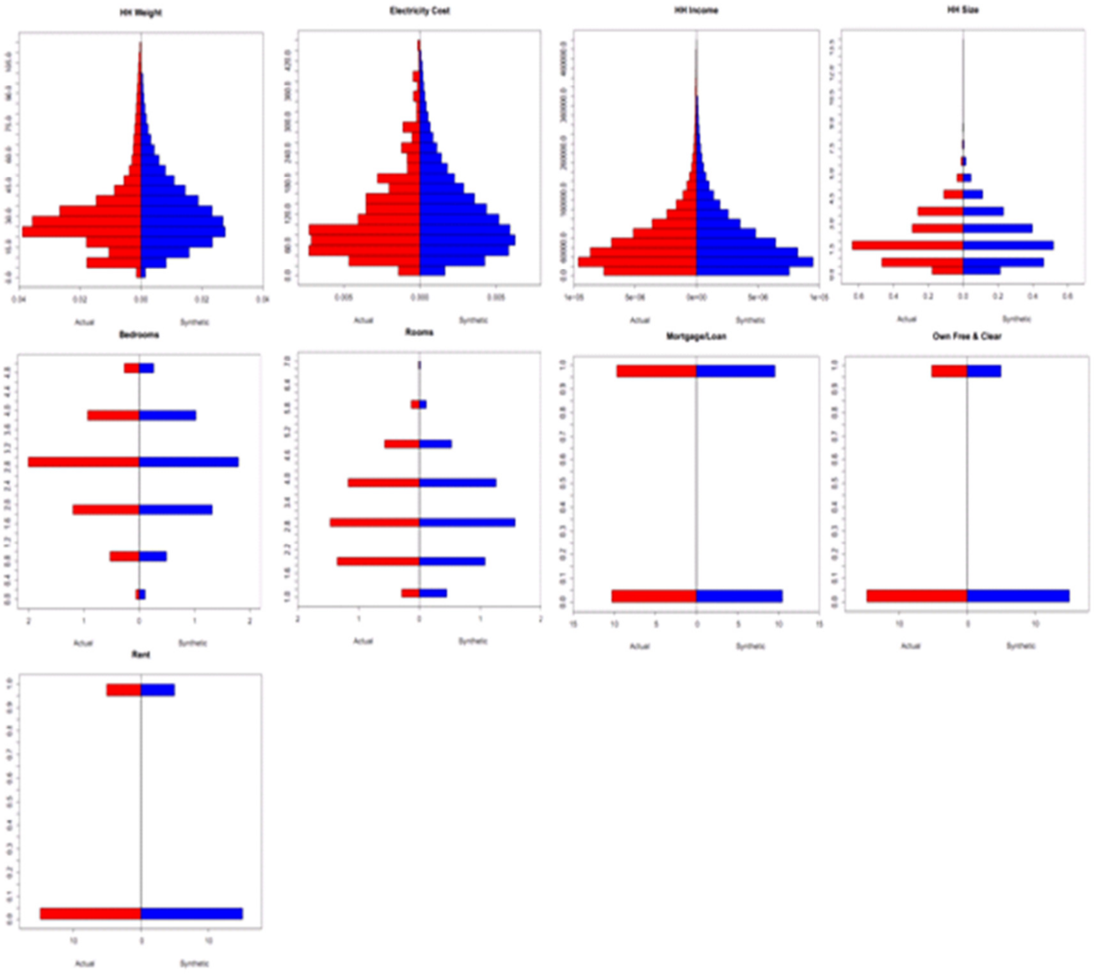
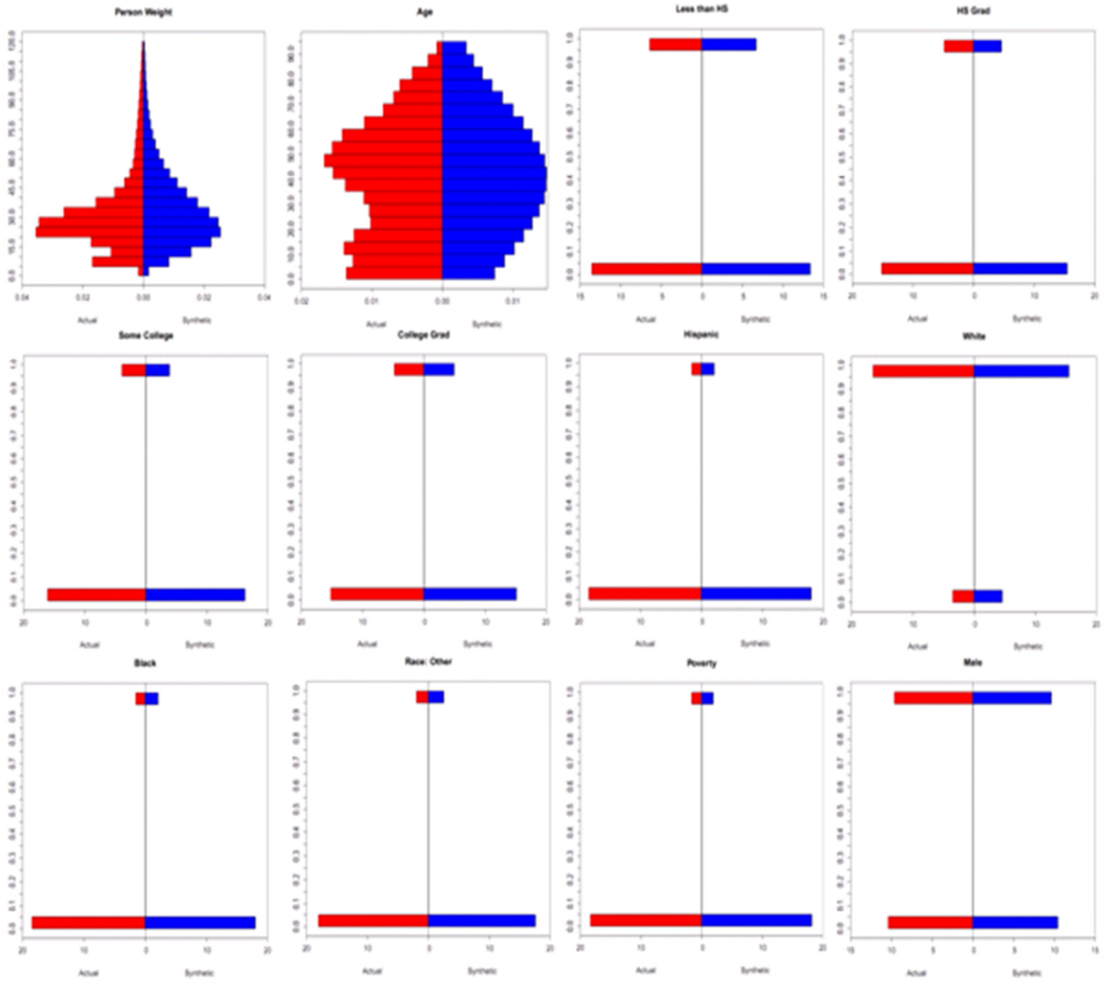


Figure 2.2 Back-to-Back Histograms of Actual (Red) and Synthetic (Blue) Distributions for ACS Person-Level Variables in the Northeast Region.



Although it is useful to compare synthetic and actual variable distributions, data users are ultimately interested in the validity of the estimates obtained from the synthetic data. Tables 2.1, 2.2, and 2.3 provide summary measures at the PUMA-, state- and region-levels, respectively, for univariate estimands obtained from the synthetic and actual data. The list of variables in column 1 includes the original set of ACS variables as well as recoded variables (income percentiles) and subgroups (income x tenure; poverty x race/ethnicity). The second column of Table 2.1 shows the average PUMA mean obtained from the synthetic and actual data. The third and fourth columns show the average PUMA standard deviation and standard error of the mean. The last column contains the intercept and slope values obtained from regressing the actual PUMA means against the corresponding synthetic means. Intercept values close to 0 and slope values close to 1 indicate strong correspondence between the synthetic and actual means.

For the 15 household-level estimands, all but 2 of them yield an average synthetic PUMA mean lying within one average standard error from the average actual PUMA mean. Although these results should not be treated as a full endorsement of the synthetic data, they do provide some reassurance that the synthetic data yield valid estimates for most PUMAs. The two outlying averages correspond to the recoded income variables representing the 75th and 90th percentiles, which tend to be overestimated in the synthetic data, on average. For the person-level variables, only 1 estimand out of a total of 16 yielded an average synthetic PUMA mean which differed from the average actual PUMA mean by more than one average standard error. The average standard errors of the PUMA means tend to be similar with a slight overestimation of the synthetic standard errors. It should be noted that the synthetic standard deviations tend to be smaller than the actual standard deviations, on average, for the transformed continuous variables (sampling weight, electricity costs, income). The underestimation could be due to the failure of the imputation model and transformation in preserving the tail-end of the distribution in the synthetic data, a problem which has been highlighted in earlier research on the estimation of totals in skewed populations (Rubin, 1983).

Differences between the synthetic and actual estimands are more apparent for state- and region-level inferences. Many of the synthetic means differ from the corresponding actual means by more than one standard error, on average. Rare

Table 2.1 Summary Measures of Actual and Synthetic PUMA 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	2.34	2.34	1.48	1.53	0.04	0.04	0.00	1.00
Sampling weight	33.71	33.45	20.03	17.80	0.55	0.49	0.19	1.00
Total bedrooms	2.77	2.80	1.01	1.01	0.03	0.04	-0.07	1.01
Electricity bill/mo.	125.08	126.80	85.39	82.80	2.32	2.49	1.55	0.97
Total rooms	3.12	3.14	1.13	1.13	0.03	0.04	-0.01	1.00
Income	80588.33	81688.73	75075.73	65523.63	2020.90	2097.89	2645.00	0.95
Tenure (%)								
Mortgage/loan	47.35	48.07	47.69	47.73	1.28	1.57	-0.01	1.00
Own free & clear	24.64	24.61	41.85	41.86	1.12	1.21	0.00	0.99
Rent	28.01	27.32	40.17	40.00	1.09	1.55	-0.00	1.03
Recodes & Subgroups								
Income > 50th pctile, %	50.00	50.87	47.69	47.67	1.29	1.25	-0.01	0.99
Income > 75th pctile, %	25.72	27.85	40.80	41.74	1.10	1.05	-0.00	0.94
Income > 90th pctile, %	10.12	12.18	27.15	28.93	0.73	0.66	0.00	0.81
Income (Mortgage=1)	103894.40	104244.70	80618.94	72318.61	3536.80	2991.05	2274.00	0.98
Income (Own=1)	74032.48	71863.12	79477.16	54722.52	4844.70	3182.60	-1749.00	1.06
Income (Rent=1)	47159.14	48156.89	43253.29	42830.20	2495.06	2609.88	3437.00	0.91
Person variables								
Sampling weight	35.37	35.74	21.53	21.20	0.37	0.63	0.49	0.98
Gender (%)	47.92	48.05	49.93	49.93	0.85	0.73	0.03	0.94
Education (%)								
< 12 years	32.46	33.24	46.30	46.69	0.79	1.06	-0.04	1.10
12 years	23.56	22.99	41.64	41.32	0.71	0.86	-0.00	1.03
13-15 years	19.48	19.25	39.38	39.19	0.67	0.79	0.01	0.98
16+ years	24.50	24.53	41.07	41.30	0.70	0.93	-0.02	1.06
Hispanic (%)	9.46	10.32	23.23	25.15	0.41	1.02	-0.01	1.02
Age	39.44	38.85	22.76	31.00	0.39	0.62	9.92	0.76
Race (%)								
White	79.14	77.34	31.68	34.31	0.56	1.18	-0.01	1.04
Race	9.73	10.46	20.86	23.39	0.37	0.86	-0.01	1.04
Other	11.13	12.20	20.86	29.07	0.48	1.07	-0.01	1.02
Poverty (%)	9.08	9.59	26.66	27.55	0.46	0.90	-0.01	1.01
Subgroups								
Poverty (White=1)	8.09	8.39	25.09	25.67	0.60	1.02	-0.00	1.01
Poverty (Black=1)	15.90	16.63	32.76	33.24	3.59	5.36	-0.01	0.99
Poverty (Other=1)	14.88	15.93	32.70	33.60	2.29	3.90	-0.00	0.96
Poverty (Hispanic=1)	16.84	17.83	34.20	34.92	3.12	5.20	-0.01	0.99

characteristics tend to be overestimated in synthetic data. For example, the region-level estimate of the percentage of Hispanics in the synthetic and actual data is 10.31% and 7.97%, respectively; the average percentages of 10.32% and 7.72%, respectively; and the combined percentage of all other races is 12.35% and 9.82%, respectively. The overestimation of the higher-level inferences, though also present to a lesser degree in the PUMA estimates, is likely due to the pooling of neighboring PUMAs when the number of cases with the attribute of interest did not meet the threshold required for producing reliable direct estimates of the regression parameters in Step 1.

Table 2.2 Summary Measures of Actual and Synthetic State Means.

	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Household variables				
Household size	2.22	2.22	0.01	0.01
Sampling weight	33.45	33.06	0.14	0.12
Total bedrooms	2.80	2.81	0.006	0.008
Electricity bill/mo.	117.34	118.35	0.44	0.55
Total rooms	3.14	3.15	0.007	0.009
Income	78316.75	78833.83	431.89	498.09
Tenure (%)				
Mortgage/loan	50.80	51.22	0.31	0.37
Own free & clear	25.43	25.59	0.27	0.29
Rent	23.77	23.19	0.24	0.38
Income > 50th pctile (%)	49.77	50.36	0.30	0.33
Income > 75th pctile (%)	24.50	26.60	0.25	0.28
Income > 90th pctile (%)	9.16	10.90	0.16	0.16
Income (Mortgage=1)	97833.55	98362.78	18149.30	19303.22
Income (Own=1)	70704.19	68179.32	892.40	598.15
Income (Rent=1)	45081.03	45784.78	514.83	602.48
Person variables				
Sampling weight	34.70	34.94	0.09	0.15
Gender (%)	48.27	48.38	0.19	0.15
Education (%)				
< 12 years	31.18	31.90	0.18	0.22
12 years	23.86	23.26	0.17	0.19
13-15 years	20.14	19.72	0.16	0.15
16+ years	24.82	25.11	0.17	0.20
Hispanic (%)	6.63	7.40	0.07	0.17
Age	40.03	39.63	0.87	1.70
Race (%)				
White	85.95	84.17	0.10	0.21
Black	5.84	6.51	0.06	0.12
Other	8.21	9.31	0.09	0.19
Poverty (%)	8.14	8.62	0.10	0.18
Poverty (White=1)	7.32	7.63	0.10	0.19
Poverty (Black=1)	16.13	16.32	1.62	1.98
Poverty (Other=1)	14.41	15.77	0.64	1.19
Poverty (Hispanic = 1)	15.40	15.84	0.91	1.55

Table 2.3 Actual and Synthetic Region Mean.

	Mean		Standard Error of Mean	
	Actual	Synthetic	Actual	Synthetic
Household variables				
Household size	2.29	2.30	0.002	0.002
Sampling weight	32.05	33.49	0.03	0.02
Total bedrooms	2.81	2.79	0.001	0.002
Electricity bill/mo.	124.80	125.63	0.12	0.13
Total rooms	3.17	3.13	0.002	0.002
Income	80670.94	81559.97	113.32	171.51
Tenure (%)				
Mortgage/loan	48.47	47.65	0.07	0.12
Own free & clear	26.11	24.86	0.06	0.06
Rent	25.42	27.49	0.06	0.09
Income > 50th pctile (%)	50.00	50.43	0.07	0.09
Income > 75th pctile (%)	25.47	27.58	0.06	0.07
Income > 90th pctile (%)	10.00	12.08	0.04	0.05
Income (Mortgage=1)	103512.60	106186.80	175.28	264.66
Income (Own=1)	69698.64	68948.59	221.26	128.81
Income (Rent=1)	48384.96	50286.02	145.61	152.03
Person variables				
Sampling weight	33.42	35.81	0.02	0.02
Gender (%)	48.13	48.05	0.04	0.03
Education (%)				
< 12 years	32.15	33.29	0.04	0.07
12 years	24.21	22.94	0.04	0.05
13-15 years	19.62	19.17	0.03	0.04
16+ years	24.02	24.60	0.04	0.05
Hispanic (%)	7.97	10.31	0.02	0.07
Age	39.69	38.85	0.02	0.04
Race (%)				
White	82.46	77.33	0.03	0.08
Black	7.72	10.32	0.02	0.05
Other	9.82	12.35	0.02	0.07
Poverty (%)	8.32	9.39	0.02	0.04
Poverty (White=1)	6.63	7.06	0.02	0.04
Poverty (Black=1)	17.15	17.86	0.11	0.20
Poverty (Other=1)	15.52	16.92	0.09	0.19
Poverty (Hispanic = 1)	19.63	20.85	0.12	0.27

The variability in the synthetic household- and person-level estimates across PUMAs is depicted via scatter plots of actual and synthetic estimates in Figures 2.3 and 2.4, respectively. The estimates lie closely along the 45 degree line for most household- and person-level variables. However, some PUMA estimates deviate from the 45 degree line by a significant margin. For example, synthetic estimates of age (Figure 2.4, middle right-most plot) are overestimated at the extreme values and underestimated in between. This is not surprising due to the bimodal nature of age (see Figure 2.2) which is inadequately accounted for in the parametric imputation model. The bias introduced due to pooling of nearby PUMAs is evident for low-prevalence PUMA estimates (e.g., prevalence of African-Americans, Figure 2.4, bottom row, second-to-left plot) which are significantly overestimated in the synthetic data. Despite a few limitations of the synthetic data, it is encouraging that these imputations result in reasonable PUMA estimates for a wide range of variables.

Figure 2.3 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Household-Level Variables.

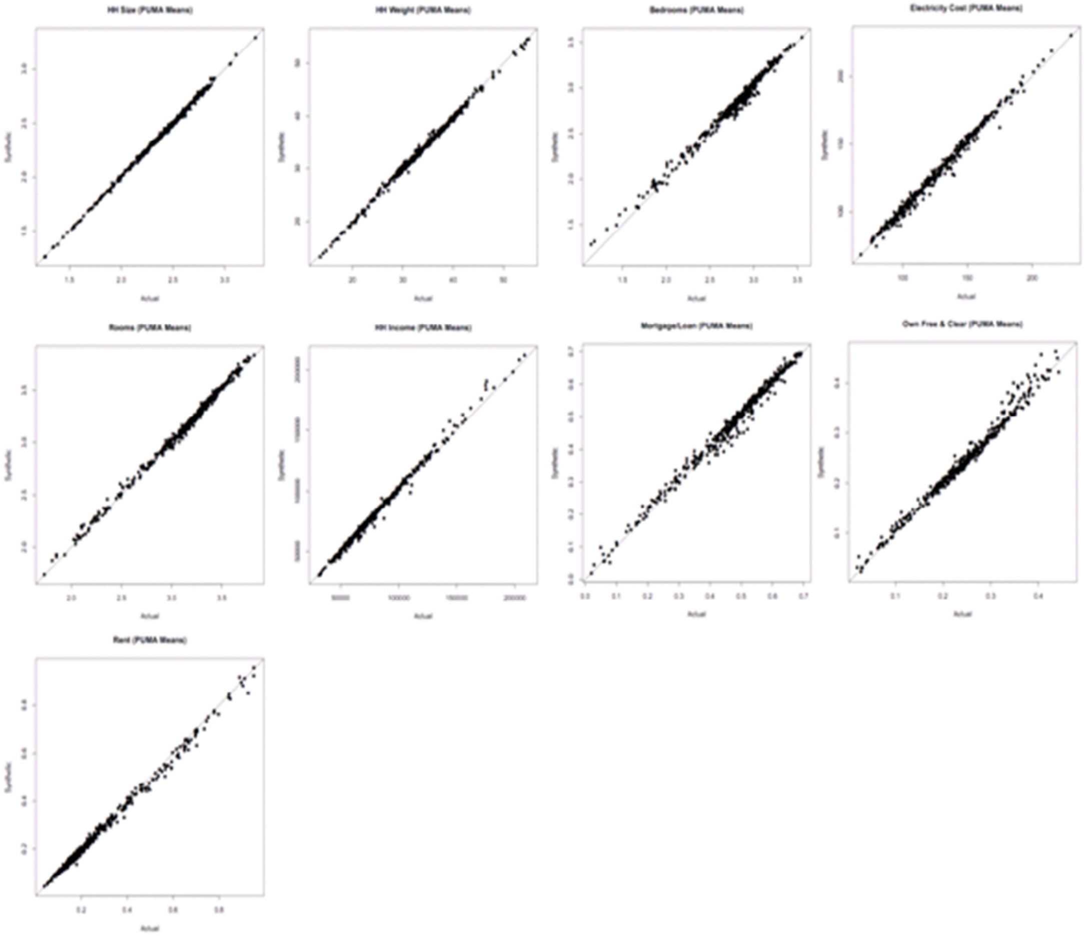
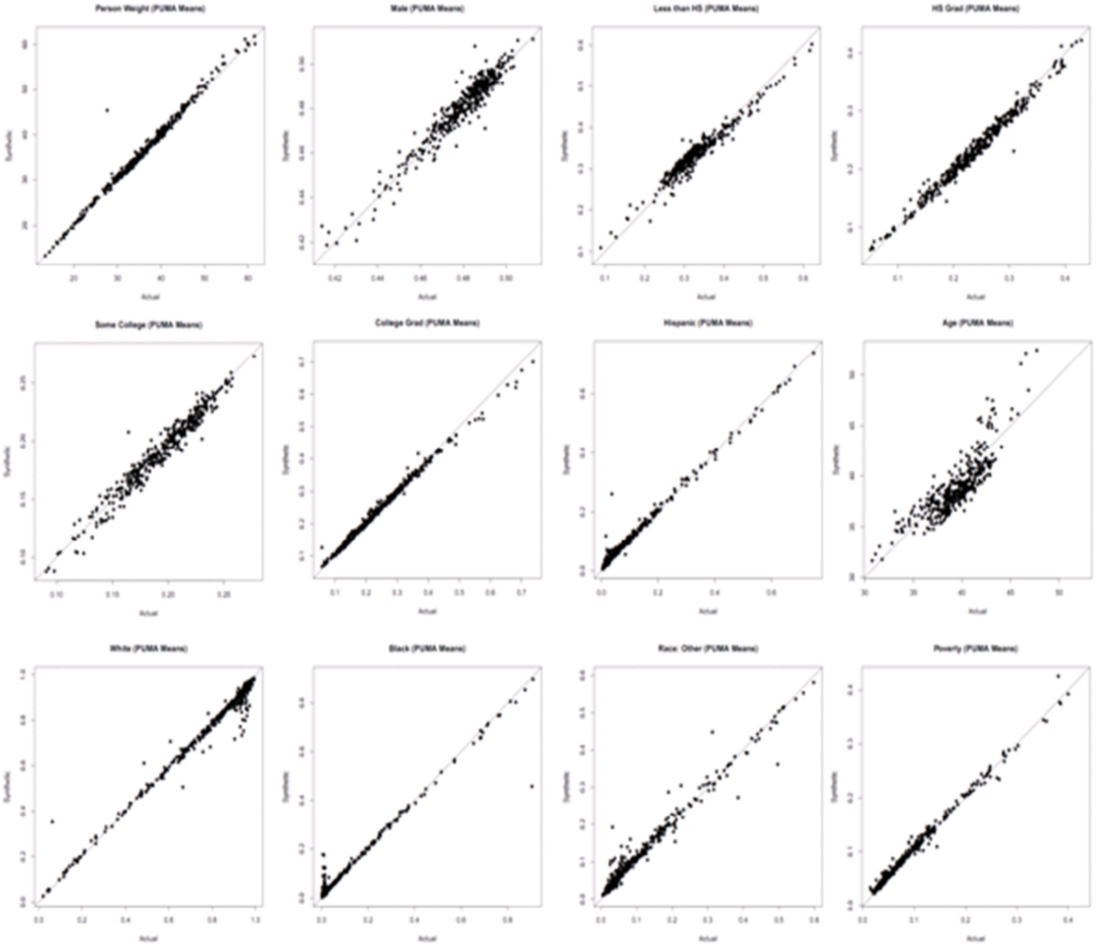


Figure 2.4 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Person-Level Variables.



Scatter plots of synthetic and actual standard deviations of PUMA means shown in Figures 2.5 and 2.6 are another indicator of the quality of the synthetic data. Ideally, each scatter plot point should fall directly on the 45 degree line if the synthetic data accurately reflects the variability of the actual distribution. The results are mixed across variables. For some variables, the synthetic standard deviations are tightly clustered around the 45 degree line, but for other variables, the points exhibit significant variable (or systematic, in the case of age) departures from the line, indicating poor model fit due to a failure of the imputation model.

Figure 2.5 Scatter Plot of Standard Deviations of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Household-Level Variables.

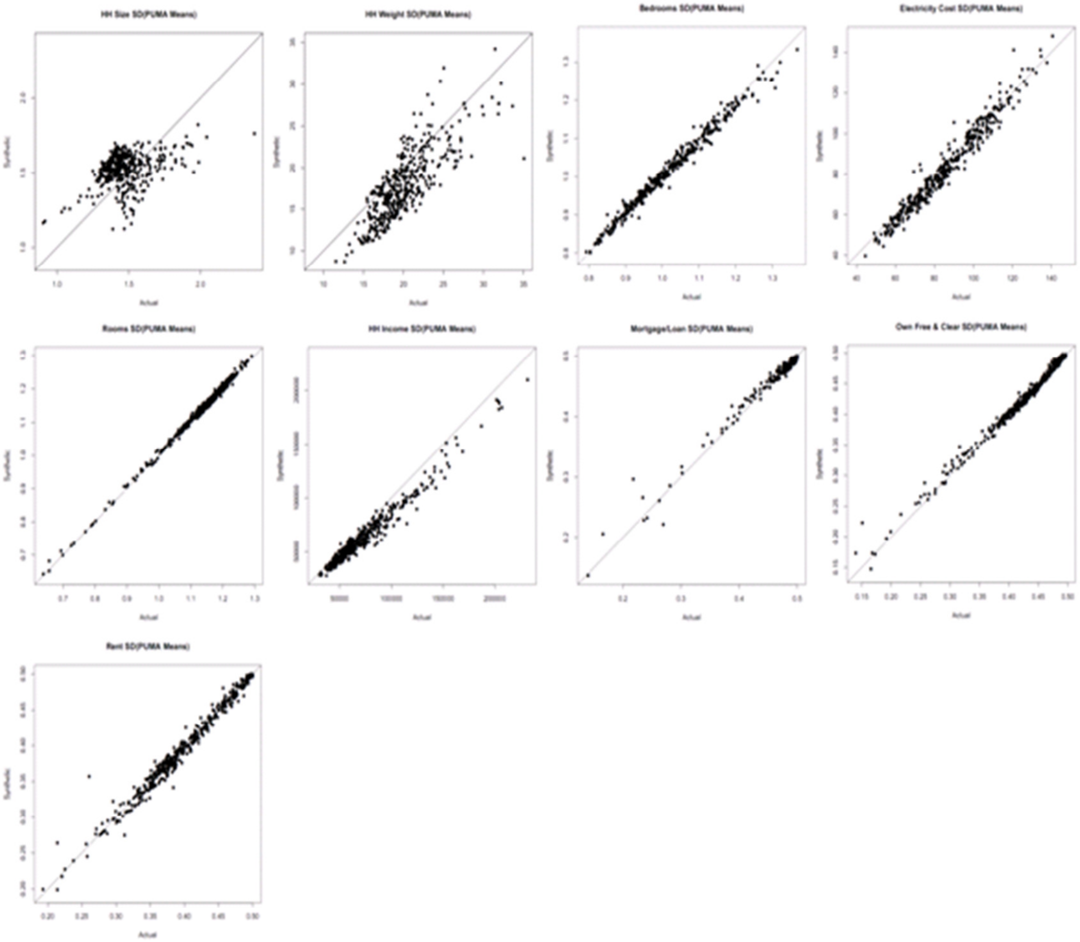
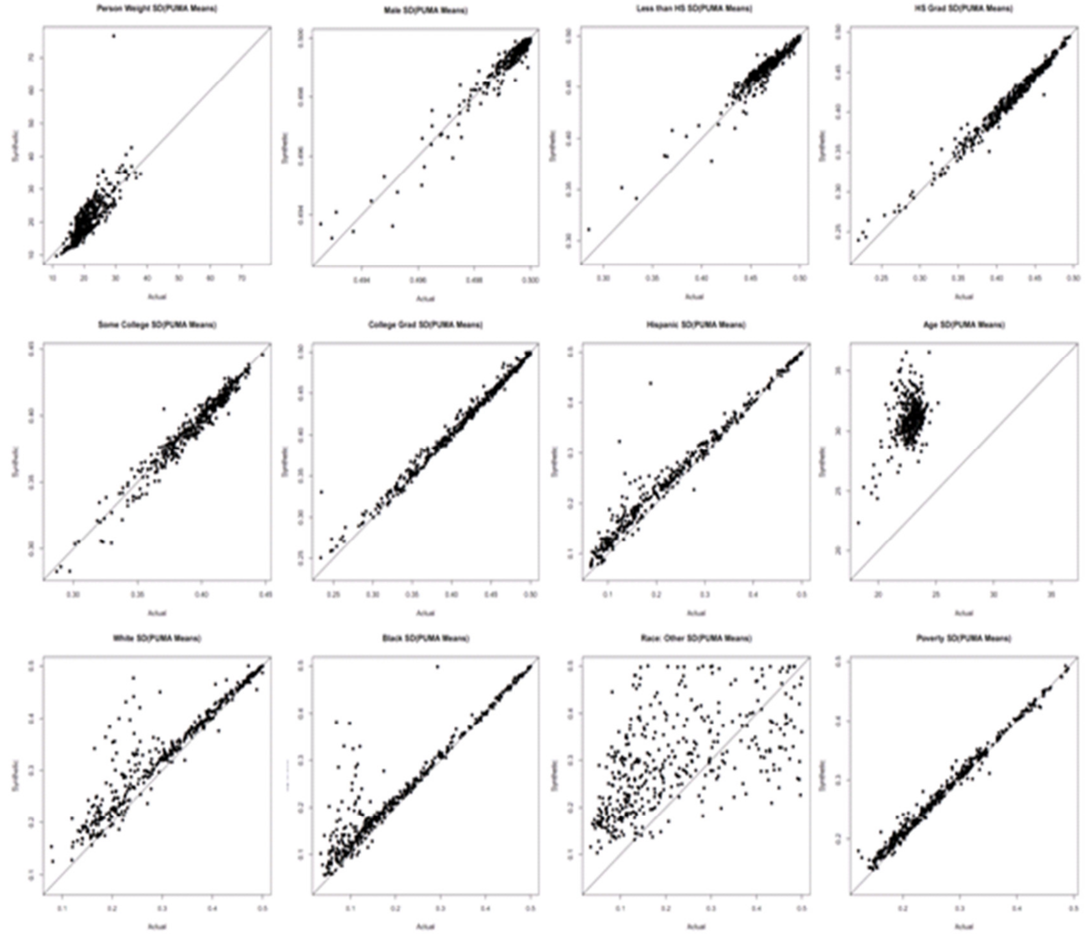


Figure 2.6 Scatter Plot of Standard Deviations of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Person-Level Variables.



Unlike the standard deviations, we would expect the standard errors of the synthetic PUMA means to be larger than the actual standard errors. No auxiliary information was incorporated into the imputation model – all variables used in the imputation model underwent synthesization, yielding a fully synthetic design. Figures 2.7 and 2.8 show scatter plots of the synthetic and actual standard errors. As expected, the synthetic data yield larger standard errors on average for these univariate estimates.

Figure 2.7 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Household-Level Variables.

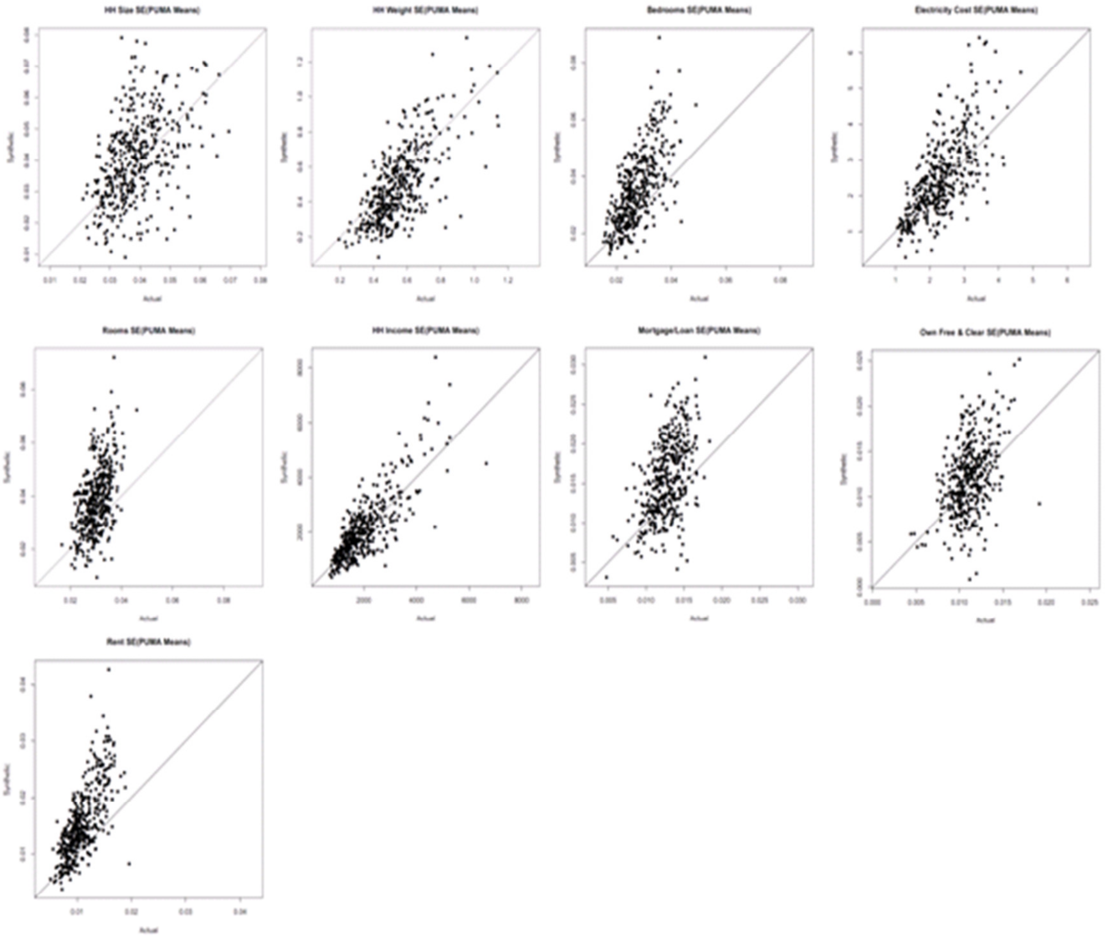
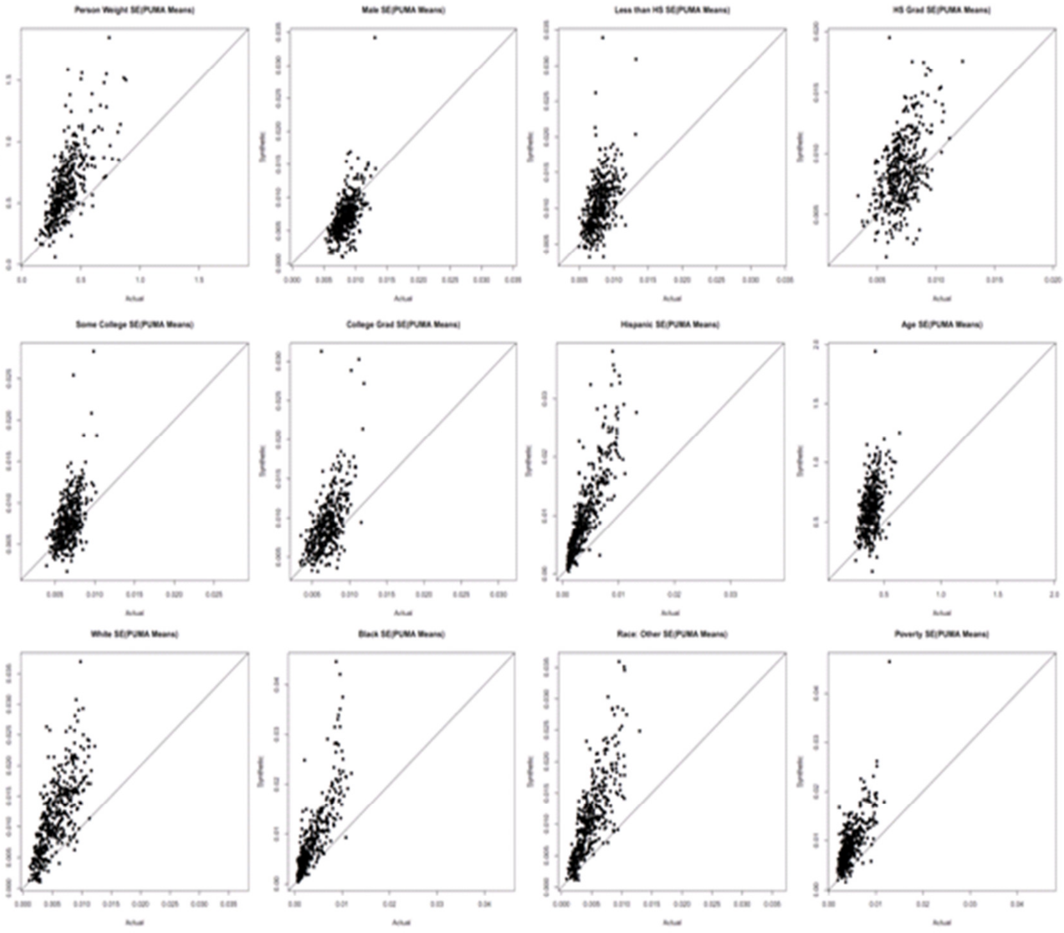


Figure 2.8 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Person-Level Variables.



Next we focus on recoded variable and subgroup estimates. Such estimates are important to data users who may have interest in analyzing particular subsets of the population. Obtaining valid subgroup estimates from synthetic data can be tricky. If the subgroups of interest are not accounted for in the imputation model, then it is unlikely that the resulting subgroup inferences will be useful to the analyst. Thus, obtaining valid subgroup inferences requires that the imputer's model is in agreement or is "congenial" with the analyst's model of interest (Meng, 1994).

Figures 2.9 and 2.10 show scatter plots of PUMA means and standard errors for various recoded variables, including binary indicators of household income greater than the 50th, 75th, and 90th percentiles and subgroups which denote household income by tenure status (mortgage/loan, own free & clear, and rent), and poverty status by race/ethnicity (white, black, other, Hispanic). Starting with the recoded income estimates, the synthetic 50th and 75th percentile estimates correspond well with the actual estimates as indicated by the tightly clustered points which lie about the 45-degree line. On the other hand, for the 90th percentile plot, the points tend to lie above for the 45-degree line for PUMAs with the highest income proportions. Thus, the synthetic data performs reasonably well for estimating less moderate income percentiles, but is somewhat poor for the extreme percentiles.

For both subgroup estimates (income by tenure; poverty by race), the synthetic data does reasonably well. Except for a few outlying points associated with the extreme PUMAs, the majority of points lie along the 45-degree line. This is remarkable considering the joint probabilities associated with these subgroups were not explicitly accounted for in the imputation model, i.e., the imputation model consisted of main effects only and did not include any interactions. The fact that the imputation model can still produce valid subgroup estimates despite being uncongenial to the analyst's model is a reassuring for several reasons. It is difficult for the imputer to foresee how the analyst will use the data. It might not be practical for an imputer, from a computational perspective, to account for all interactions and higher-order terms in the imputation model. Although it is most wise to ensure that all relevant interactions are accounted for during the imputation process, these results suggest that omitted interactions may still yield valid estimates for certain subgroups.

Figure 2.9 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Subgroups and Recoded Variable Groups.

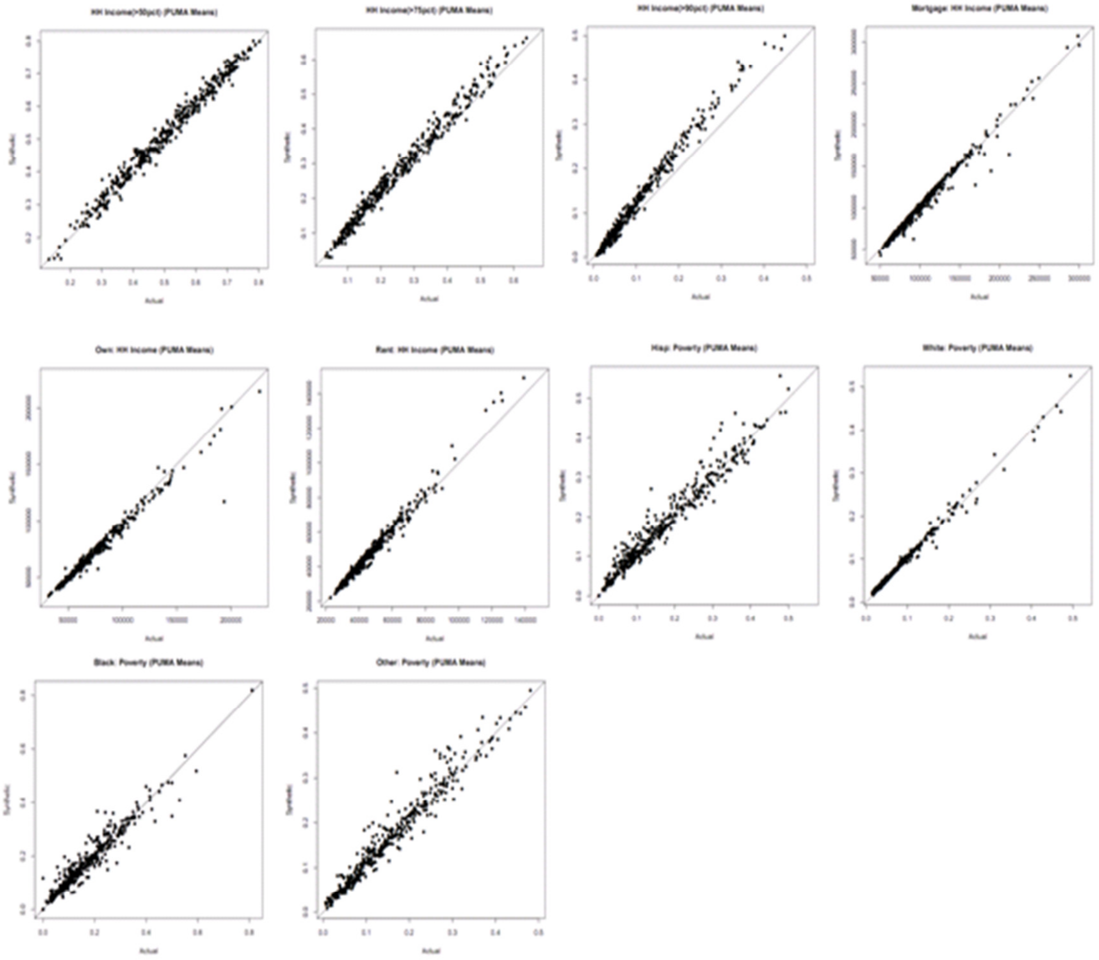
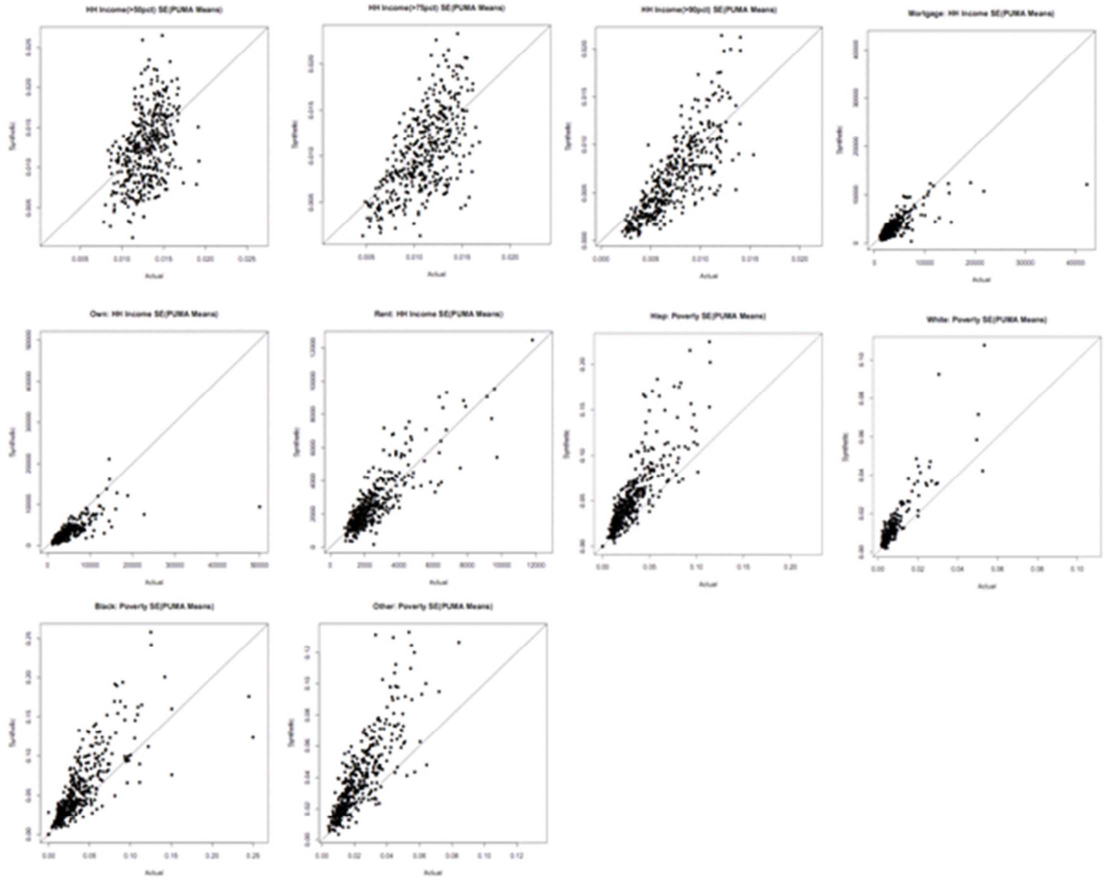


Figure 2.10 Scatter Plot of Standard Errors for Synthetic (y-axis) and Actual (x-axis) PUMA Means for Subgroups and Recoded Variable Groups.



4.2 Validity of Multivariate Estimates

The next set of analyses assesses the analytic validity of synthetic multivariate estimates obtained from several multiple regression models. Tables 2.4, 2.5, and 2.6 show coefficient estimates (and their standard errors) for six regression models (3 household and 3 person) fit at the PUMA-, state-, and region-level, respectively. The state- and region-level models were fitted on data that was aggregated from the PUMAs up to the state and region, respectively. The dependent variable for the household regression models include the continuous income variable and two recoded binary income variables indicating whether the income value met or exceeded the 50th and 75th percentiles, respectively. For the person-level regression models, the binary outcome variables include poverty status and earning a college diploma. Two models are fit for the latter outcome with a squared age term incorporated into one of the models (Model 6).

The synthetic PUMA coefficient estimates correspond reasonably well with the actual estimates, on average. Out of the 24 average synthetic PUMA household-level coefficient estimates, only 5 differ from the actual estimates by more than one average standard error. For the person-level main effects models, all of the PUMA estimates are reliably close to the actual estimates, on average. The squared term model (Model 6) yields a few synthetic estimates that notably differ from the actual estimates, including the squared term itself which is essentially zero (the actual estimate happens to be close to zero as well). This is not surprising considering that the squared term was deliberately omitted from the imputation model; thus, we should expect the synthetic term to be biased towards zero. The average state- and region-level synthetic coefficient estimates are similar to the actual estimates, on average, but like the synthetic univariate estimates obtained from aggregate areas (see section 4.1), they often differ by more than one standard error from the actual estimate indicating weaker correspondence relative to the PUMA estimates.

Table 2.4. PUMA-Level Linear and Logistic Regression Coefficients and Standard Errors Obtained from Actual and Synthetic Data Sets.

	Model 1: Y= Income (linear)		**Model 2: Y= Income (> 50pct; logistic)**		**Model 3: Y=Income (> 75pct; logistic)**	
Household-level covariates	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)
Intercept	26.83 (2.55)	27.28 (2.46)	-2.21 (0.64)	-2.15 (0.44)	-4.15 (0.77)	-3.60 (0.46)
Household size	1.56 (0.24)	1.62 (0.22)	0.40 (0.07)	0.29 (0.04)	0.24 (0.06)	0.29 (0.04)
Sampling weight	-0.39 (0.56)	-0.38 (0.56)	-0.07 (0.15)	-0.06 (0.10)	-0.09 (0.17)	-0.07 (0.10)
Total bedrooms	1.30 (0.35)	1.16 (0.32)	0.19 (0.08)	0.20 (0.06)	0.32 (0.10)	0.21 (0.06)
Electricity bill/mo.	1.05 (0.30)	1.02 (0.27)	0.17 (0.07)	0.18 (0.05)	0.22 (0.08)	0.19 (0.05)
Total rooms	1.31 (0.28)	1.27 (0.27)	0.23 (0.07)	0.22 (0.05)	0.29 (0.08)	0.24 (0.05)
Tenure						
Mortgage/loan	Ref	Ref	Ref	Ref	Ref	Ref
Own free & clear	-3.92 (0.79)	-3.38 (0.70)	-0.87 (0.17)	-0.59 (0.13)	-0.59 (0.20)	-0.62 (0.13)
Rent	-5.81 (0.82)	-6.49 (0.79)	-1.30 (0.21)	-1.14 (0.15)	-1.27 (0.31)	1.24 (0.17)
	Model 4: Y= Poverty (logistic)		**Model 5: Y= College graduate (logistic)**		**Model 6: Y=College graduate (logistic)**	
Person-level covariates	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)
Intercept	-3.14 (0.54)	-3.02 (0.85)	-1.37 (0.34)	-1.00 (0.38)	-4.12 (0.40)	-1.13 (0.38)
Sampling weight	0.29 (0.15)	0.28 (0.23)	-0.12 (0.09)	-0.11 (0.11)	-0.22 (0.10)	-0.11 (0.11)
Gender	-0.32 (0.14)	-0.33 (0.13)	-0.00 (0.10)	-0.03 (0.08)	-0.02 (0.10)	-0.03 (0.08)
Education						
< 12 years	Ref	Ref	--	--	--	--
12 years	-0.29 (0.22)	-0.29 (0.23)	--	--	--	--
13-15 years	-0.56 (0.23)	-0.56 (0.25)	--	--	--	--
16+ years	-1.35 (0.29)	-1.33 (0.32)	--	--	--	--
Hispanic	0.32 (0.34)	0.27 (0.55)	-0.74 (0.32)	0.79 (0.28)	-0.79 (0.34)	-0.78 (0.28)
Age	-0.00 (0.00)	-0.00 (0.01)	0.02 (0.00)	0.01 (0.00)	0.19 (0.01)	0.02 (0.002)
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	-0.01 (0.37)	-0.09 (0.55)	-0.66 (0.34)	-0.53 (0.29)	-0.73 (0.35)	-0.52 (0.29)
Other	0.39 (0.28)	0.37 (0.42)	0.15 (0.21)	0.16 (0.20)	0.17 (0.22)	0.18 (0.21)
Poverty	--	--	-1.14 (0.26)	-1.13 (0.28)	-1.03 (0.26)	-1.12 (0.28)
Age^2	--	--	--	--	-0.00 (0.00)	-0.00 (0.00)

Scatter plots of synthetic and actual PUMA regression coefficients and their standard errors for each of the six models are shown in Appendix 3 (Figures A3.1-A3.12). In general, the synthetic coefficient estimates tend to agree with the actual estimates for the models that do not include recoded variables or higher-order terms (Models 1, 4, and 5). However, agreement tends to decline for models 2, 3, and 6 that involve the modeling of recoded variables (income percentiles) and squared terms (age), which were not explicitly accounted for in the imputation models. For example, it is clear from Figures A3.11 and A3.12 that the squared age term is attenuated towards zero in the synthetic data. To avoid attenuation it is recommended that all relevant squared terms be included in the imputation process. For the percentile regression coefficients, many of the synthetic estimates appear to be valid but the validity tends to decrease for the extreme PUMAs, where the points tend to depart the furthest from the 45 degree line.

Table 2.5. State-Level Linear and Logistic Regression Coefficients and Standard Errors Obtained from Actual and Synthetic Data Sets.

	Model 1: Y= Income (linear)		**Model 2: Y= Income (> 50pct; logistic)**		**Model 3: Y=Income (> 75pct; logistic)**	
Household-level covariates	Avg. Actual $\hat{\beta}_s$ (SE)	Avg. Synthetic $\hat{\beta}_s$ (SE)	Avg. Actual $\hat{\beta}_s$ (SE)	Avg. Synthetic $\hat{\beta}_s$ (SE)	Avg. Actual $\hat{\beta}_s$ (SE)	Avg. Synthetic $\hat{\beta}_s$ (SE)
Intercept	25.64 (0.51)	26.36 (0.54)	-2.43 (0.12)	-2.33 (0.09)	-4.43 (0.15)	-3.80 (0.11)
Household size	1.57 (0.06)	1.65 (0.08)	0.40 (0.01)	0.30 (0.01)	0.23 (0.01)	0.30 (0.01)
Sampling weight	-0.28 (0.10)	-0.26 (0.10)	-0.05 (0.03)	-0.04 (0.02)	-0.07 (0.03)	-0.05 (0.02)
Total bedrooms	1.26 (0.08)	1.13 (0.07)	0.19 (0.02)	0.20 (0.01)	0.33 (0.02)	0.21 (0.01)
Electricity bill/mo.	1.07 (0.07)	1.01 (0.08)	0.17 (0.02)	0.18 (0.01)	0.22 (0.02)	0.19 (0.01)
Total rooms	1.42 (0.06)	1.33 (0.06)	0.25 (0.01)	0.24 (0.01)	0.32 (0.01)	0.25 (0.01)
Tenure						
Mortgage/loan	Ref	Ref	Ref	Ref	Ref	Ref
Own free & clear	-3.66 (0.15)	-3.15 (0.14)	-0.84 (0.03)	-0.56 (0.02)	-0.53 (0.04)	-0.60 (0.03)
Rent	-5.64 (0.17)	-6.36 (0.21)	-1.31 (0.04)	-1.14 (0.04)	-1.29 (0.07)	-1.25 (0.05)
	Model 4: Y= Poverty (logistic)		**Model 5: Y= College graduate (logistic)**		**Model 6: Y=College graduate (logistic)**	
Person-level covariates	Avg. Actual $\hat{\beta}_s$ (SE)	Avg. Synthetic $\hat{\beta}_s$ (SE)	Avg. Actual $\hat{\beta}_s$ (SE)	Avg. Synthetic $\hat{\beta}_s$ (SE)	Avg. Actual $\hat{\beta}_s$ (SE)	Avg. Synthetic $\hat{\beta}_s$ (SE)
Intercept	-3.30 (0.09)	-3.17 (0.19)	-1.60 (0.06)	-1.11 (0.09)	-4.32 (0.07)	-1.27 (0.09)
Sampling weight	0.33 (0.03)	0.31 (0.03)	-0.09 (0.02)	-0.10 (0.35)	-0.18 (0.02)	-0.10 (0.40)
Gender	-0.30 (0.03)	-0.29 (0.03)	-0.02 (0.02)	-0.04 (0.02)	-0.03 (0.02)	-0.04 (0.02)
Education						
< 12 years	Ref	Ref	--	--	--	--
12 years	-0.28 (0.05)	-0.29 (0.05)	--	--	--	--
13-15 years	-0.52 (0.05)	-0.48 (0.06)	--	--	--	--
16+ years	-1.31 (0.06)	-1.30 (0.07)	--	--	--	--
Hispanic	0.36 (0.09)	0.21 (0.16)	-0.55 (0.08)	-0.61 (0.07)	-0.55 (0.09)	-0.60 (0.07)
Age	-0.001 (0.001)	-0.003 (0.001)	0.02 (0.004)	0.01 (0.003)	0.19 (0.002)	0.02 (0.0004)
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	-0.49 (0.10)	-0.64 (0.30)	-0.66 (0.11)	-0.51 (0.10)	-0.67 (0.11)	-0.49 (0.11)
Other	0.42 (0.06)	0.44 (0.10)	0.09 (0.05)	0.05 (0.05)	0.12 (0.05)	0.07 (0.05)
Poverty	--	--	-1.10 (0.05)	-1.10 (0.07)	-0.99 (0.05)	-1.09 (0.07)
Age^2	--	--	--	--	-0.002 (0.000)	-0.0e-7 (0.0e-8)

Table 2.6. Region-Level Linear and Logistic Regression Coefficients and Standard Errors Obtained from Actual and Synthetic Data Sets.

	Model 1: Y= Income (linear)		**Model 2: Y= Income (> 50pct; logistic)**		**Model 3: Y=Income (> 75pct; logistic)**	
Household-level covariates	Actual $\hat{\beta}$ (SE)	Synthetic $\hat{\beta}$ (SE)	Actual $\hat{\beta}$ (SE)	Synthetic $\hat{\beta}$ (SE)	Actual $\hat{\beta}$ (SE)	Synthetic $\hat{\beta}$ (SE)
Intercept	20.78 (0.11)	22.94 (0.12)	-3.00 (0.02)	-2.55 (0.02)	-5.13 (0.03)	-4.12 (0.02)
Household size	1.40 (0.01)	1.61 (0.01)	0.32 (0.00)	0.26 (0.00)	0.19 (0.00)	0.24 (0.00)
Sampling weight	0.91 (0.02)	0.51 (0.03)	0.19 (0.01)	0.09 (0.00)	0.22 (0.01)	0.11 (0.01)
Total bedrooms	1.08 (0.02)	0.83 (0.02)	0.12 (0.00)	0.12 (0.00)	0.26 (0.00)	0.15 (0.00)
Electricity bill/mo.	1.58 (0.02)	1.53 (0.02)	0.23 (0.00)	0.21 (0.00)	0.29 (0.00)	0.27 (0.00)
Total rooms	1.37 (0.01)	1.30 (0.02)	0.23 (0.00)	0.21 (0.00)	0.28 (0.00)	0.22 (0.00)
Tenure						
Mortgage/loan	Ref	Ref	Ref	Ref	Ref	Ref
Own free & clear	-4.27 (0.04)	-3.79 (0.04)	-0.90 (0.01)	-0.63 (0.01)	-0.64 (0.00)	-0.62 (0.01)
Rent	-5.91 (0.04)	-6.53 (0.04)	-1.21 (0.01)	-1.07 (0.01)	-0.96 (0.01)	-0.96 (0.01)
	Model 4: Y= Poverty (logistic)		**Model 5: Y= College graduate (logistic)**		**Model 6: Y=College graduate (logistic)**	
Person-level covariates	Actual $\hat{\beta}$ (SE)	Synthetic $\hat{\beta}$ (SE)	Actual $\hat{\beta}$ (SE)	Synthetic $\hat{\beta}$ (SE)	Actual $\hat{\beta}$ (SE)	Synthetic $\hat{\beta}$ (SE)
Intercept	-2.52 (0.02)	-2.60 (0.04)	-2.45 (0.01)	-1.61 (0.01)	-5.23 (0.01)	-1.76 (0.01)
Sampling weight	0.15 (0.01)	0.19 (0.01)	0.19 (0.00)	0.08 (0.00)	0.15 (0.00)	0.08 (0.00)
Gender	-0.30 (0.01)	-0.31 (0.01)	0.01 (0.00)	-0.02 (0.01)	-0.02 (0.00)	-0.02 (0.01)
Education						
< 12 years	Ref	Ref	--	--	--	--
12 years	-0.37 (0.01)	-0.33 (0.01)	--	--	--	--
13-15 years	-0.66 (0.01)	-0.59 (0.01)	--	--	--	--
16+ years	-1.47 (0.01)	-1.34 (0.02)	--	--	--	--
Hispanic	0.70 (0.01)	0.70 (0.02)	-1.00 (0.01)	-1.03 (0.01)	-1.09 (0.01)	1.03 (0.01)
Age	-0.00 (0.00)	-0.00 (0.00)	0.02 (0.00)	0.01 (0.00)	0.18 (0.00)	0.02 (0.00)
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	0.89 (0.01)	0.88 (0.02)	-0.61 (0.01)	-0.64 (0.01)	-0.68 (0.01)	-0.63 (0.01)
Other	0.51 (0.01)	0.54 (0.01)	0.42 (0.01)	0.34 (0.01)	0.42 (0.01)	0.36 (0.01)
Poverty	--	--	-1.19 (0.01)	-1.11 (0.02)	-1.08 (0.01)	-1.10 (0.02)
Age^2	--	--	--	--	-0.00 (0.00)	-0.00 (0.0)

4.3 Propensity Score Balance

Another indicator of the quality of the synthetic data is to assess the covariate balance between the synthetic and actual data. This is most easily performed using propensity scores (Rubin and Rosenbaum, 1983). Propensity scores are often used to identify imbalances in in two or more groups (e.g., treatment and control groups) based on the distribution of a set of observed covariates. Biases caused by covariate imbalances may be adjusted by performing a weighted analysis with weights inversely proportional to the propensity scores (Ekholm and Laaksonen, 1991).

To assess the covariate balance between the synthetic and actual data sets, a randomly selected synthetic data set and the actual data are stacked vertically. Then an actual data indicator variable is regressed against all synthetic and actual variables using a logistic regression model. The fitted model is used to obtain estimates of the propensity of a record belonging to the actual data. The propensity scores are then sorted and classified into deciles and the proportions of synthetic and actual records are compared. If the synthetic and actual covariates are fully balanced, then the proportion of synthetic versus actual data should be the same for each decile group. A chi-squared test with 9 degrees of freedom (if deciles are used) can be performed to assess the equivalence of the actual data proportions across the groups.

We use the propensity score balance method to assess the similarity of the synthetic and actual data in each PUMA. Tables 2.7 and 2.8 show summary statistics of the estimated probabilities of belonging to the actual data in each PUMA obtained from the household- and person-level propensity models, respectively, and associated test statistics. The overall mean estimated propensity score was 0.13, which reflects the true proportion of actual data in each PUMA and the oversampling of synthetic data. Within each PUMA, the propensity scores were sorted and grouped into deciles and a chi-square statistic was computed. Small chi-square values indicate that the synthetic and actual data sets are balanced or statistically independent from each other, based on the set of covariates, while large values indicate poor covariate balance between the two data sets. The mean chi-square p-value for the household- and person-level data was 0.02 and 0.001, respectively. This suggests that the synthetic data is not statistically balanced with the actual data based on the set of synthetic covariates. These results should be interpreted with caution, however, as the large sample sizes tend to produce overpowered tests. In addition, the independence assumption is violated between the two data sets (Raghunathan, 2008).

Table 2.7 Estimated Propensities of Belonging to the Actual Household-Level Data

Households; PUMAs	Mean	Min	Max
Estimated probabilities $\hat{p}$	0.13	0.08	0.20
χ^2 statistic	45.38	27.03	182.09
P-value	0.02	<0.000	0.14

Table 2.8 Estimated Propensities of Belonging to the Actual Person-Level Data

Persons; PUMAs	Mean	Min	Max
Estimated probabilities $\hat{p}$	0.13	0.06	0.17
χ^2 statistic	216.71	96.95	625.27
P-value	0.001	< 0.000	0.003

5 ACS-Based Simulation

This section evaluates the repeated sampling properties for small area inferences drawn from the synthetic data based on a simulation application. In this simulation, the 2005-2007 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. A total of 50,000 synthetic data sets are generated.

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 most commonly encountered 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 is obtained for each variable in each PUMA using the combining rule equations [1] and [2] in 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 estimates are obtained again 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. $Q_{CIC} = 1$ 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 by,

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

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 from the 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 the synthetic interval completely covers the actual interval. Calculating the confidence interval overlap is only possible for conditional, not unconditional, 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 2.9 shows the average confidence interval coverage (CIC) and confidence interval overlap (CIO) across all PUMAs for each household-level estimate. The conditional CIC is high for basic (non-recoded) estimates ranging from 0.86-0.99. The income/tenure subgroup estimates also yield relatively high conditional CIC values (range: 0.89-0.97). For the income percentile estimates, the CIC values 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 2.9 Simulation-Based Confidence Interval Results for PUMA Means.

	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
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
Income > 50th pctile	0.89	0.92	0.94	0.98
Income > 75th pctile	0.71	0.71	0.80	0.98
Income > 90th pctile	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

Although the synthetic PUMA means exhibit good confidence interval properties, the CIC and CIO values are less impressive for the state-level means. Table 2.10 shows the average CIC and CIO values across all states. The conditional CIC and CIO measures range from 0.18-0.88 and 0.29-0.99, respectively. The CIO values tend to be relatively higher than the CIC values suggesting that these estimates have higher data utility than their corresponding CIC values might indicate. The same pattern is generally true for the unconditional inference. The unconditional synthetic CIC values fail to reach the actual CIC values by a notable margin for all estimates.

Table 2.10 Simulation-Based Confidence Interval Results for State Means.

	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
Household size	0.88	0.89	0.90	0.98
Sampling weight	0.67	0.99	0.76	0.99
Bedrooms	0.25	0.60	0.37	0.98
Electricity cost/mo.	0.18	0.60	0.31	0.98
Rooms	0.67	0.76	0.75	0.98
Household income	0.42	0.75	0.56	0.99
Tenure				
Own free & clear	0.64	0.70	0.73	0.98
Rent	0.60	0.85	0.70	0.98
Income > 50th pctile	0.68	0.85	0.68	0.99
Income > 75th pctile	0.27	0.29	0.34	0.98
Income > 90th pctile	0.40	0.47	0.36	0.98
Income (Mortgage=1)	0.58	0.75	0.61	0.98
Income (Own=1)	0.53	0.99	0.63	0.98
Income (Rent=1)	0.81	0.80	0.87	0.98

5.2 Validity of Multivariate Estimates

Multivariate simulation results are shown in Table 2.11. This table shows average CIC and CIO values for regression coefficient estimates obtained within each PUMA from a household-level regression model. 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 the CIC values obtained from the actual data (0.98). Because the analytic model being evaluated here is the same model used to impute the synthetic data, it is not surprising that the analytic validity of the estimates is high. This result underscores the importance of ensuring that the imputation model sufficiently overlaps with the analytic small area model of interest.

Additional simulation-based summary measures PUMA-, state-, and region-level estimands can be found in Appendix 4.

Table 2.11 Simulation-Based Confidence Interval Results for PUMA Regression Coefficients

	Conditional Inference		Unconditional Inference	
Covariates	CIC	CIO	CIC	CIC (Actual)
Regression of income (cube root) on				
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 Application: Restricted ACS County-Level Data

In addition to the public-use microdata, restricted ACS microdata for years 2005-2009 were obtained from the Michigan Census Research Data Center and used to demonstrate the proposed synthetic data method. The restricted data contain identifiers for all counties in the United States. We restrict the data to the Northeast region which contains 217 counties, in contrast to the public-use microdata which contains 405 public-use microdata areas (PUMAs). Although 3 years of microdata were used in the public-use application, we use the restricted 5-year data set to facilitate the disclosure review and allow the publication of estimates for all counties. The same variables shown in Table 2.0 were synthesized in this application. The synthetic data estimates are based on $M = 10$ imputations.

Tables 2.12 and 2.13 show summary measures of actual and synthetic county means and regression coefficients. In general and without going into great detail, the synthetic means and regression coefficients correspond relatively closely to the actual estimates, on average, as was found for the public-use application (Sections 4.1 and 4.2).

Figures 2.11 and 2.12 present scatter plots of the actual and synthetic means for all counties in the Northeast region. (Plots of county-level regression estimates are not shown, but yielded similar correspondence as was shown for the PUMA estimates). In general, the correspondence between actual and synthetic means is reasonably good as indicated by the points lying closely along the 45-degree line. Overall, the results of the restricted-data application are similar to the public-use application.

As in the public-use application, the actual and synthetic point estimates correspond relatively closely when applied to actual counties, with the aforementioned exceptions (e.g., bimodal age variable). This finding should give confidence to the synthetic data methodology, as the method is more practically useful when applied to actual small areas, such as counties, as opposed to combined counties or PUMAs.

Table 2.12 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
Income > 50th pctile,%	44.65	44.56	48.24	48.19	0.80	0.56	0.01	0.97
Income > 77th 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
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

Table 2.13 County-Level Linear and Logistic Regression Coefficients and Standard Errors Obtained from Actual and Synthetic Data Sets

	Y=Income (linear)		**Y=Income (>50pct; logistic)**		**Y=Income (>75pct; logistic)**	
Household-level covariates	Actual Beta (SE)	Synthetic Beta (SE)	Actual Beta (SE)	Synthetic Beta (SE)	Actual Beta (SE)	Synthetic Beta (SE)
Intercept	24.34 (1.11)	24.26 (1.09)	-2.86 (0.29)	-2.82 (0.23)	-5.15 (0.39)	-4.42 (0.28)
Household size	1.52 (0.14)	1.44 (0.14)	0.37 (0.04)	0.28 (0.03)	0.21 (0.04)	0.29 (0.03)
Sampling weight	-0.04 (0.24)	-0.05 (0.26)	0.006 (0.07)	-0.01 (0.05)	0.03 (0.09)	-0.01 (0.07)
Total bedrooms	1.15 (0.19)	1.23 (0.18)	0.19 (0.05)	0.24 (0.04)	0.34 (0.06)	0.25 (0.05)
Electricity bill/mo.	0.99 (0.18)	1.04 (0.17)	0.18 (0.05)	0.20 (0.04)	0.24 (0.06)	0.21 (0.04)
Total rooms	1.25 (0.14)	1.26 (0.13)	0.25 (0.04)	0.24 (0.03)	0.32 (0.05)	0.26 (0.04)
Tenure						
Mortgage/loan	Ref	Ref	Ref	Ref	Ref	Ref
Own free & clear	-3.47 (0.37)	-3.05 (0.34)	-0.80 (0.09)	-0.57 (0.08)	-0.52 (0.12)	-0.62 (0.10)
Rent	-6.01 (0.44)	-6.84 (0.47)	-1.45 (0.14)	-1.31 (0.14)	-1.45 (0.26)	-1.57 (0.31)
	Y=Poverty (logistic)		**Y=College graduate (logistic)**		**Y=College graduate (logistic)**	
Person-level covariates	Actual Beta (SE)	Synthetic Beta (SE)	Actual Beta (SE)	Synthetic Beta (SE)	Actual Beta (SE)	Synthetic Beta (SE)
Intercept	-2.39 (0.16)	-2.32 (0.24)	-2.27 (0.12)	-2.17 (0.13)	-4.99 (0.18)	-2.18 (0.14)
Sampling weight	0.25 (0.07)	0.25 (0.10)	0.03 (0.05)	0.03 (0.05)	-0.00 (0.05)	0.03 (0.05)
Gender: Male	-0.33 (0.08)	-0.34 (0.08)	-0.06 (0.06)	-0.06 (0.05)	-0.08 (0.06)	-0.06 (0.05)
Education						
<12 years	Ref	Ref	--	--	--	--
12 years	-0.36 (0.12)	-0.35 (0.13)	--	--	--	--
13-15 years	-0.62 (0.13)	-0.63 (0.15)	--	--	--	--
16+years	-1.52 (0.18)	-1.59 (0.30)	--	--	--	--
Hispanic	0.36 (0.29)	0.27 (0.63)	-0.70 (0.34)	-0.66 (0.67)	-0.66 (0.36)	-0.66 (0.67)
Age	-0.00 (0.00)	0.01 (0.07)	0.02 (0.001)	0.02 (0.05)	0.17 (0.007)	0.02 (0.05)
Race						
White	Ref	Ref	Ref	Ref	Ref	Ref
Black	0.28 (0.34)	0.22 (0.87)	-1.06 (0.36)	-0.65 (0.80)	-1.01 (0.38)	-0.65 (0.80)
Other	0.41 (0.25)	0.41 (0.56)	0.23 (0.24)	0.33 (0.36)	0.21 (0.25)	0.33 (0.36)
Poverty	--	--	-1.26 (0.17)	-1.26 (0.28)	-1.15 (0.17)	-1.26 (0.28)
Age (squared)	--	--	--	--	-0.00 (0.00)	-0.01 (0.02)

Figure 2.11 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) County Means for Household-Level Variables.

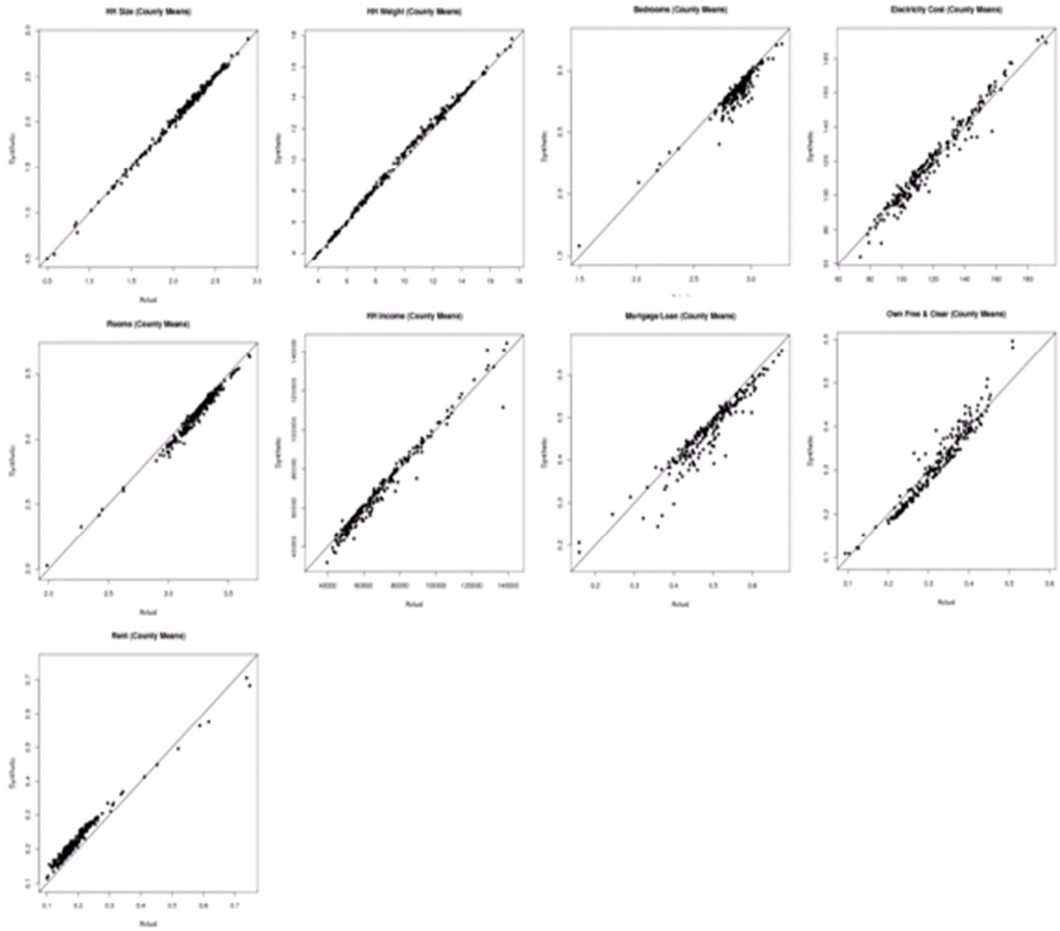
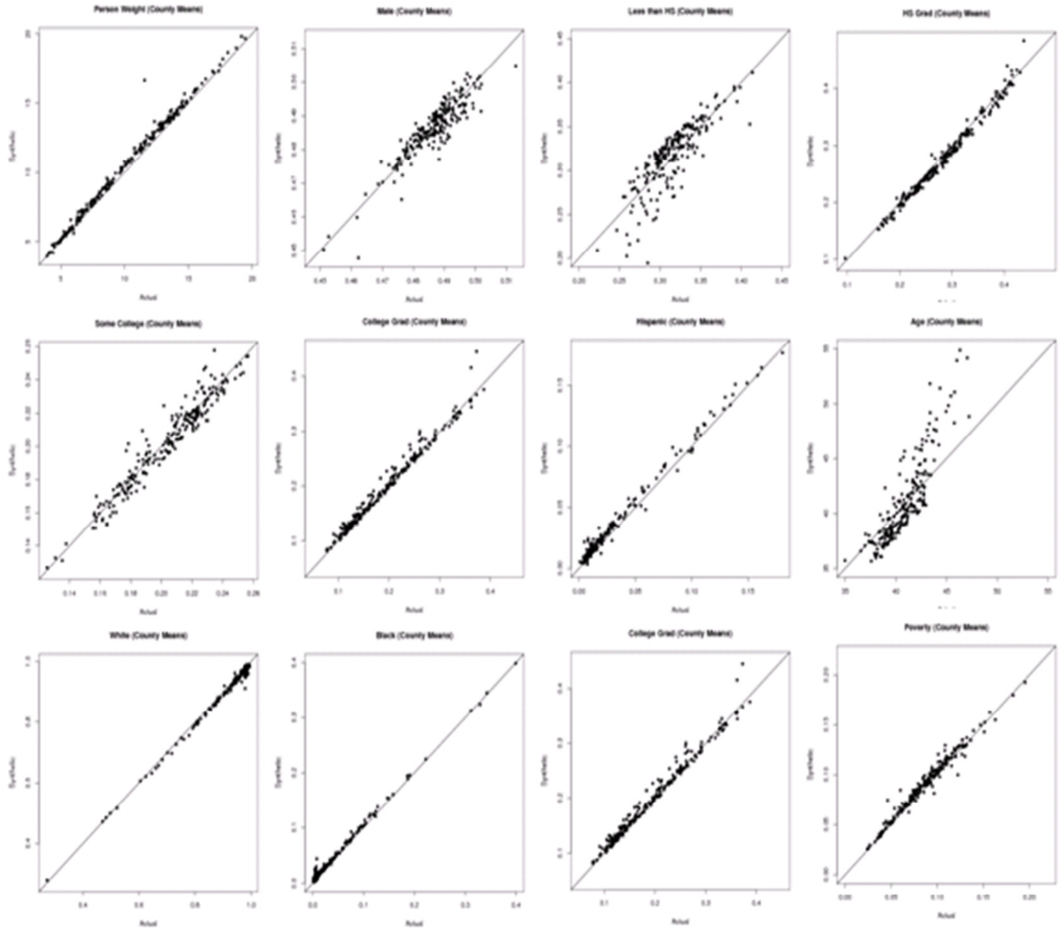


Figure 2.12 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) County Means for Person-Level Variables.



7 Conclusions

In this chapter, I demonstrated a new synthetic data methodology for disseminating public-use microdata for small geographic areas. Data users are increasingly interested in producing small area estimates, but statistical agencies are prevented from releasing these data due to disclosure concerns. Compared with current practices of disseminating small area data via research data centers, geographically suppressed public-use microdata, and summary/aggregate tables, 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 the Bayesian hierarchical model produces both valid univariate and multivariate statistics computed within the smallest geographic areas. However, limitations of the method were apparent when producing estimates for larger (aggregate) areas and simulating synthetic data for non-standard distribution; both situations yielded low analytic validity. The low analytic validity for state- and region-level estimates could be attributable to the choice of covariates incorporated into the hierarchical model. Only 3 state-level covariates (number of metropolitan and micropolitan areas, and log population size) were used in this demonstration, but a broader set of variables that are highly correlated with the variables undergoing synthesis may yield improvements. In addition, the “empirical” Bayesian approach considered here by fixing the hyperparameters at their maximum likelihood estimates may have underestimated the synthetic standard errors and shortened confidence intervals to the extent that they did not adequately cover the actual estimate of interest at reasonable rates. A fully-Bayesian approach, accounting for the variation in the hyperparameters, might improve confidence interval coverage of estimates computed for aggregate areas.

Regarding the preservation of skewed and non-standard distributions, parametric imputation models are inherently limited in this task as demonstrated in this study. Extending the proposed methodology to handle nonparametric distributions is a natural next step and a fruitful area for future work. Although the ACS samples all geographically-relevant areas, another possible extension is the generation of synthetic data for non-sampled small areas in complex sample surveys (e.g., NHIS).

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

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 $p = (1,2,\ldots,P)$ 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,

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

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

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,

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

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

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

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,

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

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

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

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 α_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,

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

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

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,

$\tilde{Z}_{hcs} \sim Poisson\left(\tilde{\lambda}_{cs}\right)$.

Appendix 3 Scatter Plots of Synthetic and Actual PUMA Regression Coefficients

Figure A3.1 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Household Income on Basic Household Characteristics.

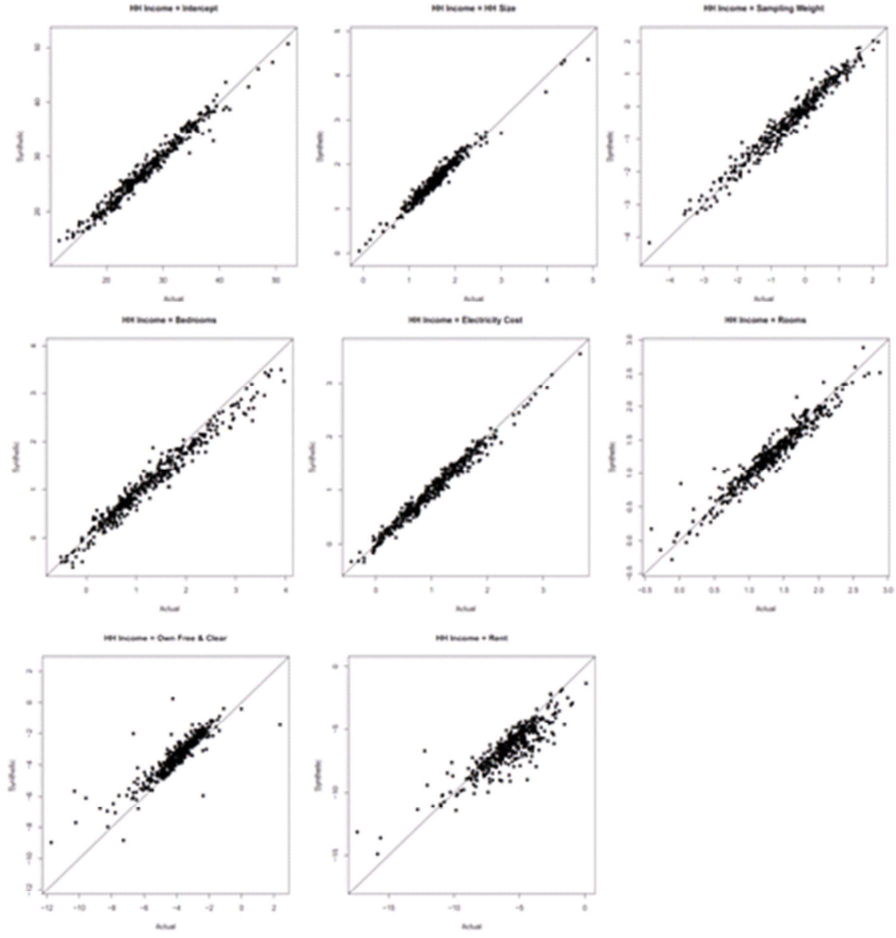


Figure A3.2 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Household Income on Basic Household Characteristics.

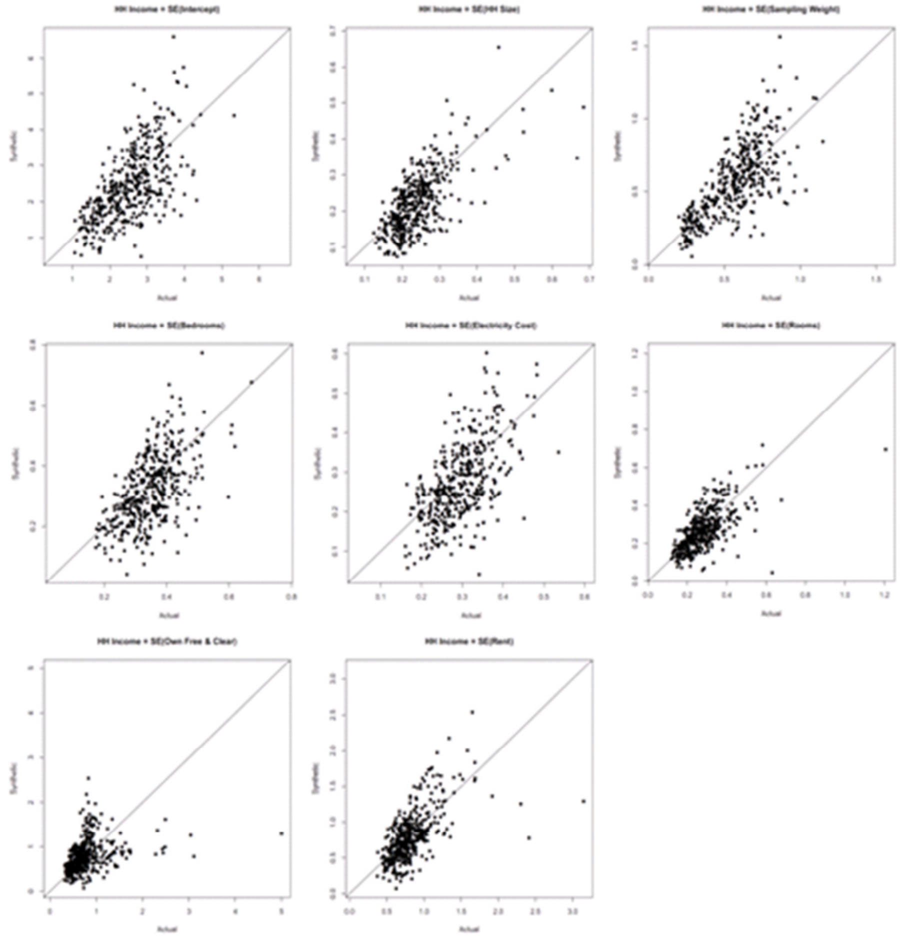


Figure A3.3 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Household Income Greater than the 50th Percentile on Basic Household Characteristics.

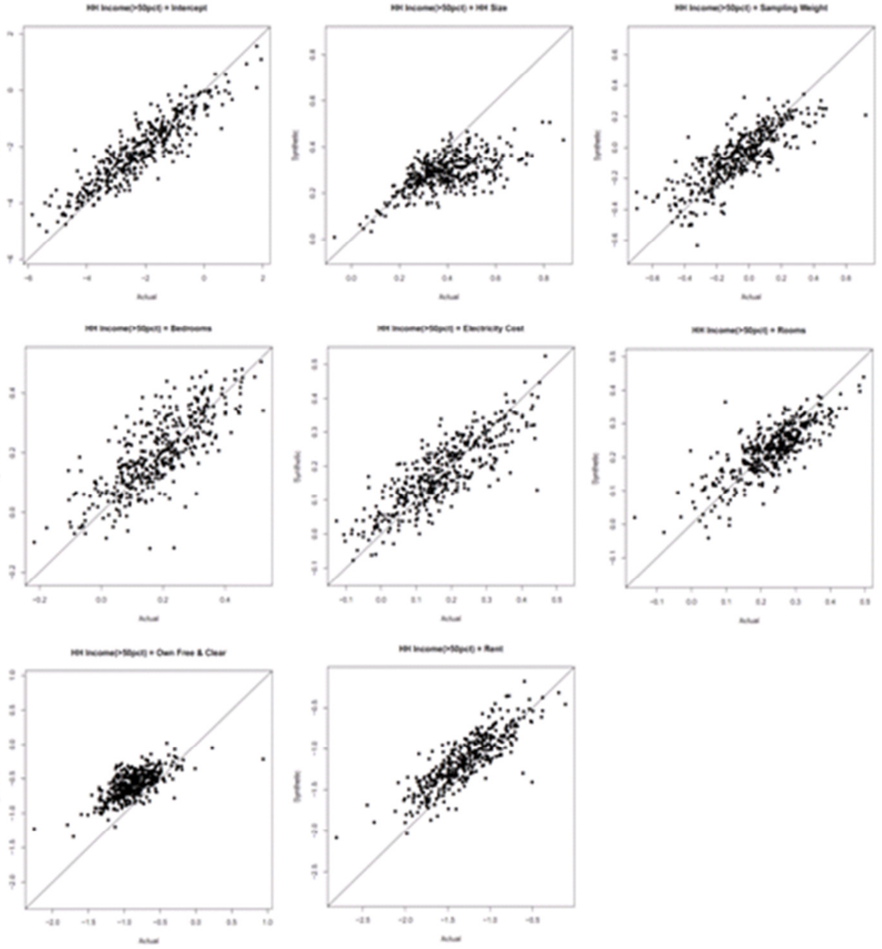


Figure A3.4 Scatter Plot of Standard errors of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Household Income Greater than the 50^{th} Percentile on Basic Household Characteristics.

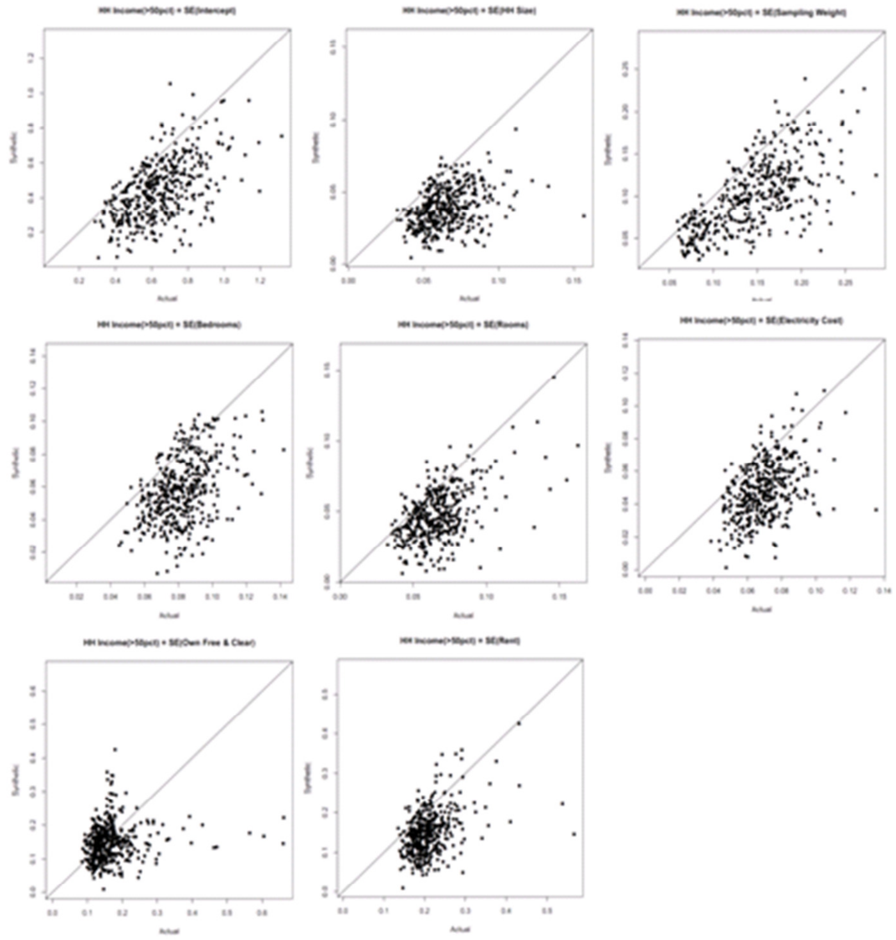


Figure A3.5 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Household Income Greater than the 75th Percentile on Basic Household Characteristics.

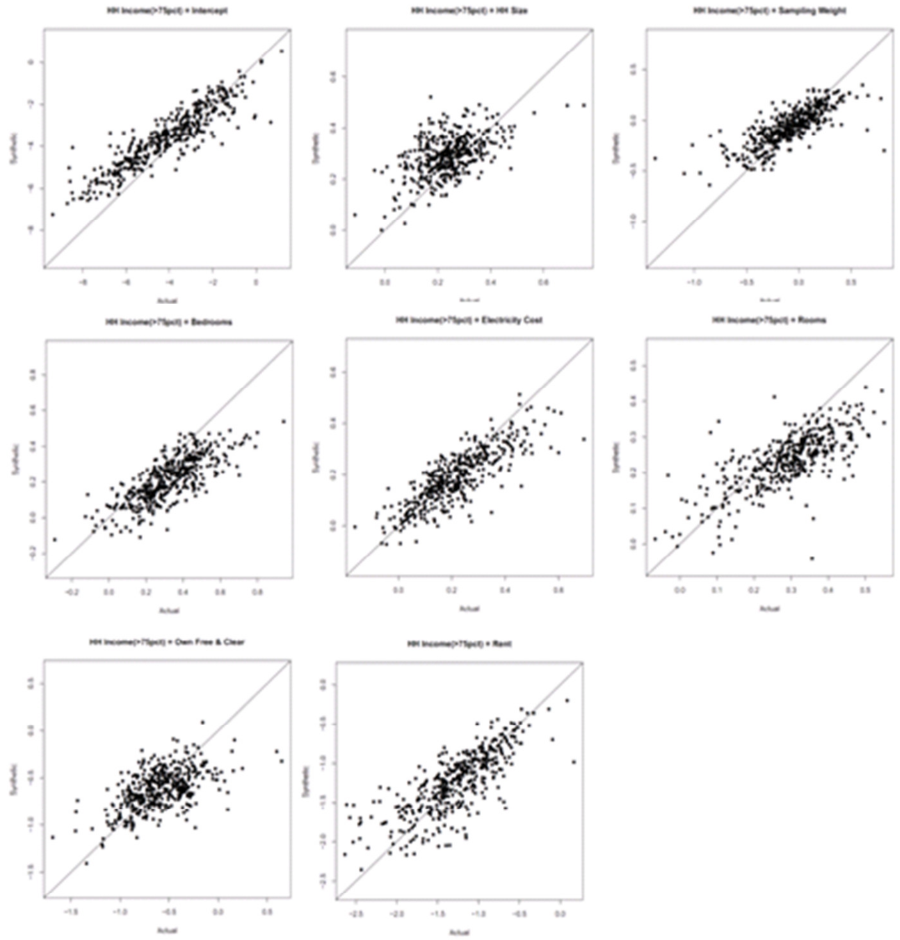


Figure A3.6 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Household Income Greater than the 75th Percentile on Basic Household Characteristics.

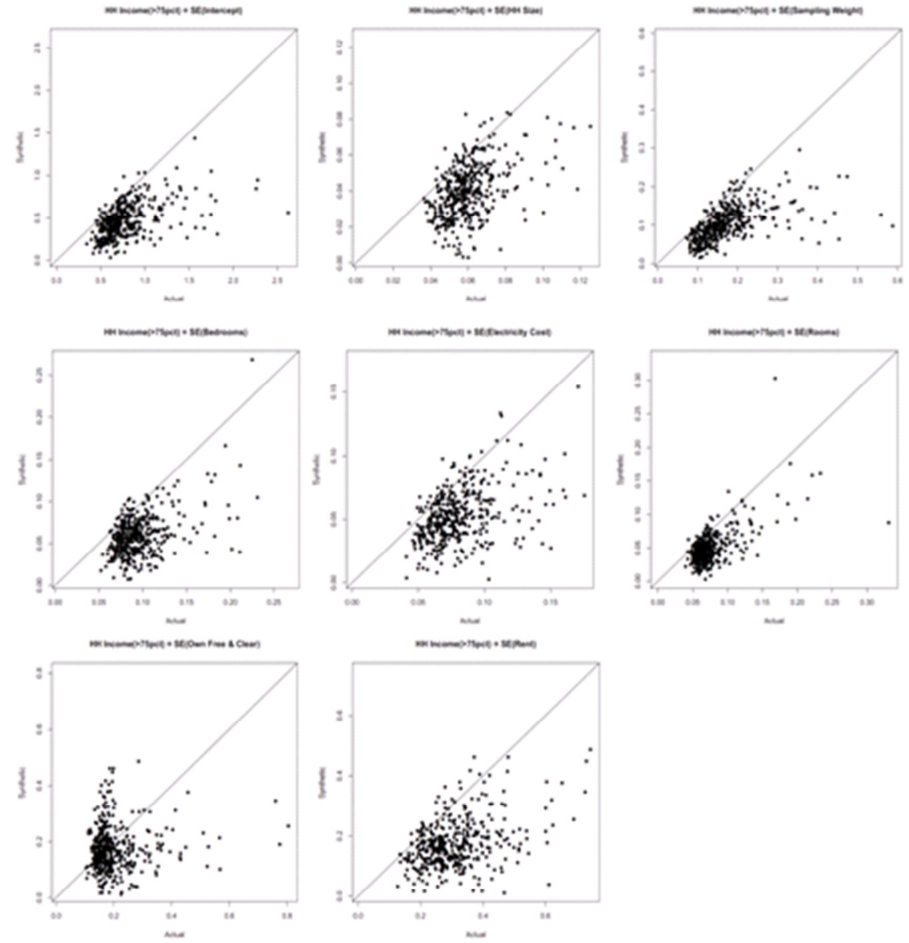


Figure A3.7 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Poverty Status on Personal Demographics.

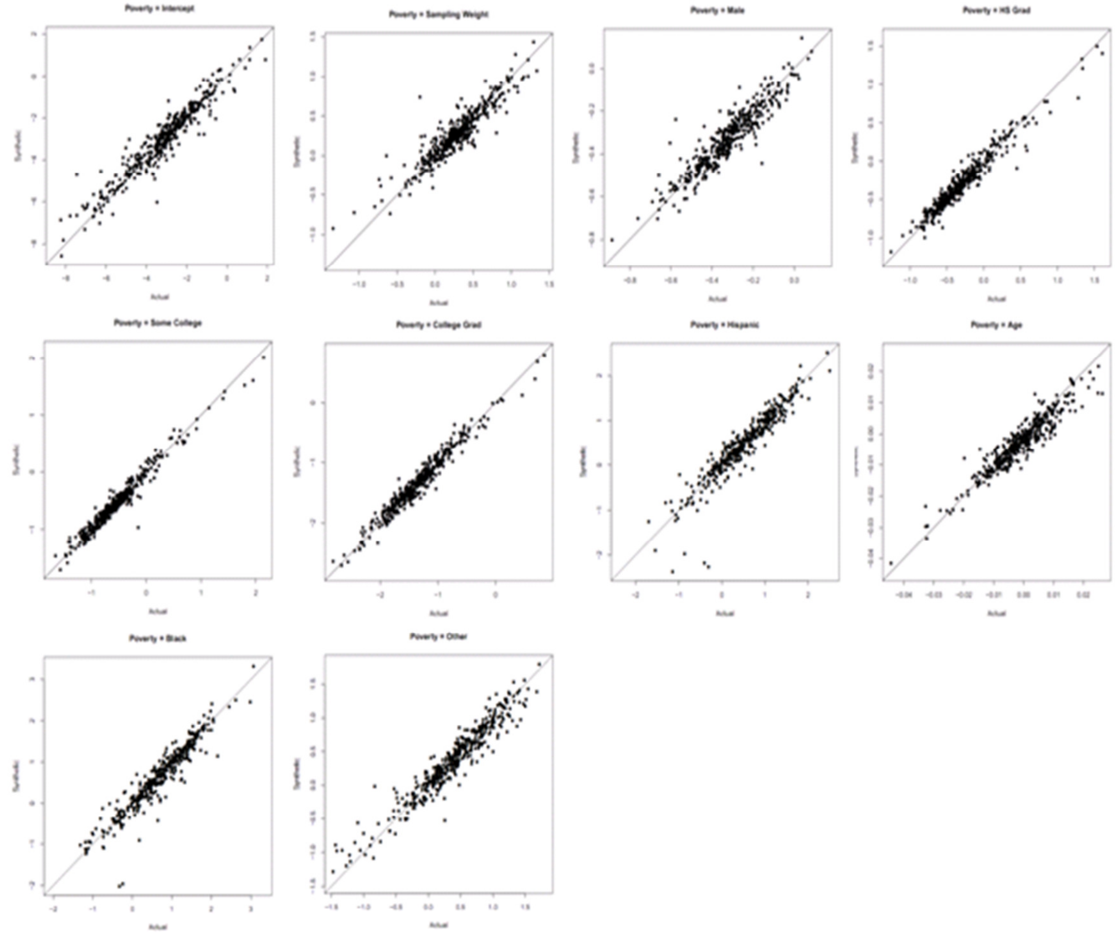


Figure A3.8 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of Poverty Status on Personal Demographics.

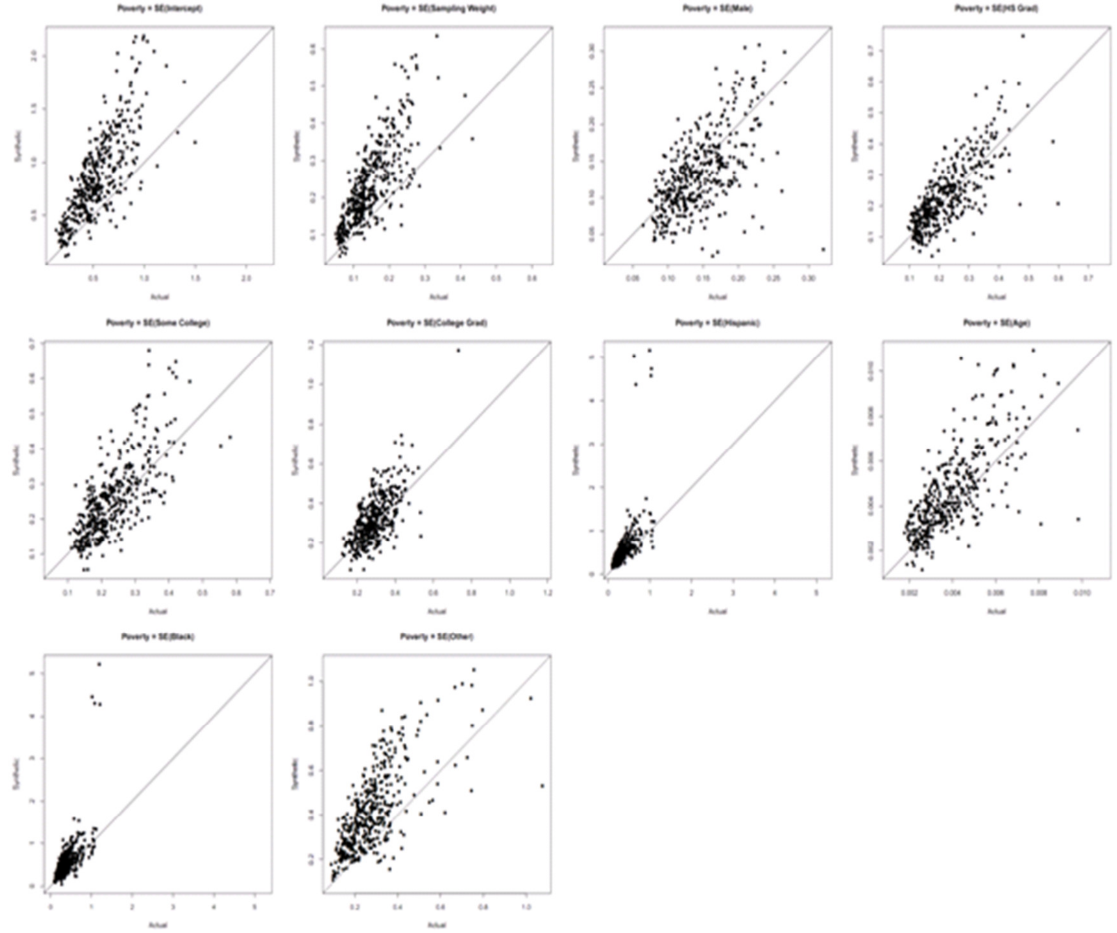


Figure A3.9 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of College Graduation on Personal Demographics.

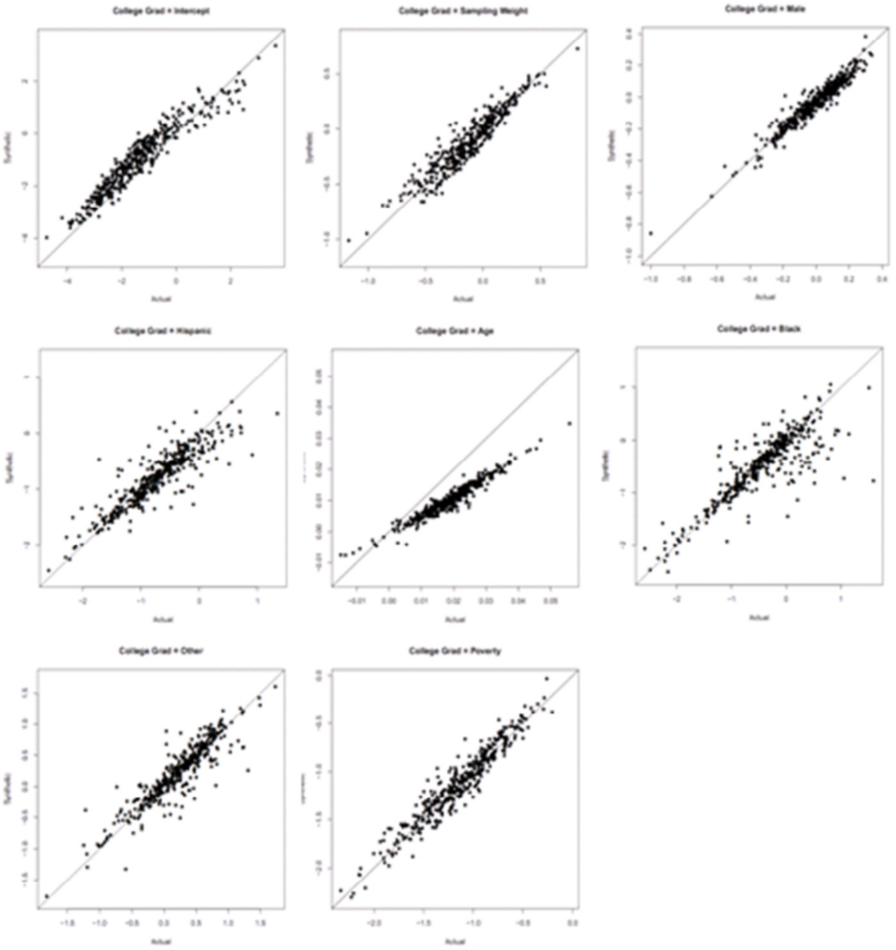


Figure A3.10 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of College Graduation on Personal Demographics.

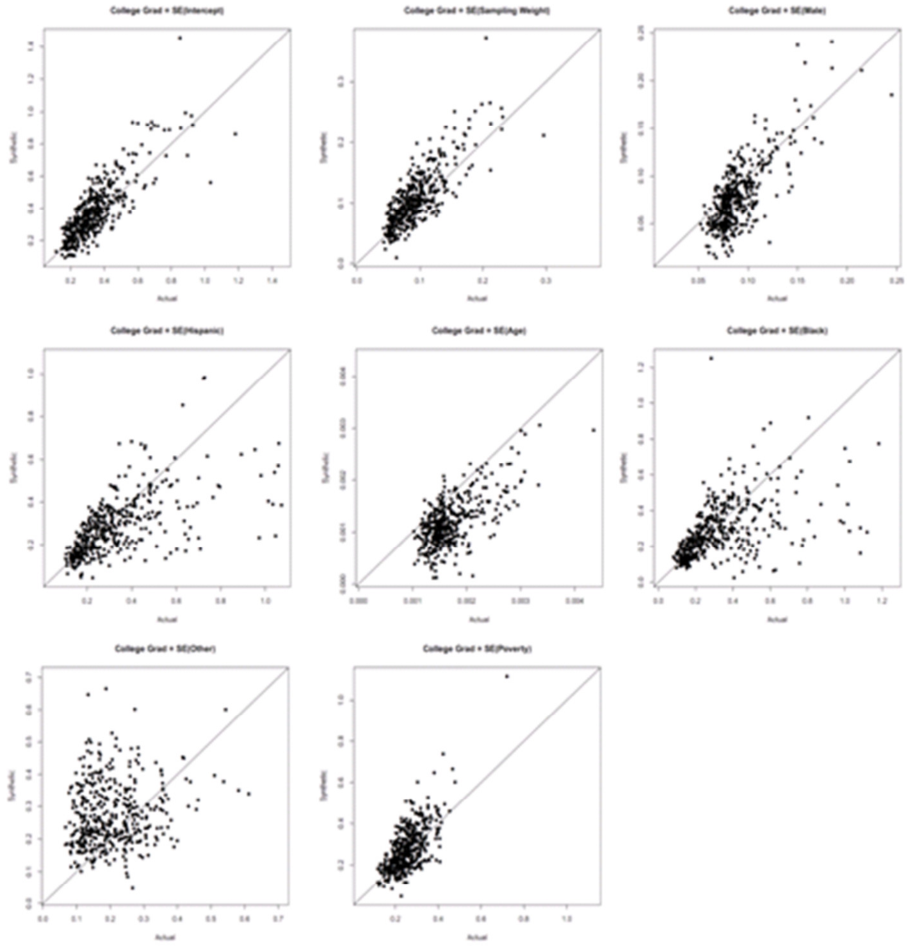


Figure A3.11 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of College Graduation on Personal Demographics and Age Squared Term.

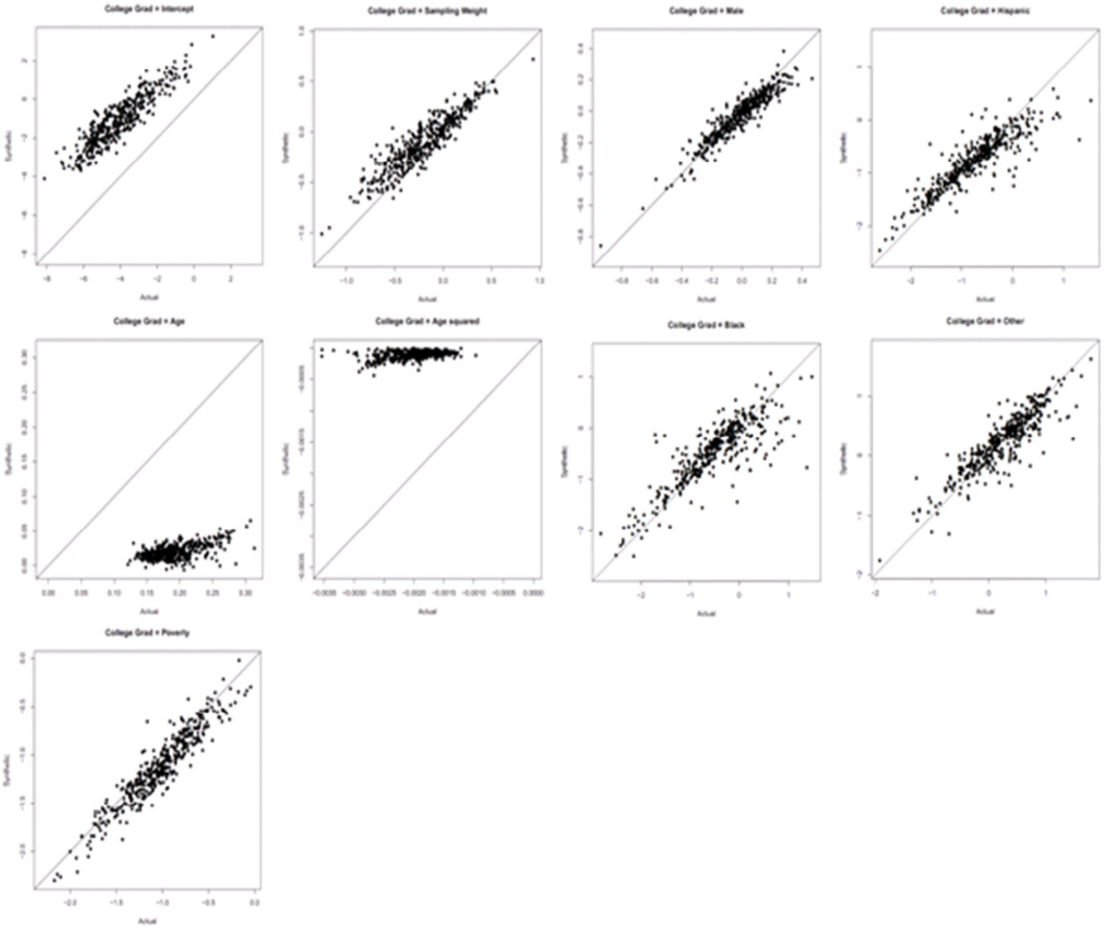
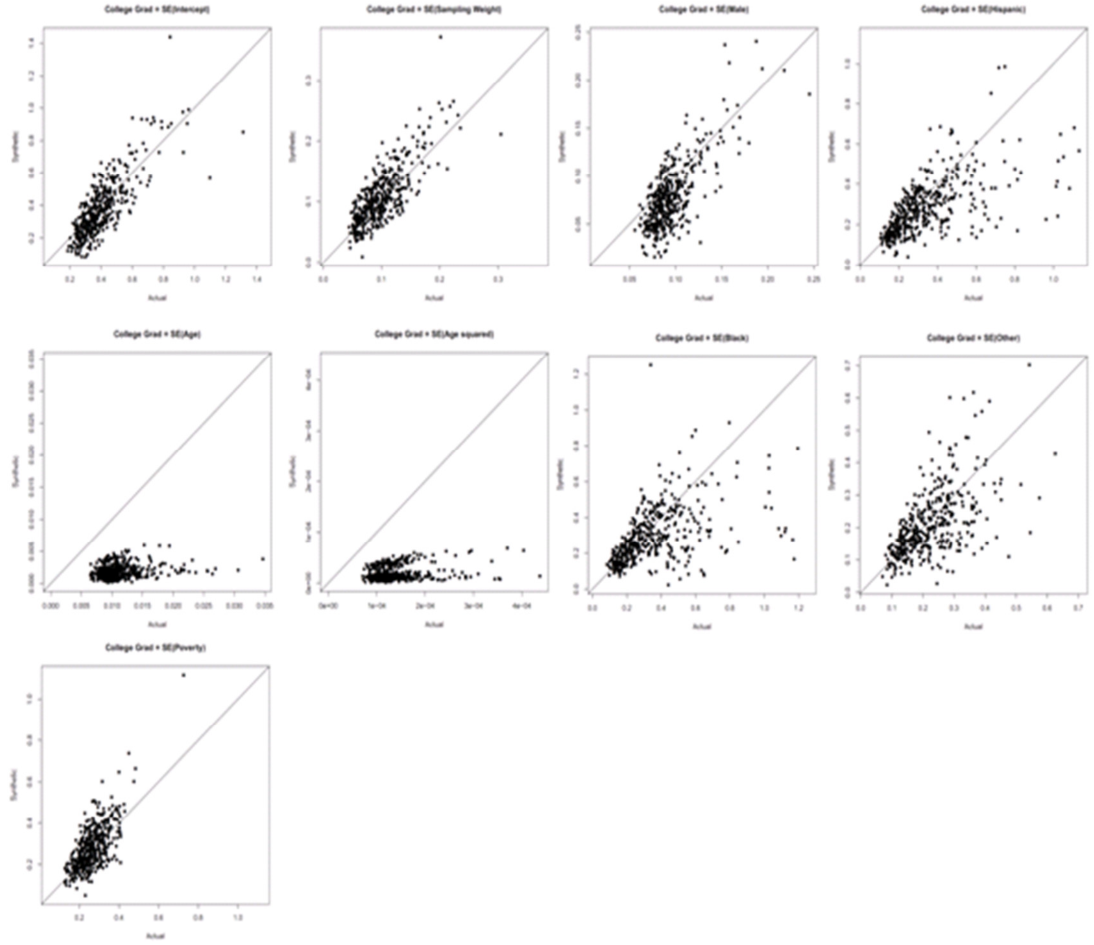


Figure A3.12 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients of College Graduation on Personal Demographics and Age Squared Term.



Appendix 4 Simulation Results

Table A4.1 Conditional Simulation-Based Summary Measures of Actual and Synthetic PUMA 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	2.23	2.23	1.42	1.49	0.07	0.07	0.02	0.99
Sampling weight	33.81	33.50	20.27	17.25	0.99	0.85	0.67	0.99
Total bedrooms	2.79	2.81	0.98	0.99	0.05	0.05	0.22	0.91
Electricity bill/mo.	132.4	134.23	83.70	82.62	3.97	4.02	10.94	0.91
Total rooms	3.14	3.15	1.15	1.16	0.06	0.06	0.16	0.93
Income	82675.8	83847.9	78151.7	68654.3	3631.6	3277.0	2571.0	0.96
Tenure (%)								
Own free & clear	24.78	25.10	42.47	42.58	2.02	2.09	0.02	0.91
Rent	22.74	22.25	39.51	39.60	1.91	2.00	-0.03	1.14
Income > 50th pctile,%	51.26	51.44	47.96	47.76	2.30	2.07	0.02	0.96
Income > 75th pctile,%	25.96	28.47	41.30	42.13	1.96	1.79	0.01	0.89
Income > 90th pctile,%	10.15	12.75	27.59	29.67	1.30	1.22	0.00	0.77
Income (Mortgage=1)	101587.8	103392.6	80397.4	74279.4	5255.6	4726.8	4086.2	0.94
Income (Own=1)	74266.1	71587.9	81705.7	55286.5	8070.6	5407.1	-237.8	1.04
Income (Rent=1)	45652.0	46677.1	42759.2	41501.9	4544.0	4690.3	2313.6	0.93

Table A4.2 Unconditional Simulation-Based Summary Measures of Actual and Synthetic PUMA 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	2.23	2.23	1.42	1.49	0.04	0.08	0.04	0.98
Sampling weight	33.82	33.50	20.35	17.28	0.54	1.13	0.69	0.99
Total bedrooms	2.79	2.81	0.98	0.99	0.03	0.06	0.24	0.91
Electricity bill/mo.	132.45	134.23	83.87	82.69	2.18	5.03	11.34	0.90
Total rooms	3.14	3.15	1.15	1.16	0.03	0.07	0.17	0.94
Income	82696.2	83849.7	78585.2	68746.9	2000.2	4305.6	2738.6	0.95
Tenure (%)								
Own free & clear	24.78	25.10	42.52	42.59	1.11	2.58	0.02	0.91
Rent	22.74	22.25	39.56	39.61	1.05	2.51	-0.03	1.14
Income > 50th pctile,%	51.18	51.50	48.01	47.76	1.26	2.37	0.02	0.96
Income > 75th pctile,%	25.99	28.47	41.37	42.15	1.08	2.15	0.01	0.89
Income > 90th pctile,%	10.16	12.74	27.71	29.70	0.71	1.49	0.00	0.77
Income (Mortgage=1)	101624.5	103421.6	81163.2	74480.3	2906.1	6003.4	4608.3	0.94
Income (Own=1)	74242.0	71593.8	83850.3	55457.5	4527.4	7157.2	979.2	1.02
Income (Rent=1)	45694.9	46681.0	44233.6	41647.6	2561.8	5896.7	2747.8	0.92

Table A4.3 Conditional Simulation-Based Summary Measures of Actual and Synthetic State Means

	Avg. Mean		Avg. Standard Error of Mean	
	Actual	Synthetic	Actual	Synthetic
Household variables				
Household size	2.08	2.08	0.01	0.02
Sampling weight	32.86	32.47	0.16	0.30
Total bedrooms	2.82	2.81	0.01	0.02
Electricity bill/mo.	124.96	124.99	0.50	1.12
Total rooms	3.14	3.14	0.01	0.02
Income	78077.48	78222.90	493.05	998.70
Tenure (%)				
Own free & clear	27.71	28.52	0.30	0.72
Rent	19.64	19.68	0.26	0.62
Income > 50th pctile (%)	47.68	47.43	0.32	0.58
Income > 75th pctile (%)	23.03	25.04	0.26	0.48
Income > 90th pctile (%)	9.06	10.94	0.17	0.31
Income (Mortgage=1)	94908.97	95882.56	716.69	1419.67
Income (Own=1)	69422.16	66704.91	981.42	1611.02
Income (Rent=1)	42773.76	43427.09	583.99	1555.05

Table A4.4 Unconditional Simulation-Based Summary Measures of Actual and Synthetic State Means

	Avg. Mean		Avg. Standard Error of Mean	
	Actual	Synthetic	Actual	Synthetic
Household variables				
Household size	2.11	2.11	0.02	0.02
Sampling weight	34.19	33.84	0.29	0.25
Total bedrooms	2.80	2.80	0.01	0.01
Electricity bill/mo.	124.32	124.52	0.90	0.90
Total rooms	3.13	3.13	0.01	0.02
Income	77999.08	78228.85	899.06	788.85
Tenure (%)				
Own free & clear	26.83	27.61	0.54	0.56
Rent	20.57	20.49	0.47	0.50
Income > 50th pctile (%)	47.94	47.72	0.58	0.52
Income > 75th pctile (%)	23.05	25.12	0.47	0.43
Income > 90th pctile (%)	8.98	10.85	0.30	0.28
Income (Mortgage=1)	95182.15	96080.19	1306.66	1126.27
Income (Own=1)	69556.97	66873.80	1788.97	1274.80
Income (Rent=1)	43088.50	43822.45	1061.24	1259.94

Table A4.5 Conditional and Unconditional Simulation-Based Summary Measures of Actual and Synthetic PUMA Regression Coefficients

	Conditional		Unconditional	
Covariates	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)	Avg. Actual $\hat{\beta}_{cs}$ (SE)	Avg. Synthetic $\hat{\beta}_{cs}$ (SE)
Regression of income (cube root) on				
Intercept	26.80 (4.63)	26.73 (4.75)	26.83 (2.55)	26.74 (4.54)
Household size	1.56 (0.43)	1.66 (0.43)	1.56 (0.24)	1.65 (0.41)
Sampling weight	-0.39 (1.02)	-0.41 (1.07)	-0.39 (0.56)	-0.41 (1.03)
Total bedrooms	1.30 (0.63)	1.23 (0.62)	1.30 (0.35)	1.24 (0.60)
Electricity bill/mo.	1.05 (0.53)	1.05 (0.52)	1.05 (0.30)	1.05 (0.50)
Total rooms	1.31 (0.50)	1.30 (0.51)	1.31 (0.28)	1.30 (0.49)
Tenure				
Mortgage/loan	Ref	Ref	Ref	Ref
Own free & clear	-3.92 (1.44)	-3.28 (1.32)	-3.92 (0.79)	-3.28 (1.24)
Rent	-5.79 (1.48)	-6.18 (1.48)	-5.81 (0.82)	-6.16 (1.40)

Table A4.6 Conditional and Unconditional Simulation-Based Summary Measures of Actual and Synthetic State Regression Coefficients

	Conditional		Unconditional	
Covariates	Avg. Actual $\hat{\beta}_{s}$ (SE)	Avg. Synthetic $\hat{\beta}_{s}$ (SE)	Avg. Actual $\hat{\beta}_{s}$ (SE)	Avg. Synthetic $\hat{\beta}_{s}$ (SE)
Regression of income (cube root) on				
Intercept	25.23 (0.92)	25.35 (0.97)	25.27 (0.51)	25.34 (0.92)
Household size	1.58 (0.10)	1.68 (0.12)	1.58 (0.06)	1.68 (0.07)
Sampling weight	-0.24 (0.18)	-0.26 (0.21)	-0.25 (0.10)	-0.26 (0.20)
Total bedrooms	1.29 (0.14)	1.21 (0.13)	1.28 (0.08)	1.21 (0.13)
Electricity bill/mo.	1.08 (0.13)	1.08 (0.12)	1.08 (0.07)	1.08 (0.12)
Total rooms	1.42 (0.10)	1.39 (0.10)	1.43 (0.06)	1.39 (0.10)
Tenure				
Mortgage/loan	Ref	Ref	Ref	Ref
Own free & clear	-3.61 (0.28)	-3.04 (0.26)	-3.61 (0.15)	-3.04 (0.25)
Rent	-5.57 (0.30)	-6.08 (0.32)	-5.58 (0.17)	-6.07 (0.31)

Chapter 3

Synthetic Data for Continuous Non-Normal Distributions: A Nonparametric Simulation Approach for Small Area Estimation

1 Introduction

One of the primary functions of a statistical agency is to collect high quality survey data and make these data widely available to data users in the public domain. Scientific surveys serve as the principal data sources for many academic researchers, analysts, and policy-makers who use these data to test theories of human behavior and, in turn, inform important policy decisions. The greatest impact of policy decisions and interventions is arguably felt at the local level where people are most likely to be exposed to changes in infrastructure and resource availability. Several studies have shown that neighborhood- and community-level factors are associated with numerous health and behavioral outcomes (Diez Roux, 2001; Mujahid et al., 2008; Auchincloss et al., 2008; Fisher et al., 2004). These findings underscore the need for high quality survey data which is being demanded by researchers interested in studying how small area factors influence the characteristics and well-being of the population.

Many statistical agencies release estimates for various levels of geography. For example, the Census Bureau releases summary tables containing estimates of demographic, social, and economic characteristics of people, households, and housing units for large areas (e.g., national, region, division), small areas (e.g., tracts, block groups), and many intermediate areas (e.g., state, county, census tract) (U.S. Census

Bureau, 2011). The Census Bureau also administers specialized programs for producing updated estimates of income and poverty statistics for school districts, counties, and states (Bell et al., 2007), and health insurance estimates for counties and states (Fisher and Turner, 2004).

The production of these small area estimates can be quite useful for many research and evaluation purposes, but oftentimes these estimates are too limiting for data users who require microdata to perform their own customizable geographical analyses. Such data is needed to test complex hypotheses which require analytic estimates and sophisticated modeling approaches. The Census Bureau and other statistical agencies try to meet this demand by releasing public-use microdata files. However, the usefulness of these public-use files for geographically-based analyses is limited, because geographic identifiers are suppressed for areas with fewer than 100,000 residents. Disclosure concerns prohibit the release of small area identifiers for areas that do not meet this pre-specified threshold. To overcome this limitation, data users may access the suppressed geographic identifiers in a Research Data Center (RDC). However, working in an RDC is not always ideal for prospective data users for several reasons. First, prospective users are usually required to submit a research proposal that is subject to approval by the agency responsible for the collection and storage of the restricted data. This requirement may be too burdensome for users whose analytic objectives are exploratory in nature and whose research questions are not yet well-defined. Second, there is a significant cost burden associated with using the RDC. Many federal RDCs charge a usage fee upward of $20,000 per year, which can be difficult to cover for data users who lack external funds.

Finally, there is no guarantee that small area outputs generated from the RDC will pass disclosure review and be permitted for publication.

1.1 Multiple Imputation for Disclosure Avoidance

To facilitate access to public-use microdata for small geographic areas while maintaining confidentiality protections, we propose the dissemination of synthetic data. As originally described by Rubin (1993), synthetic data consists of multiply-imputed data values that overwrite the observed data values. The synthetic values are drawn from a posterior predictive distribution based on the observed data, similar to how multiply-imputed values are generated for handling survey nonresponse (Rubin, 1987). In the general synthetic data framework, we treat the unobserved portion of the population as missing data to be multiply-imputed using values generated from a predictive model fitted using the observed data. Random samples of arbitrary size are then drawn from the synthetic populations and are released as public-use microdata files. Valid inferences are obtained by analyzing each synthetic data set separately and combining the point estimates and standard errors using standard combining rules developed by Raghunathan, Reiter, and Rubin (2003). Several statistical agencies have experimented with releasing synthetic data files in practical survey applications. (Abowd, Stinson, and Benedetto, 2006; Rodriguez, 2007; Kinney and Reiter, 2008), but no study has considered the synthetic data approach for the purpose of disseminating public-use microdata for small geographic areas.

1.2 The Inferential Validity and Utility of Synthetic Data

A key requirement for obtaining high analytic validity from the synthetic data is that the imputation model is correctly specified and reflects all of the key relationships and variables that are of interest to data users. That is, the synthetic data reflect only those relationships included in the data generation models. When the imputer’s model corresponds to the analyst’s model, then the models are said to be “congenial” in the context of multiple imputation for survey nonresponse (Meng, 1994). The lack of correspondence (or congeniality) between the two models can lead to biased synthetic data inferences. This is an important point of contention among data users, who may be interested in analyzing complex relationships, interactions, and higher-order terms in the synthetic data that are usually unbeknownst to the data imputer prior to synthesis (Reiter, 2009). This issue has raised skepticism among the data user community who fear that the synthetic data will not yield valid inferences.

Ideally, the data imputer will know in advance the types of relationships and estimands that are of potential interest to data users, and will incorporate those features into the data generation process to protect against bias. However, knowing exactly how the synthetic data will be used is not always possible, and the imputer must guess as to which relationships to include in the model. One approach to protecting against bias is to incorporate as many variables, interactions, and higher-order terms as possible into the synthetic data generation model. However, incorporating all-possible analytic features into the model may not be practically feasible in all cases and compromises may be needed. Such compromises should be chosen to maximize analytic validity for the majority of data usages, while simultaneously ensuring a high level of validity for complex analytic objectives that are of interest to a small percentage of data users.

A second approach to protecting against bias is to relax the distributional assumptions associated with parametric imputation models in order to improve model fit and protect against model misspecification. This approach has led to several innovations in the use of semi-parametric and non-parametric imputation models for the purpose of generating synthetic data. Raghunathan, Reiter, and Rubin (2003) evaluated a multivariate normal and a nonparametric Bayesian bootstrap procedure to generate synthetic data sets based on the 1994 Consumer Expenditure Survey. In simulations, the authors found that the sampling properties of inferences from synthetic data sets and the actual data sets were very similar for both the parametric and nonparametric synthetic data generation methods. The authors note, however, that the parametric approach should protect confidentiality more effectively because the values are drawn from a smooth distribution and do not contain any fully observed records, unlike the Bayesian bootstrap which samples from observed records. Reiter (2005) presented a nonparametric imputation method based on classification and regression tree (CART) models to generate synthetic data. In most cases, the repeated sampling properties of the synthetic data mimicked those of the corresponding actual data for both descriptive and analytic estimands. However, the author warns that CART models may not be suitable when trees are built from only a small number of units, in which case they may fail to split on certain variable categories. There is also the concern that nonparametric synthesizers may replicate the data too well and fail to provide sufficient protection for cases with a particularly high of disclosure. In the context of CART, imputers can prune branches from the trees or otherwise coarsen the imputations for these cases. Caiola and Reiter (2010) considered imputation models based on random forests (RF), which are collections of CARTs based on random subsamples of the original data where each tree is grown using random samples of predictors. They found the RF synthesizer to be an effective method for preserving descriptive estimands, non-linear relationships, interactions, and subgroup analyses based on three categorical variables. Disclosure risk assessments also indicated sufficient reduction in the risk of re-identification. Woodcock and Benedetto (2009) developed a new imputation strategy based on kernel density estimation for variables with very skewed and multimodal distributions, which they found to deliver better data utility and lower disclosure risk compared to alternative nonparametric methods.

The synthetic data methods discussed in the above literature review focus on preserving statistics about the entire sample. The development of nonparametric methods for generating synthetic data for small domains and small geographic areas is an underdeveloped area, but one that shows great promise. If created with special care, the dissemination of synthetic data sets for small geographic areas may offer an appealing alternative to working in research data centers and may even generate broader interest and utilization of survey data sets in schools and local organizations.

1.3 Organization of Chapter

In this chapter, I propose a Bayesian hierarchical model for the purpose of creating fully-synthetic continuous variables for small geographic areas. A hierarchical version of the sequential multivariate regression procedure (Raghunathan et al, 2001) is implemented that accounts for multiple levels of geography and borrows strength across related areas. We introduce a nonparametric component of the procedure that is implemented at the final stage of the data generation process when the synthetic continuous values are drawn. The random effect terms are modeled parametrically. The analytic validity of the method is evaluated using public-use and restricted microdata from the American Community Survey (ACS) for years 2005-2007 and 2005-2009, respectively. We focus the evaluation on a selection of skewed and bimodal variables obtained from the ACS. Fully-synthetic data inferences are compared against the actual data inferences for both descriptive and analytic statistics. The disclosure risk properties of the synthetic data are not addressed in this chapter.

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, Reiter, and Rubin (2003) and Reiter (2004). 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, I assume there are no item missing data in the observed survey data set, but methods exist for handling this situation (Reiter, 2004).

Fully synthetic data sets are constructed in two steps. First, M synthetic populations $P^{(l)} = \{(X, Y^{(l)}); l = 1,2, \ldots, M\}$ 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, 10, or 20) 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 $l = 1,2, \ldots, M$ synthetic data populations, $D^{(l)} = \left(x_i, y_i^{(l)}, i = 1,2, \ldots, n_{syn}\right)$. The corresponding M synthetic samples $D_{syn} = \left(D^{(l)}; l = 1,2, \ldots, M\right)$ 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, Reiter, and Rubin (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, Reiter, and Rubin (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 the units that comprise the synthetic samples from each multiply-imputed synthetic population. Because of this additional sampling step, the between imputation variance already reflects the within imputation variability, which is not the case in the usual multiple imputation framework.

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) and Reiter (2005).

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 M or n_{syn}. Numerical routines can be used to calculate the integrals involved in the construction of T_M, yielding more precise variance estimates (Raghunathan, Reiter, and Rubin, 2003). A simpler variance approximation that is always positive is shown in Reiter (2002).

3 Extension to Small Geographic Areas

In this section, we first introduce a fully-parametric synthetic data generation procedure for continuous variables that is based on a hierarchical Bayesian model to generate synthetic data for small geographic areas. The procedure involves three stages. In the first stage, the joint density of the variables under consideration is approximated by fitting a series of sequential linear 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 stage 1 is approximated and the between-area variation is modeled using auxiliary information for larger geographic areas. In the final stage, the unknown regression parameters are simulated from the posterior distribution and used to draw synthetic values from the posterior predictive distribution. We then introduce a modification of the procedure to allow for nonparametric simulation of the synthetic values. The modification is designed to handle skewed and bimodal continuous variable distributions. To simply explanation of the method, I define “small areas” to be counties nested with states.

3.1 Parametric Approach

3.1.1 Stage 1: Approximation to the Joint Density via Sequential Regression

For descriptive purposes, I introduce the following notation. I define “small areas” as counties, nested within states, which could also be nested within even larger areas (e.g., regions). In specific terms, 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 $c = (1,2,\dots,C_s)$ nested within state $s = (1,2,\dots,S)$. Let $Y_{cs} = \left(Y_{ics,p}; i = 1,2,\dots,n_{cs};\ p = 1,2,\dots,P\right)$ represent the $n_{cs} \times P$ matrix of continuous survey variables collected from each survey respondent located in county c and state s. Let $X_{cs} = \left(X_{ics,j}; i = 1,2,\dots,n_{cs}, n_{cs}+1,\dots,N_{cs};\ j = 1,2,..,J\right)$ represent the $N_{cs} \times J$ matrix of auxiliary or administrative variables known for every population member in a particular county and state. Although I 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 among the observed variables are maintained in the synthetic data, i.e., the joint distribution of variables given the auxiliary information $f\left(Y_{cs,1}, Y_{cs,2}, \ldots, Y_{cs,P} | X_{cs,j}\right)$ 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 $f\left(Y_{cs,1}, Y_{cs,2}, \ldots, Y_{cs,P} | X_{cs,j}\right)$ can be factored into the following conditional densities: $f\left(Y_{cs,1} | X_{cs,j}\right)$, $f\left(Y_{cs,2} | Y_{cs,1}, X_{cs,j}\right)$,…,$f\left(Y_{cs,P} | Y_{cs,1}, \ldots, Y_{cs,P-1}, X_{cs,j}\right)$. 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., $Y_{ics,1} = (X_{ics})\beta_{cs,1} + \varepsilon_{ics}$, $Y_{ics,2} = \left(X_{ics}, Y_{ics,1}\right)\beta_{cs,2} + \varepsilon_{ics}$,…, $Y_{ics,P} = \left(X_{ics}, Y_{ics,1}, Y_{ics,2}, \ldots, Y_{ics,P-1}\ \right)\beta_{cs,P} + \varepsilon_{ics}$. The choice of model (e.g., Gaussian, binomial) is dependent on the type of variable to be synthesized, but we only consider continuous variables and corresponding linear regression models. It is assumed that any complex survey design features are incorporated into the generalized linear models. 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. $p = (1,2, \ldots, P)$ 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.

3.1.2 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 $Z_s = \left(Z_{s,k}; k = 1,2, \ldots, K\right)$ 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 β_p and Σ_p, but for computational simplicity I assume that β_p and Σ_p are fixed at their respective maximum likelihood estimates (MLE), 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,

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

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

3.1.3 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 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, under the assumption of normality, the first variable to be synthesized $Y_{cs,1}$ 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 $f\left(\tilde{\sigma}_{cs,1}^2|Y_{cs,1},X_{cs},\sigma_{cs,1}^2\ \right)$, or fixed at the maximum likelihood estimate $\hat{\sigma}_{cs,1}^2$ (obtainable from Stage 1). Once the first synthetic variable $\tilde{Y}_{cs,1}$ is generated, a second (normally distributed) synthetic variable $\tilde{Y}_{cs,2}$ can be drawn from the posterior predictive distribution $f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1},X_{cs},\tilde{\beta}_{cs,2}\right)$, which 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 $\tilde{Y}_{cs,P}{\sim}N\left[\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,P-1}\right)\tilde{\beta}_{cs,P},\sigma_{cs,P}^2\ \right]$. The iterative process continues until all synthetic variables $\left(\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\dots,\tilde{Y}_{cs,P}\right)$ are generated. The procedure is repeated *M* times to create multiple populations of synthetic variables $\left(\tilde{Y}_{cs,1}^{(l)},\tilde{Y}_{cs,2}^{(l)},\dots,\tilde{Y}_{cs,P}^{(l)}; l = 1,2,\dots,M\right)$.

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 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 small areas. Inferences for a variety of estimands can be obtained using the combining rules in Section 2.2.

3.2 Nonparametric Approach

We now consider a modified approach to the parametric framework described in 3.1 that does not require the synthetic values to be drawn from a univariate normal distribution. The final stage in the parametric approach (stage 3) described in 3.1.3 is replaced with a distribution-free simulation procedure, while the first two stages (Sections 3.1.1 and 3.1.2) remain the same. The method still relies on multivariate normality to model the random effects and to obtain the posterior distribution of $\tilde{\beta}_{cs,p}$ in equation (5).

3.2.1 Method

Recall from 3.1.3 the fully-parametric iterative simulation procedure proceeds as follows. The first continuous and normally distributed observed variable $Y_{cs,1} = (Y_{ics,1}; i = 1,2, \dots, n_{cs})$ was simulated from a normal distribution with location and scale parameters $X_{cs}\tilde{\beta}_{cs,1}$ and $\sigma^2_{cs,1}$, respectively, i.e.,

$$\tilde{Y}_{cs,1} \sim N\left[X_{cs}\tilde{\beta}_{cs,1}, \sigma^2_{cs,1}\ \right],$$

where X_{cs} is an $N_{cs} \times J$ matrix of auxiliary or administrative variables known for every population member in a particular county and state. The second observed variable to be synthesized $Y_{cs,2}$, is simulated by drawing from a normal distribution with location and scale parameters$\left(X_{cs}, \tilde{Y}_{cs,1}\right)\tilde{\beta}_{cs,2}$ and $\sigma_{cs,2}^{2}$, respectively, i.e.,

$$\tilde{Y}_{cs,2} \sim N\left[\left(X_{cs}, \tilde{Y}_{cs,1}\right)\tilde{\beta}_{cs,2}, \sigma_{cs,2}^{2}\ \right]$$

where the location parameter $\left(X_{cs}, \tilde{Y}_{cs,1}\right)\tilde{\beta}_{cs,2}$ conditions on the previously synthesized variable $\tilde{Y}_{cs,1}$. The iterative procedure continues until the final variable $Y_{cs,P}$ is synthesized,

$$\tilde{Y}_{cs,P} \sim N\left[\left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,P-1}\right)\tilde{\beta}_{cs,P}, \sigma_{cs,P}^{2}\ \right].$$

The general form of the simulation procedure for the $p^{th}(p = 1,2, \dots, P)$ synthetic variable can therefore be written as,

$$\tilde{Y}_{cs,p} \sim N\left[\left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,p}, \sigma_{cs,p}^{2}\ \right]. \tag{6}$$

The nonparametric simulation procedure that we now describe removes the assumption of univariate normality. The general procedural steps for synthesizing the p^{th} variable are implemented as follows. First, we use the location parameter from (6) to obtain predicted values based on the vector of simulated beta coefficients $\tilde{\beta}_{cs,p}$, any previously synthesized variables $\left(\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)$, and any auxiliary information X_{cs} that is known for each population member in county c nested within state s. Specifically, we refer to these synthetically-based predicted values as those obtained from the following equation,

$$\hat{Y}_{cs,p,syn} = \left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,p} \tag{7}$$

which is computed for population unit $l = (1,2, \dots, N_{cs})$ located in the small area (or county) of interest.

Second, we modify (7) to obtain another set of predicted values that are based on the set of observed variables Y_{cs} instead of the synthetically-generated ones $\tilde{Y}_{cs}$,

$$\hat{Y}_{cs,p,obs} = \left(X_{cs}, Y_{cs,1}, Y_{cs,2}, \dots, Y_{cs,p-1}\right)\tilde{\beta}_{cs,p} \tag{8}$$

In the third step, the differences between the observed survey values $Y_{cs,p}$ and the observed predicted values $Y_{cs,p}$ are obtained to create a $n_{cs} \times 1$ vector of deviations,

$$\Delta_{cs,p} = Y_{cs,p} - \hat{Y}_{cs,p,obs} \tag{9}$$

In the fourth step, we account for the uncertainty associated with the distribution of deviated values by resampling the vector $\Delta_{cs,p}$ using an approximate Bayesian Bootstrap (ABB) procedure (Rubin and Schenker, 1996), which is a more computationally direct procedure than the original Bayesian Bootstrap (Rubin, 1981). The ABB procedure is implemented by drawing the components of an n_{cs}-dimensional vector $\Delta_{cs,p,SRSWR}$ from $\Delta_{cs,p}$ with replacement, i.e., $\Delta_{cs,p,SRSWR} = SRSWR(\Delta_{cs,p})$. The final part of the ABB procedure is to draw the components of a N_{cs}-dimensional vector $\Delta_{cs,p,ABB}$ from $\Delta_{cs,p,SRSWR}$ with replacement, i.e., $\Delta_{cs,p,ABB} = SRSWR(\Delta_{cs,p,SRSWR})$.

The final step of the simulation process is to generate synthetic variables using the components from the previous steps. Specifically, the p^{th}synthetic variable is generated using the following equation,

$$\begin{aligned} \tilde{Y}_{cs,p} &= \left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,p} + \Delta_{cs,p,ABB} \\ &= \tilde{Y}_{cs,p,syn} + \Delta_{cs,p,ABB} \end{aligned} \tag{10}$$

The resulting synthetic data may be analyzed per usual using the combining rules presented in Section 2.2.

A few general remarks can be made about this simulation method. Firstly, generating the synthetic values does not rely on any standard distributions as it relaxes the assumption of univariate normality. However, the method still relies on multivariate normality in the hierarchical model, which may not be an adequate assumption if the random effects follow a non-normal distribution. Secondly, due to the nonparametric nature of the method, and the fact that the synthetic values are based on deviations from the actual values, means there is no need to apply a transformation to the variables, prior to the synthesis, in order to achieve normality. This is a useful property of the method as choosing a suitable transformation can be a difficult task, particularly when the appropriate transformation may vary across geographic areas in spatial applications. The effectiveness of the method for synthesizing non-transformed variables will be assessed in the next section. Lastly, the method can be easily implemented in a variety of hierarchical synthetic data applications involving continuous variables. It can also be applied in conjunction with parametric simulation models (e.g., binomial) in applications involving a mix of continuous and non-continuous variables.

4 Application: American Community Survey (Public-Use Microdata)

The nonparametric simulation method in 3.2.1 is evaluated using a subset of public-use microdata from the 2005-2007 U.S. American Community Survey (ACS). The ACS is an ongoing national survey that provides yearly estimates regarding income and benefits, health insurance, disabilities, family and relationships, among other topics. The ACS collects information on persons living in housing units and group quarters facilities in all 3,142 counties. Data collection is conducted using a mixed-mode design. First, questionnaires are mailed to all sampled household addresses obtained from the Master Address File. Approximately six weeks after the questionnaire is mailed the Census Bureau will attempt to conduct telephone interviews for all addresses that do not respond by mail. Following the telephone operation, a sample is taken from addresses which were not interviewed and these addresses are visited by a field interviewer. Full details of the ACS methodology can be found elsewhere (Census Bureau, 2009).

The smallest geographic unit that is identified in the public-use ACS microdata is a Public-Use Microdata Area (PUMA). PUMAs are census areas that contain at least 100,000 persons, are nested within states or equivalent entities, cover the entirety of the United States, Puerto Rico, Guam, and the U.S. Virgin Island, are built on counties and census tracts, and are contiguous. For this application, the ACS sample is restricted to the Northeast region, which contains 9 states and 405 PUMAs. ACS data was collected in each of these PUMAs during the 3-year study period. The evaluation is conducted on 5 continuous variables (three household- and two person-level variables) measured on 599,450 households and 1,506,011 persons. The variables, shown in Table 3.0, include the household- and person-level sampling weights, electricity cost/month, household income, and age of all household residents. None of these variables follows a normal distribution. The first four variables are right-skewed and the last variable (age) is bimodal. These variables were suggested by statisticians at the U.S. Census Bureau for this project.

$M = 10$ fully synthetic data sets are generated for each "small area" or PUMA. To ensure that each synthetic data set contains ample numbers of households and/or persons within PUMAs, the synthetic sample sizes are created to be larger than the observed sample sizes, and are approximately equivalent to 20% of the total number of households located in each PUMA based on the 2000 decennial census counts. This yielded a total synthetic sample size of 3,963,715 households and 10,192,987 persons in the Northeast region.

Design-based estimates of regression parameters were obtained by fitting normal linear models within each PUMA and synthetic values were drawn from the Gaussian posterior predictive distribution. To ensure the stability of the design-based regression estimates, a minimum PUMA sample size rule of $15 \cdot p$ was applied within each PUMA. If a PUMA did not meet this sample size threshold, then nearby PUMAs were pooled together until the criterion was met.

After the household variables were synthesized, the synthetic household data sets were converted to person-level data sets and the person-level variables were synthesized unconditional to the household-level variables. 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. Finally, 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.

Both the parametric and nonparametric synthetic data generation procedures presented in Sections 3.1.3 and 3.2.1, respectively, are evaluated in this analysis. Both procedures may be applied to variables that have undergone a normalizing transformation or not. We apply both synthetic data procedures to transformed and nontransformed versions of the same variables to evaluate the analytic validity of the method under different transformation scenarios. The log transformation is applied to the household- and person-level sampling weights and a cube root transformation is applied to the electricity cost and household income variables. All of these variables are right-skewed. The approximate bimodal variable age is left untransformed. All transformed variables are back-transformed in the evaluation. That is, all synthetic and observed distributions and estimates shown below are presented in actual units. All estimates are based on unweighted data.

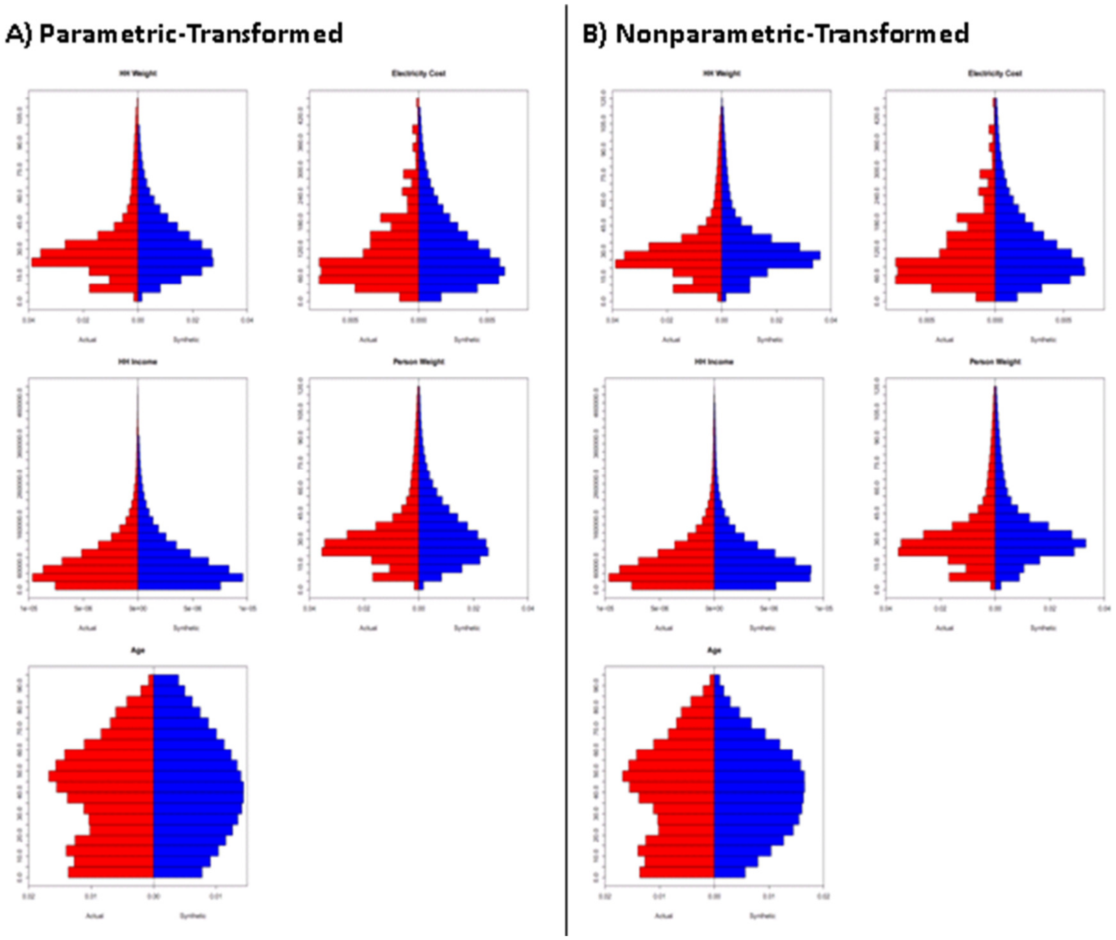
Table 3.0 List of ACS Variables Used in Synthetic Data Application. Variables Shown in the Order of Synthesis.

Variable	**Type**	**Range**	**Shape**
Household variables			
Sampling weight	continuous	1 - 516	right-skewed
Electricity bill/mo.	continuous	1 - 600	right-skewed
Income	continuous	0 – 2,158,100	right-skewed
Person variables			
Sampling weight	continuous	1 - 814	right-skewed
Age	continuous	0 - 95	bimodal

4.1 Validity of Univariate Estimates

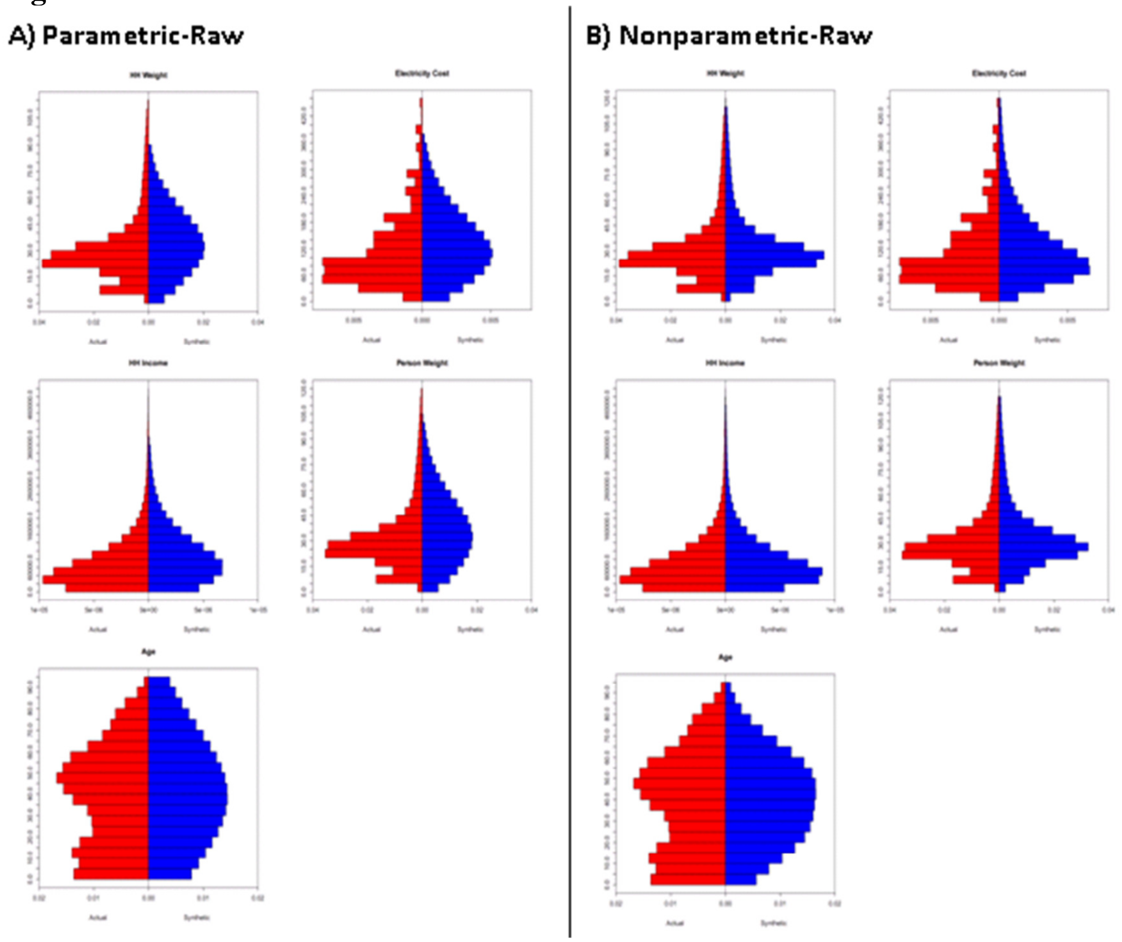
Figures 3.1 and 3.2 show back-to-back histograms of the overall synthetic and actual distributions of the transformed and non-transformed variables, respectively, for each synthetic data method. The actual distributions are shown in red and the synthetic distribution in blue. The parametric and nonparametric results are shown in panels A and B, respectively. All variables are presented on the untransformed scale. The synthetic data generated from both the parametric and nonparametric methods resemble the actual data reasonably well for the right-skewed distributions. Both methods preserve the bulk of the distributions. However, the nonparametric synthetic data tends to reflect the distributions more precisely than the parametrically-generated data. For example, the parametric data tends to smooth over the transition between the distributional mode and skewed portion of the distributions, whereas the shape of the nonparametrically-generated data is more closely aligned with the actual shape and arc of the distribution.

Figure 3.1. Back-to-Back Histograms of Actual (Red) and Synthetic (Blue) Distributions for Transformed ACS Household-Level Variables in the Northeast Region.



However, the bimodal variable distribution, age, is not reflected very well by either method. The lone bimodal variable, age (depicted on the bottom of the figures), is not reflected very well by either synthetic data method, as both methods fail to replicate the upward concavity of the distribution. However, the nonparametric data distribution still seems to reflect other portions of the distribution more precisely than the parametric data. Based on the histograms, it does not seem to matter whether a transformation was used prior to synthesis. We will examine this matter more closely when evaluating the validity of the synthetic data estimates.

Figure 3.2. Back-to-Back Histograms of Actual (Red) and Synthetic (Blue) Distributions for Nontransformed ACS Household-Level Variables in the Northeast Region.



Although the quality of the synthetic variable distributions look relatively promising, data users are most interested in the validity of estimates obtained from the synthetic data. Tables 3.1 and 3.2 contain overall averages of PUMA means (column 2), obtained from 405 PUMAs in the Northeast region, for the transformed and nontransformed variables, respectively. The means are computed for the list of variables in Table 3.0 as well as for three binary variables corresponding to the 50th, 75th, and 90th percentiles of the household income distribution. The average standard deviation and standard error, and intercept and slope of the regression of the actual point estimates on the synthetic point estimates are shown in columns 3-5, respectively. (Intercept values close to 0 and slope values close to 1 indicate strong correspondence between the synthetic and actual means.) The state- and region-level summary measures of means and standard errors are shown in Tables 3.3-3.4 and 3.5-3.6, respectively.

For the eight parametric-transformed estimands shown in upper panel of Table 3.1, six of them yield an average synthetic PUMA mean that lies within one average standard error from the actual average PUMA mean. The two discordant estimands correspond to the proportions of household incomes greater than the 75th and 90th percentiles; both estimates tend to be overestimated in the synthetic data, on average. For the parametric-nontransformed estimands shown in the lower panel of Table 3.1, five out of the eight synthetic estimands lie within one average standard error from the actual average PUMA mean; the discordant estimands consist of all three income proportions. Some of the nonparametric synthetic point estimates tend to be closer to the actual point estimates than do the parametric estimates (e.g., Avg. Household Income; Parametric: 81671 vs. Nonparametric: 81169 vs. Actual: 80588), but this is not always true as indicated by the lack of strong correspondence for all of the nonparametric income proportion estimates.

Another way to assess the analytic validity of the synthetic data is to compare its standard deviations with the actual data. If the validity of the synthetic data is high then the standard deviations obtained from the synthetic data should equal (or approximately equal) to the standard deviations obtained from the actual data. In many cases, the nonparametric synthetic data yield an average standard deviation that is much closer to the actual standard deviation. This is particularly true for the household income variable, which yields average standard deviations of 66250, 76337, and 75075 for the parametric, nonparametric, and actual PUMA estimates, respectively. The same pattern, though, more striking, is observed for the bimodal age variable as the nonparametric average standard deviation is equivalent to the corresponding actual standard deviation (Avg. SD; Parametric: 33.17 vs. Nonparametric: 22.76 vs. Actual: 22.76).

Table 3.1 Summary Measures of Actual and Synthetic PUMA Means for Transformed Variables.

	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
Parametric - Transformed								
Household variables								
Sampling weight	33.71	33.46	20.03	17.71	0.55	0.47	0.15	1.00
Electricity bill/mo.	125.08	126.75	85.39	83.22	2.32	2.27	2.35	0.97
Income	80588.3	81671.4	75075.7	66250.8	2020.9	1811.9	2616.0	0.96
Person variables								
Sampling weight	35.37	35.73	21.53	21.16	0.37	0.62	0.47	0.98
Age	39.44	39.00	22.76	33.17	0.39	0.55	10.90	0.73
Recodes								
Income > 50th pctile,%	50.00	50.62	47.69	47.71	1.29	1.05	-0.01	1.00
Income > 75th pctile,%	25.72	27.60	40.80	41.65	1.10	0.88	-0.00	0.95
Income > 90th pctile,%	10.12	12.13	27.15	28.93	0.73	0.57	0.00	0.82
Nonparametric - Transformed								
Household variables								
Sampling weight	33.71	34.10	20.03	20.21	0.55	0.59	0.13	0.99
Electricity bill/mo.	125.08	124.81	85.39	87.57	2.32	2.43	1.89	0.99
Income	80588.3	81169.5	75075.7	76337.2	2020.9	2112.1	2385.0	0.96
Person variables								
Sampling weight	35.37	35.62	21.53	20.91	0.37	0.66	-0.34	1.00
Age	39.44	38.99	22.76	22.76	0.39	0.50	11.10	0.73
Recodes								
Income > 50th pctile,%	50.00	52.07	47.69	47.46	1.29	1.13	0.00	0.96
Income > 75th pctile,%	25.72	27.09	40.80	40.92	1.10	1.01	0.02	0.89
Income > 90th pctile,%	10.12	11.15	27.15	27.74	0.73	0.71	0.01	0.83

For nontransformed variables (Table 3.2), the superiority of the nonparametric method is more evident. Under the parametric synthetic method, only two of the average

PUMA means lies within one average standard error of the actual average PUMA mean. In contrast, the nonparametric synthetic method yields five estimates which fall within a single standard error of their corresponding actual estimate, on average. Furthermore, the nonparametric approach yields income proportion estimates that are more valid than the corresponding parametric method. For example, the average PUMA proportions of income values greater than the 50th, 75th, and 90th percentiles for the parametric data are 0.57, 0.35, and 0.15, respectively, whereas the corresponding nonparametric proportions are 0.53, 0.27, and 0.11.

By comparing Tables 3.1-3.2, it is evident that the nonparametric method produces small area estimates that are comparable regardless of whether a pre-synthesis transformation is applied to the variables. Hence, the nonparametric method does not require a transformation to obtain basic descriptive estimates from the variables considered here. This is a strength of the method as it avoids the need to select a transformation which can be an imperfect and time consuming task for imputers, especially when a large number of variables are being synthesized.

In summary, these summary measures suggest that the analytic validity of the nonparametric method is high for univariate small area estimates, and in some cases, outperforms the parametric approach for transformed variables. The same pattern is observed for higher-levels of geography, including state- and region-level estimates shown in Tables 3.3-3.4 and 3.5-3.6, respectively.

Table 3.2 Summary Measures of Actual and Synthetic PUMA Means for Non-Transformed Variables.

	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
Parametric - Raw								
Household variables								
Sampling weight	33.71	36.15	20.03	17.94	0.55	0.45	-3.95	1.04
Electricity bill/mo.	125.08	126.58	85.39	85.70	2.32	2.37	1.91	0.97
Income	80588.3	81604.4	75075.7	75403.1	2020.9	2136.0	2723.0	0.95
Person variables								
Sampling weight	35.37	38.51	21.53	20.21	0.37	0.57	-3.33	1.01
Age	39.44	38.77	22.76	33.15	0.39	0.55	10.13	0.76
Recodes								
Income > 50th pctile,%	50.00	57.59	47.69	48.01	1.29	1.07	-0.22	1.25
Income > 75th pctile,%	25.72	35.33	40.80	44.35	1.10	0.98	-0.03	0.81
Income > 90th pctile,%	10.12	15.39	27.15	29.81	0.73	0.59	0.01	0.58
Nonparametric - Raw								
Household variables								
Sampling weight	33.71	33.76	20.03	20.02	0.55	0.55	-0.04	1.00
Electricity bill/mo.	125.1	126.73	85.39	85.65	2.32	2.32	1.72	0.97
Income	80588.3	82102.9	75075.7	75365.9	2020.9	2023.8	2943.0	0.95
Person variables								
Sampling weight	35.37	35.54	21.53	21.15	0.37	0.69	-0.50	1.01
Age	39.44	38.99	22.76	22.70	0.39	0.48	10.95	0.73
Recodes								
Income > 50th pctile,%	50.00	53.13	47.69	47.33	1.29	1.11	0.00	0.94
Income > 75th pctile,%	25.72	27.61	40.80	41.00	1.10	1.02	0.02	0.86
Income > 90th pctile,%	10.12	11.21	27.15	27.58	0.73	0.69	0.01	0.79

Table 3.3 Summary Measures of Actual and Synthetic State Means for Transformed Variables.

	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Parametric - Transformed				
Household variables				
Sampling weight	33.45	33.12	0.14	0.13
Electricity bill/mo.	117.34	118.25	0.44	0.43
Income	78316.75	78921.56	431.89	373.61
Person variables				
Sampling weight	34.70	34.95	0.09	0.16
Age	40.03	39.89	0.09	0.12
Recodes				
Income > 50th pctile (%)	49.77	50.02	0.73	0.75
Income > 75th pctile (%)	24.50	26.41	0.25	0.20
Income > 90th pctile (%)	9.16	10.94	0.16	0.13
Nonparametric - Transformed				
Household variables				
Sampling weight	33.45	33.77	0.14	0.17
Electricity bill/mo.	117.34	116.48	0.44	0.40
Income	78316.75	78242.80	431.89	421.00
Person variables				
Sampling weight	34.70	34.90	0.09	0.18
Age	40.03	39.89	0.09	0.11
Recodes				
Income > 50th pctile (%)	49.77	51.15	0.30	0.22
Income > 75th pctile (%)	24.50	25.64	0.25	0.20
Income > 90th pctile (%)	9.16	10.19	0.16	0.15

Table 3.4 Summary Measures of Actual and Synthetic State Means for Non-Transformed Variables.

	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Parametric - Raw				
Household variables				
Sampling weight	33.45	36.50	0.14	0.12
Electricity bill/mo.	117.34	118.02	0.44	0.48
Income	78316.60	78676.6	431.89	415.61
Person variables				
Sampling weight	34.70	38.45	0.09	0.16
Age	40.03	39.54	0.09	0.12
Recodes				
Income > 50th pctile (%)	49.77	57.43	0.30	0.22
Income > 75th pctile (%)	24.50	34.24	0.25	0.20
Income > 90th pctile (%)	9.16	13.52	0.16	0.12
Nonparametric - Raw				
Household variables				
Sampling weight	33.45	33.43	0.14	0.15
Electricity bill/mo.	117.34	118.08	0.44	0.46
Income	78316.75	79160.17	431.89	374.71
Person variables				
Sampling weight	34.70	34.81	0.09	0.19
Age	40.03	39.83	0.09	0.10
Recodes				
Income > 50th pctile (%)	49.77	52.36	0.30	0.22
Income > 75th pctile (%)	24.50	25.90	0.25	0.22
Income > 90th pctile (%)	9.16	9.97	0.16	0.14

Table 3.5 Summary Measures of Actual and Synthetic Region Means for Transformed Variables.

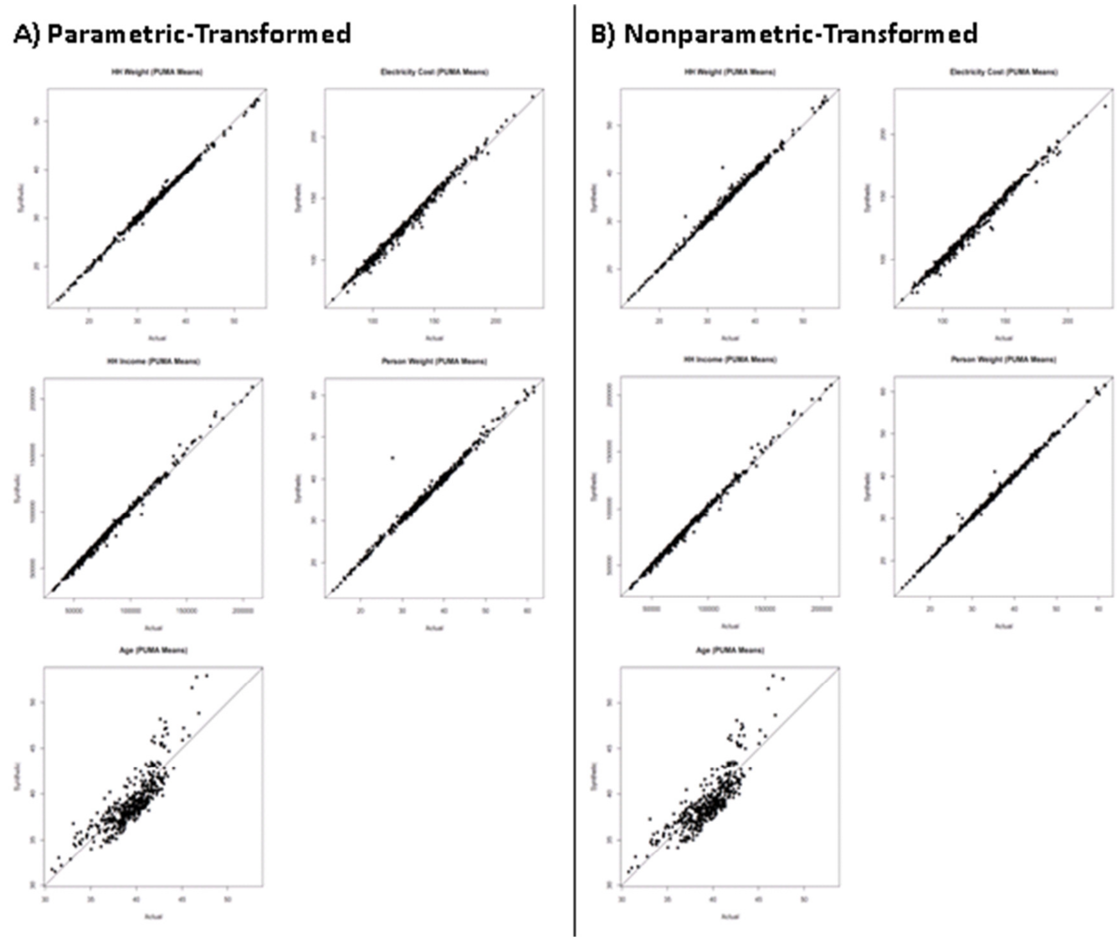
	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Parametric - Transformed				
Household variables				
Sampling weight	32.05	33.48	0.03	0.03
Electricity bill/mo.	124.80	125.58	0.12	0.09
Income	80670.94	81544.00	113.32	92.83
Person variables				
Sampling weight	33.42	35.80	0.02	0.04
Age	39.69	38.99	0.02	0.03
Recodes				
Income > 50th pctile (%)	50.00	50.31	0.07	0.05
Income > 75th pctile (%)	25.47	27.35	0.06	0.05
Income > 90th pctile (%)	10.00	12.05	0.04	0.03
Nonparametric - Transformed				
Household variables				
Sampling weight	32.05	34.12	0.03	0.03
Electricity bill/mo.	124.80	123.66	0.12	0.12
Income	80670.94	81059.38	113.32	85.04
Person variables				
Sampling weight	33.42	35.68	0.02	0.03
Age	39.69	38.98	0.02	0.02
Recodes				
Income > 50th pctile (%)	50.00	51.71	0.07	0.02
Income > 75th pctile (%)	25.47	26.81	0.06	0.03
Income > 90th pctile (%)	10.00	11.10	0.04	0.03

Table 3.6 Summary Measures of Actual and Synthetic Region Means for Non-Transformed Variables.

	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Parametric - Raw				
Household variables				
Sampling weight	32.05	36.24	0.03	0.02
Electricity bill/mo.	124.80	125.45	0.12	0.13
Income	80670.94	81531.97	113.32	132.85
Person variables				
Sampling weight	33.42	38.55	0.02	0.04
Age	39.69	38.76	0.02	0.02
Recodes				
Income > 50th pctile (%)	50.00	57.37	0.07	0.06
Income > 75th pctile (%)	25.47	35.06	0.06	0.05
Income > 90th pctile (%)	10.00	15.27	0.04	0.04
Nonparametric - Raw				
Household variables				
Sampling weight	32.05	33.80	0.03	0.03
Electricity bill/mo.	124.80	125.60	0.12	0.11
Income	80670.94	82038.81	113.32	89.76
Person variables				
Sampling weight	33.42	35.61	0.02	0.04
Age	39.69	38.98	0.02	0.01
Recodes				
Income > 50th pctile (%)	50.00	52.77	0.07	0.05
Income > 75th pctile (%)	25.47	27.33	0.06	0.04
Income > 90th pctile (%)	10.00	11.20	0.04	0.04

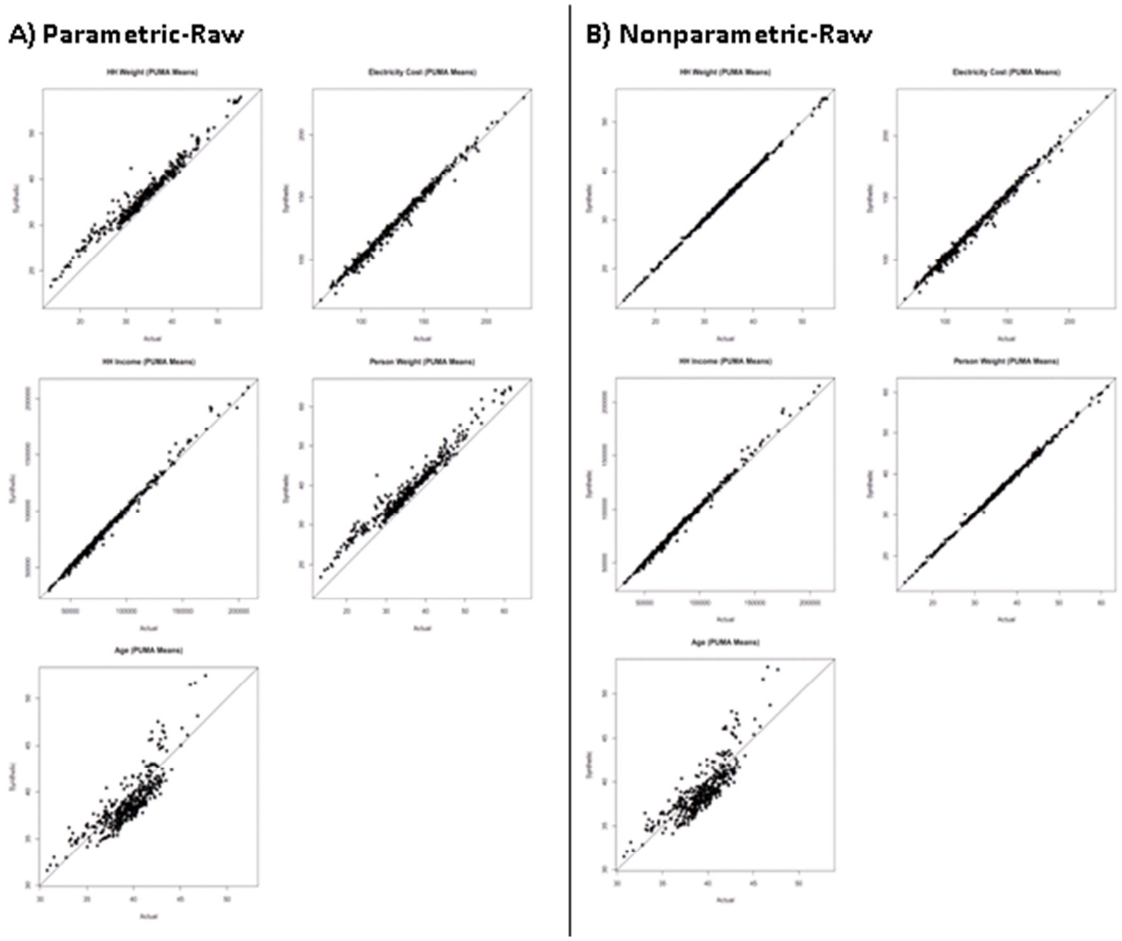
The variability in the synthetic means/percentages of across PUMAs is shown via scatter plots in Figures 3.3 and 3.4 for transformed and nontransformed variables, respectively. Panels A and B correspond to estimates obtained from the parametric and nonparametric synthetic data generation models, respectively. The transformed variable estimates in Figures 3.3a and 3.3b lie closely along the 45 degree line, which suggests strong correspondence between the synthetic and actual PUMA estimates for both the parametric and nonparametric synthetic data generation methods. Mean estimates of age yield the greatest amount of dispersion around the 45-degree line. PUMAs with the highest average ages tend to be overestimated in the synthetic data. This is not surprising due to the bimodal nature of the age distribution which is poorly reflected with both synthetic data methods.

Figure 3.3 Scatter Plot of Synthetic (y-axis) and Actual (x-acis) PUMA Means for Transformed Variables



The scatter plots for estimates obtained from the nontransformed variables (Figures 3.4a and 3.4b) yield larger differences between parametric and nonparametric methods. For example, the parametric plots for the household- and person-level sampling weight variables yield a noticeable amount of dispersion about the 45-degree line as well as overestimation compared to the actual estimates. The same plots in the nonparametric panel show point estimates that are tightly clustered about the 45-degree line with no indication of bias. In general, when the nonparametric approach is applied to the raw variables it produces synthetic point estimates that are just as close (if not closer) to the actual point estimates, than are the parametrically-based point estimates.

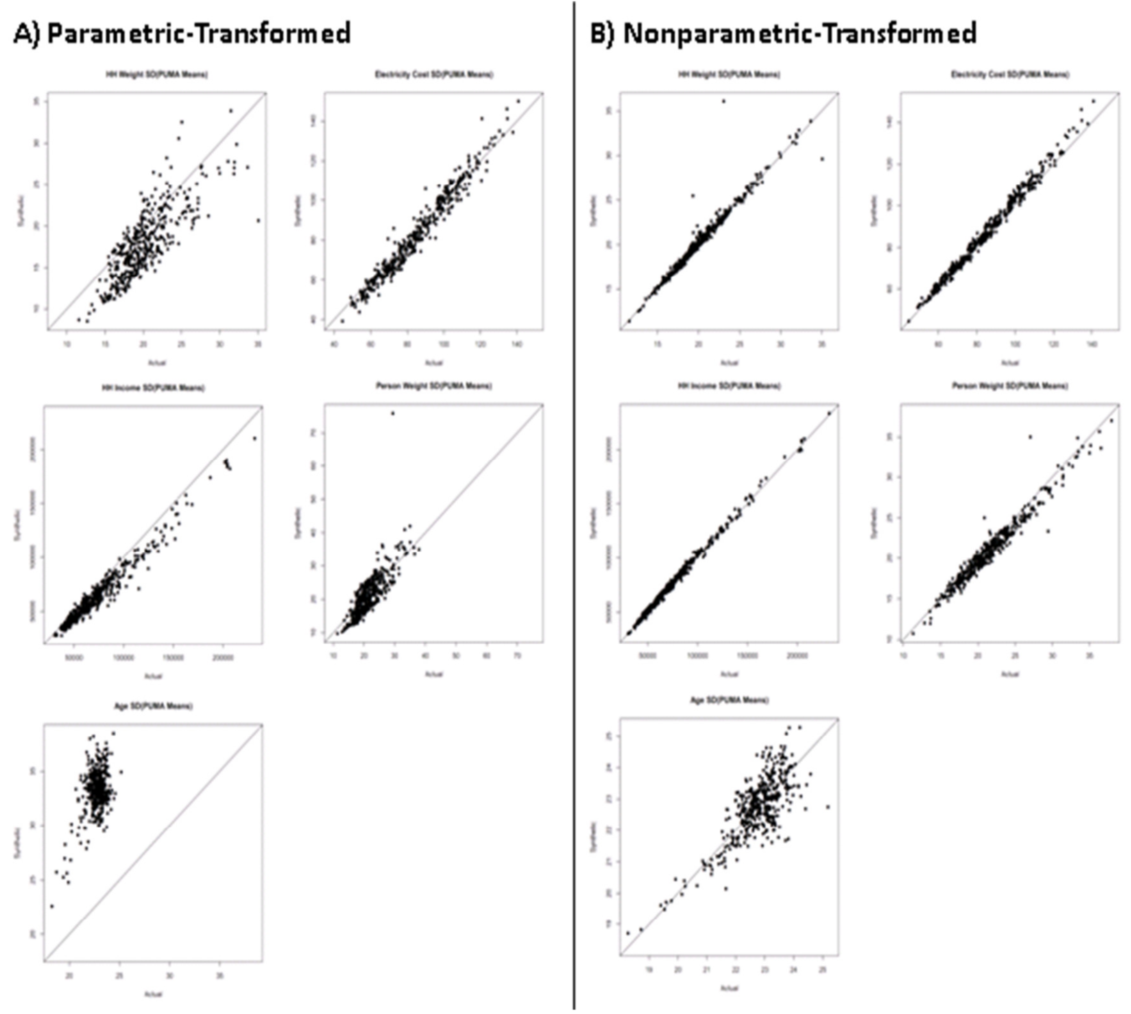
Figure 3.4 Scatter Plot of Synthetic (y-axis) and Actual (x-acis) PUMA Means for Nontransformed Variables



Scatter plots of synthetic and actual standard deviations of PUMA means are shown in Figures 3.5 and 3.6. Ideally, each scatter plot point should fall directly on the 45-degree line if the synthetic data accurately reflects the variability in the actual data. In nearly all cases, the nonparametric method yields standard deviations that are more closely aligned about the 45-degree line relative to the parametric method. The results are quite striking in some cases. For example, the standard deviations of age tend to be overestimated in the parametric-based synthetic data, but are markedly improved in the nonparametric-based synthetic data; the points are still widely dispersed but they are no longer overestimated and are centered about the 45-degree line. The parametric approach produces a significant amount of additional variation in the tail-end of the synthetic age distribution. The smoothing effect creates additional variation around the mean and causes the standard deviations to be larger than the actual standard deviations. In contrast, the tail-end of the nonparametric synthetic data distribution is more closely aligned with the actual distribution, and produces less of a smoothing effect. This results in synthetic standard deviations that correspond better with the actual standard deviations under the nonparmametric approach, than under the parametric approach.

In addition, the standard deviations for the household sampling weight tend to be widely dispersed and systematically underestimated in the parametric-based synthetic data. The dispersion and underestimation appears to be largely corrected under the nonparametric synthesization. However, there is still slight overestimation for the largest standard deviations under the nonparametric-transformed framework. This overestimation is fully corrected under the nonparametric-nontransformed framework, which suggests that the imputation procedures fail to preserve the tail-end of the transformed distribution. This result is consistent with findings from earlier research that has found problems with using imputation to adjust for item missing data for transformed totals in skewed populations (Rubin, 1983).

Figure 3.5 Scatter Plot of Standard Deviations of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Transformed Variables.



Because we adopt a fully-synthetic design and do not incorporate any auxiliary information into the imputation models, we would expect the standard errors of the synthetic PUMA estimates to be larger than the actual standard errors, on average. Figures 3.7 and 3.8 show scatter plots of the synthetic and actual standard errors of the means for the transformed and nontransformed variables, respectively. As expected, the synthetic data standard errors tend to be larger, on average, than the actual standard errors for these simple mean estimates. There does not appear to be any striking differences between the parametric and nonparametric or the transformed and nontransformed approaches. Each approach tends to reveal similar patterns in the scatter plots.

Figure 3.6 Scatter Plot of Standard Deviations of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Nontransformed Variables.

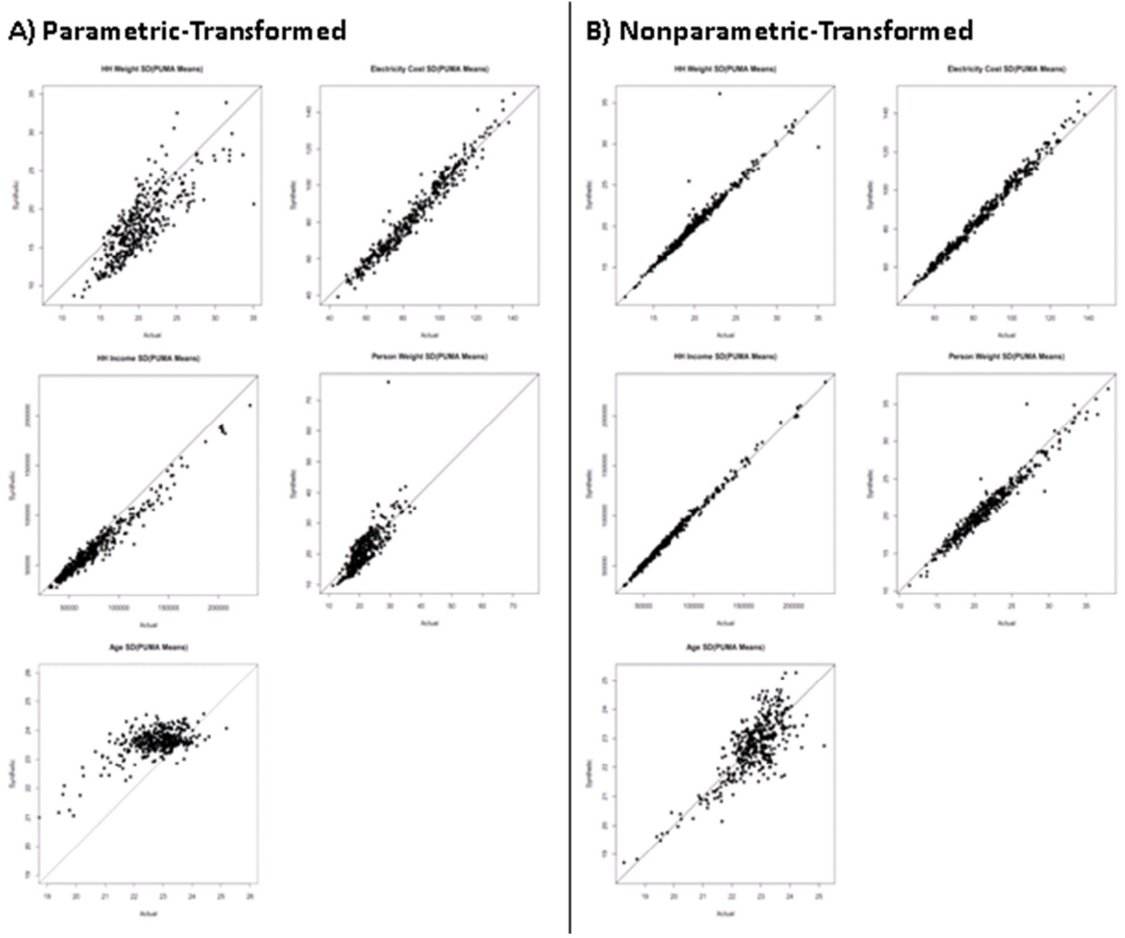
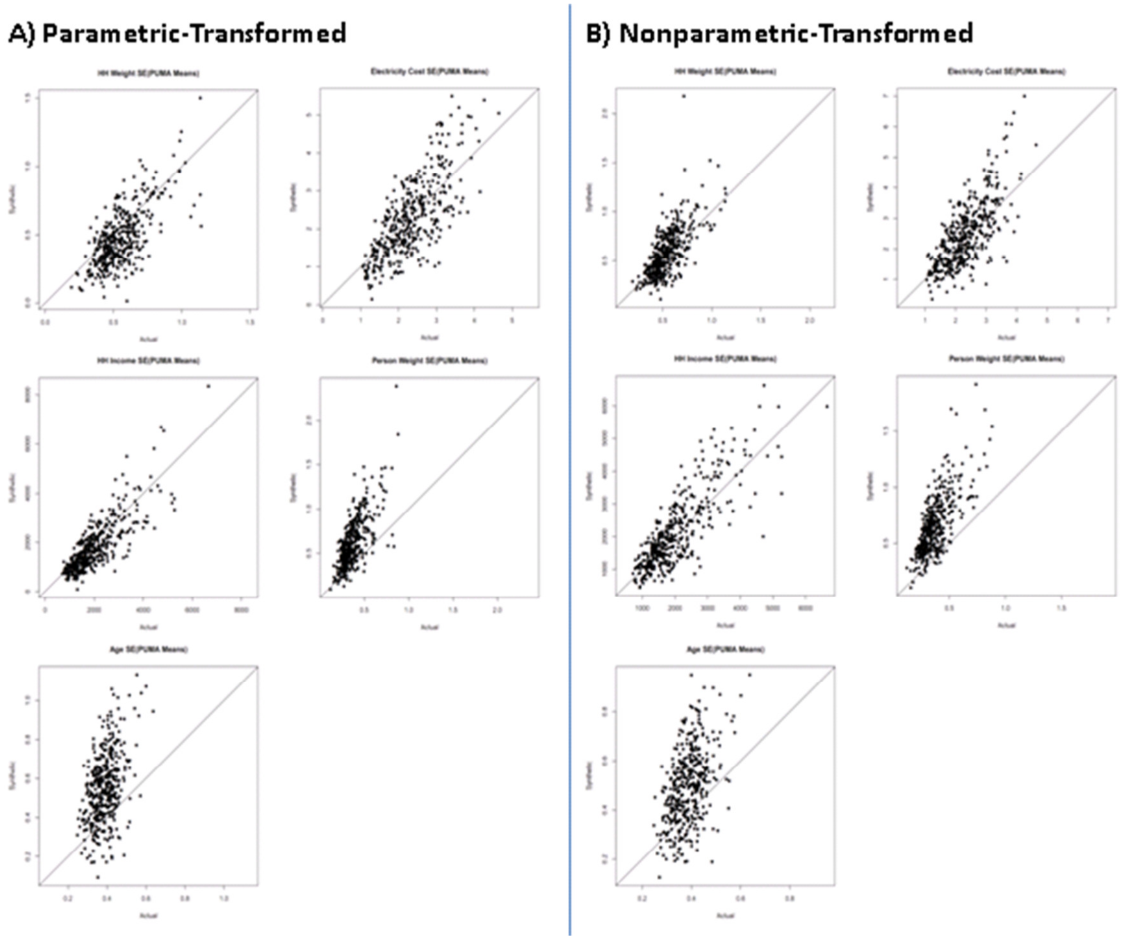


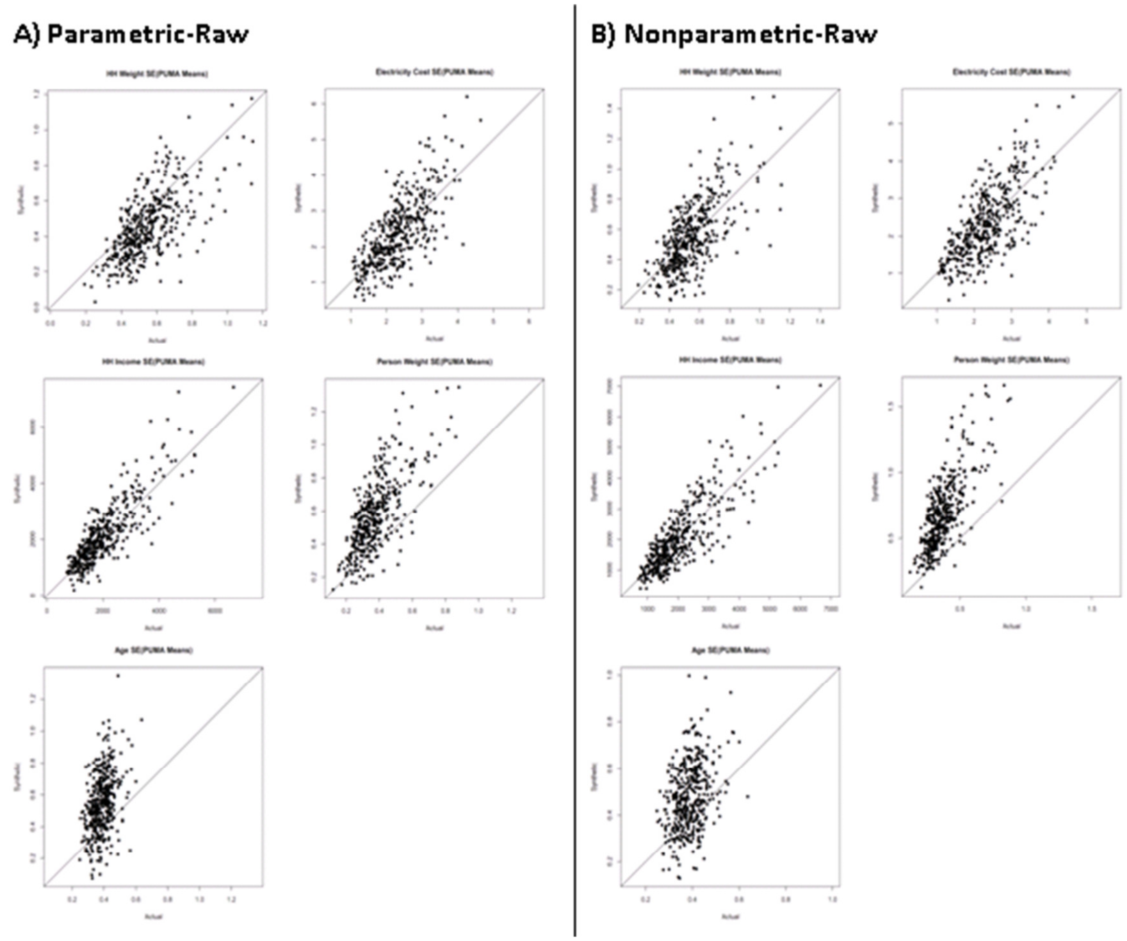
Figure 3.7 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Transformed Variables.



A) Parametric-Transformed

B) Nonparametric-Transformed

Figure 3.8 Scatter Plot of Standard Errors of Synthetic (y-axis) and Actual (x-axis) PUMA Means for Nontransformed Variables.



Next we turn our attention to recoded variable estimates, particularly percentile estimates. Such estimates are important to data users who may have interest in analyzing cases that lie within a certain portion of a distribution, including those that lie near the tail ends. Obtaining valid percentile estimates from synthetic data can be tricky, especially if the imputation model fails to adequately replicate the full range of the distribution.

Figures 3.9 and 3.10 show scatter plots of PUMA percentages of recoded household incomes greater than the 50th, 75th, and 90th percentiles for transformed and nontransformed household income variables, respectively. For the transformed variables, the synthetic 50th percentile estimates correspond well with the actual percentile estimates as indicated by the tightly clustered points that lie about the 45-degree line. For the most part, the 75th percentile estimates also lie about the equilibrium line, but the synthetic estimates tend to be overestimated as the PUMA proportions increase. For the 90th percentile estimates, there is significant departure between the actual estimates and synthetic estimates; the amount of overestimation of the synthetic estimates tends to be a positively correlated with the PUMA proportions. The analytic validity of the point estimates in the nonparametric-based synthetic data is equally poor. The same pattern is generally true for the nontransformed variables (Figure 3.10); however, the parametric data estimates are much worse than the nonparametric estimates. In general, the results suggest that the analytic validity of the percentiles estimates obtained from both the parametric and nonparametric methods tends to decrease as the percentile estimates become more extreme.

Figure 3.9 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Percentages for Transformed Household Income Percentiles (50th, 75th, and 90th).

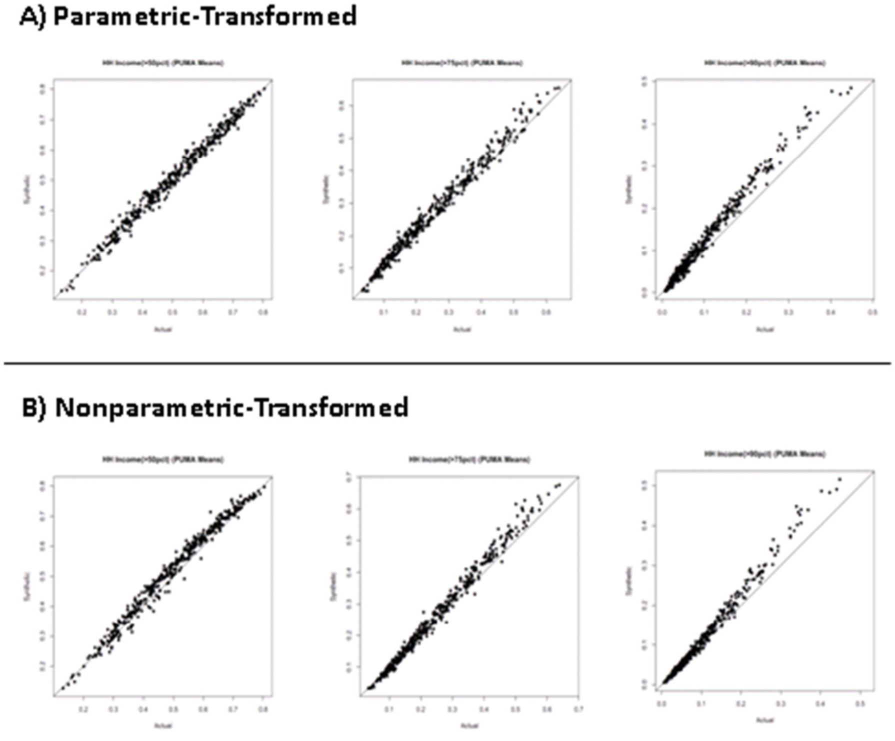
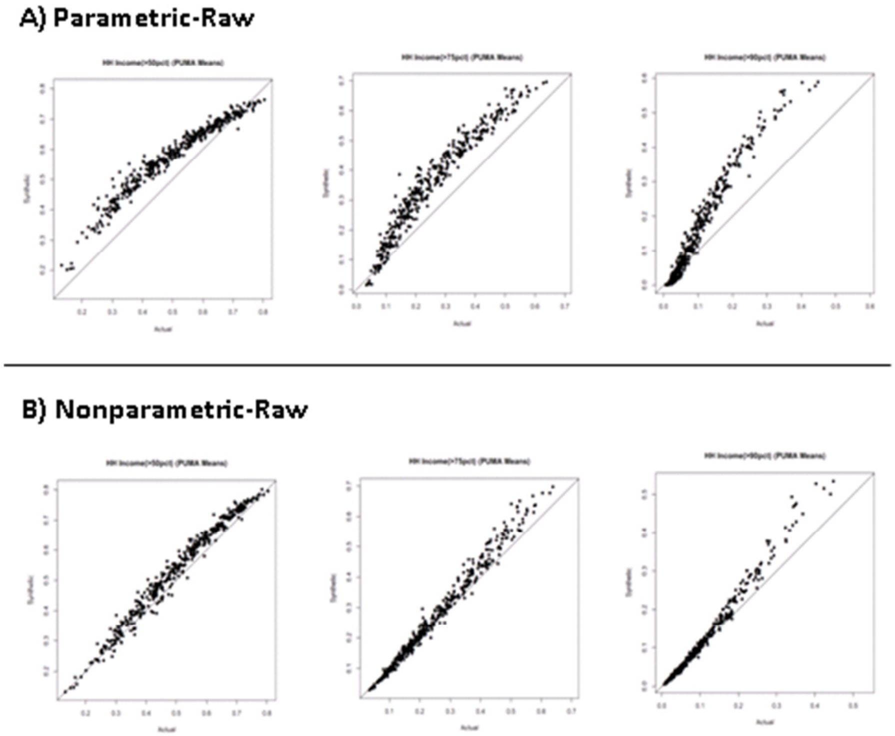


Figure 3.10 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PUMA Percentages for Nontransformed Household Income Percentiles (50th, 75th, and 90th).

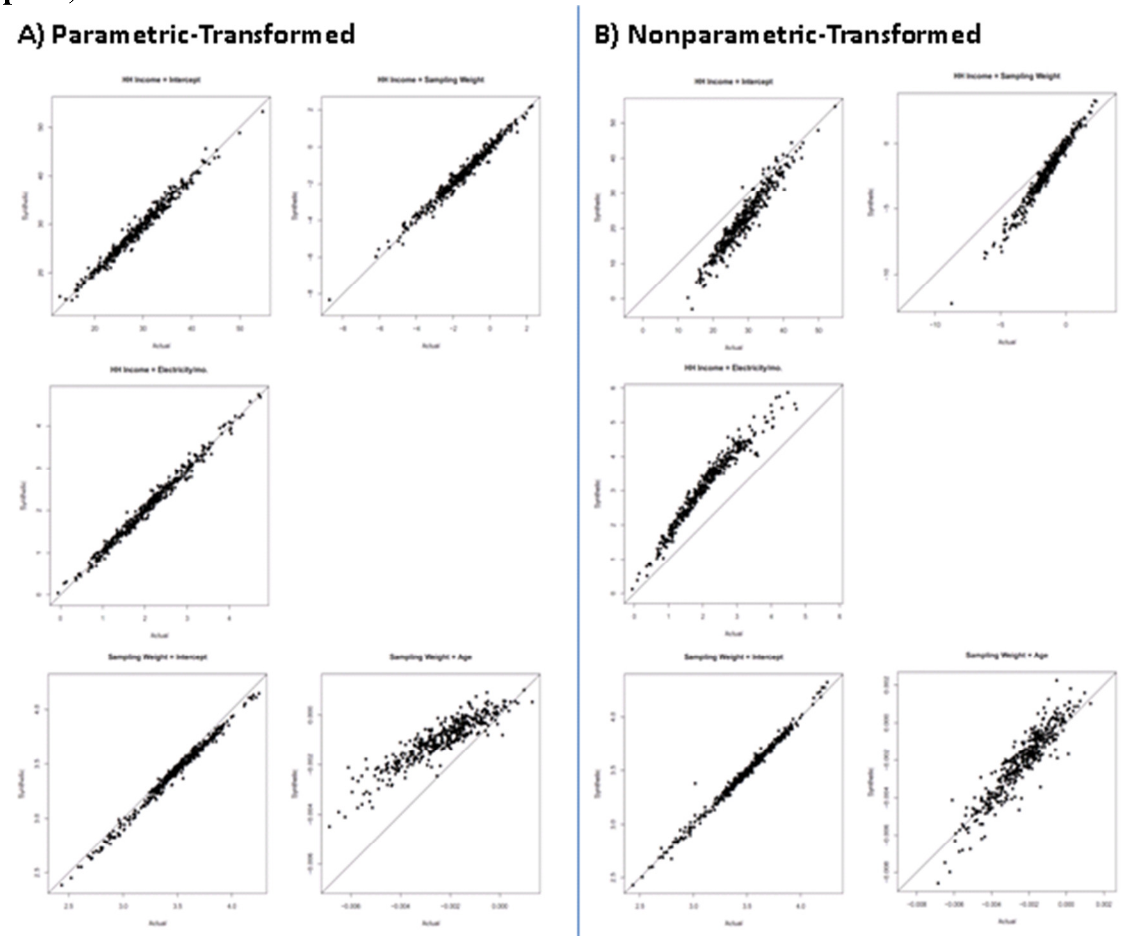


4.2 Validity of Multivariate Estimates

The next set of analyses assesses the analytic validity of synthetic multivariate estimates obtained from multiple regression models. Figures 3.11 and 3.12 show scatter plots of PUMA-level regression coefficients (and their standard errors) for very basic household- and person-level regression models fit within each PUMA, for transformed and untransformed variables, respectively. The dependent variable for the household-level regression model is household income (or log household income in the transformed model). For the bivariate person-level regression model the dependent variable is sampling weight (or log sampling weight in the transformed model). Two household-level models are fit: 1) main effects and; 2) main effects plus squared term for electricity costs. We acknowledge that these models may not be substantively appealing to analysts, but we use them strictly for the purpose of evaluating the analytic validity of the synthetic data methods.

For the transformed household-level main effects model (Figure 3.11; top 3 plots), the analytic validity of the estimated regression coefficients is higher for the parametrically-generated synthetic data than for the nonparametrically-generated data. For the nonparametric data, the regression coefficients are either severely underestimated for the smaller estimates or severely overestimated for the larger estimates. Where the nonparametric data seems to excel, however, is for the age predictor in the bivariate person-level regression model. For the nonparametric age coefficient scatter plot (Figure 3.11b; bottom-right plot), the synthetic data points are centered about the 45-degree line, in contrast to the parametric scatter plot (Figure 3.11a; bottom-right plot) which indicates that the synthetic age coefficients are severely overestimated relative to the actual coefficients. Recall that age is bimodal and was not transformed. In general, it appears that the nonparametric approach is only an improvement over the parametric approach in regression models when a predictor has a bimodal, or other non-normal shape.

Figure 3.11 Scatter Plots of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients for Transformed Household- (top 3 plots) and Person-Level (bottom 2 plots) Main Effects.

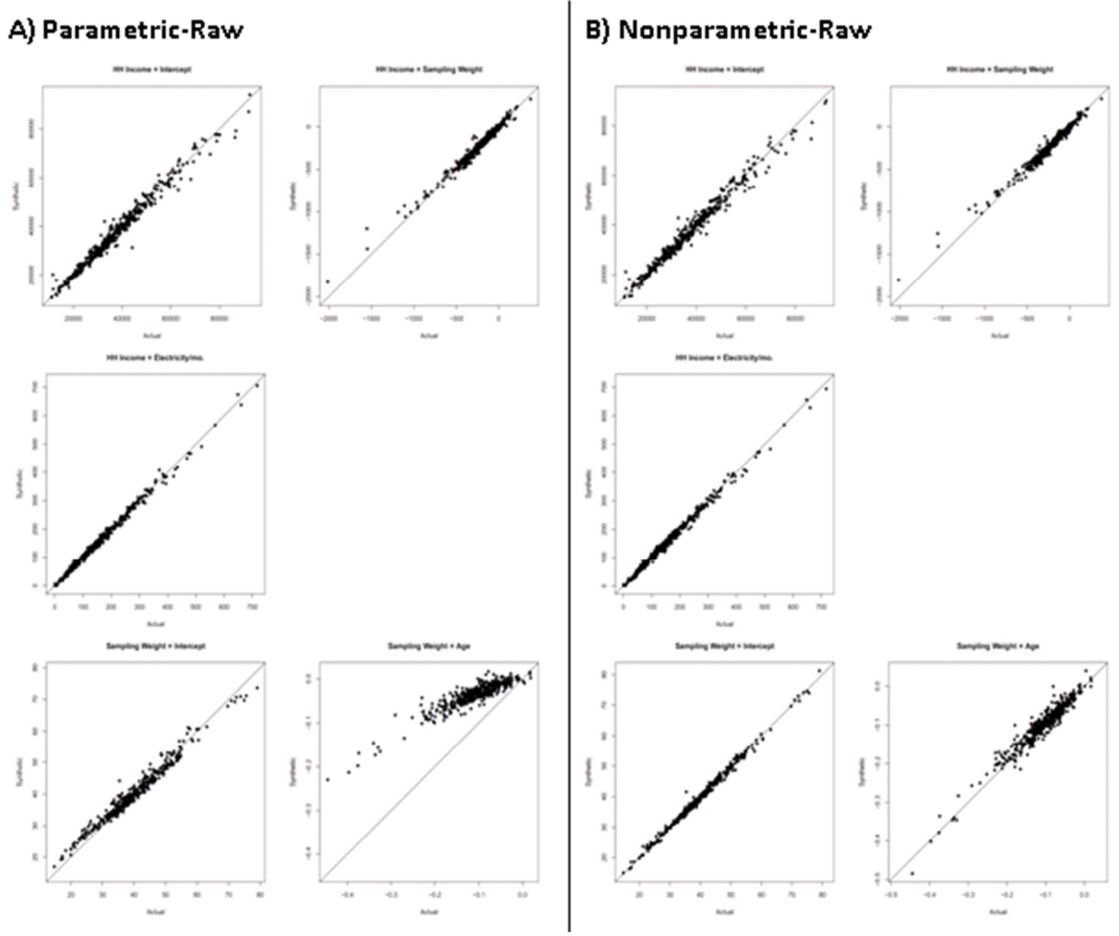


A) Parametric-Transformed

B) Nonparametric-Transformed

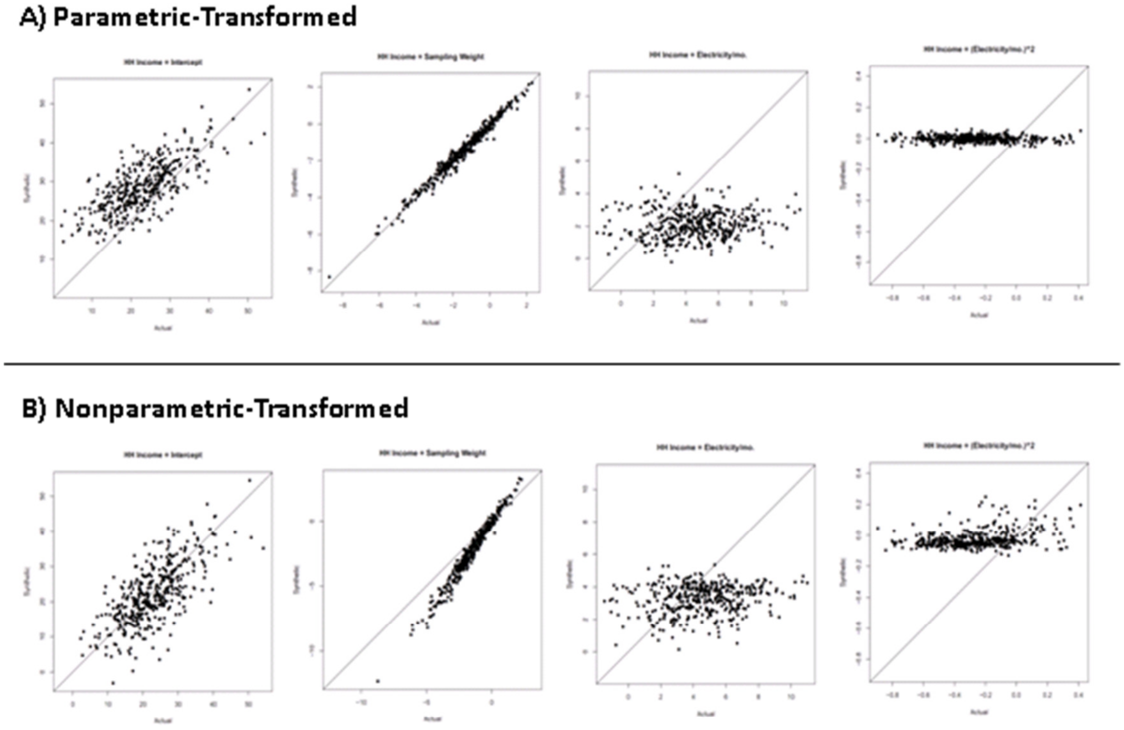
This generalization seems to hold true in the case of completely untransformed data as well. The scatter plots shown in Figure 3.12 indicate stronger correspondence between the actual and synthetic PUMA coefficient estimates under the nonparametric data approach, than under the parametric data approach. In fact, all of the nonparametric regression coefficients yield very high analytic validity. This result lends strong support to the nonparametric method in conjunction with untransformed variables, as it is the only combination that produces high analytic validity for all regression coefficients.

Figure 3.12 Scatter Plots of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients for Nontransformed Household- (top 3 plots) and Person-Level (bottom 2 plots) Main Effects.



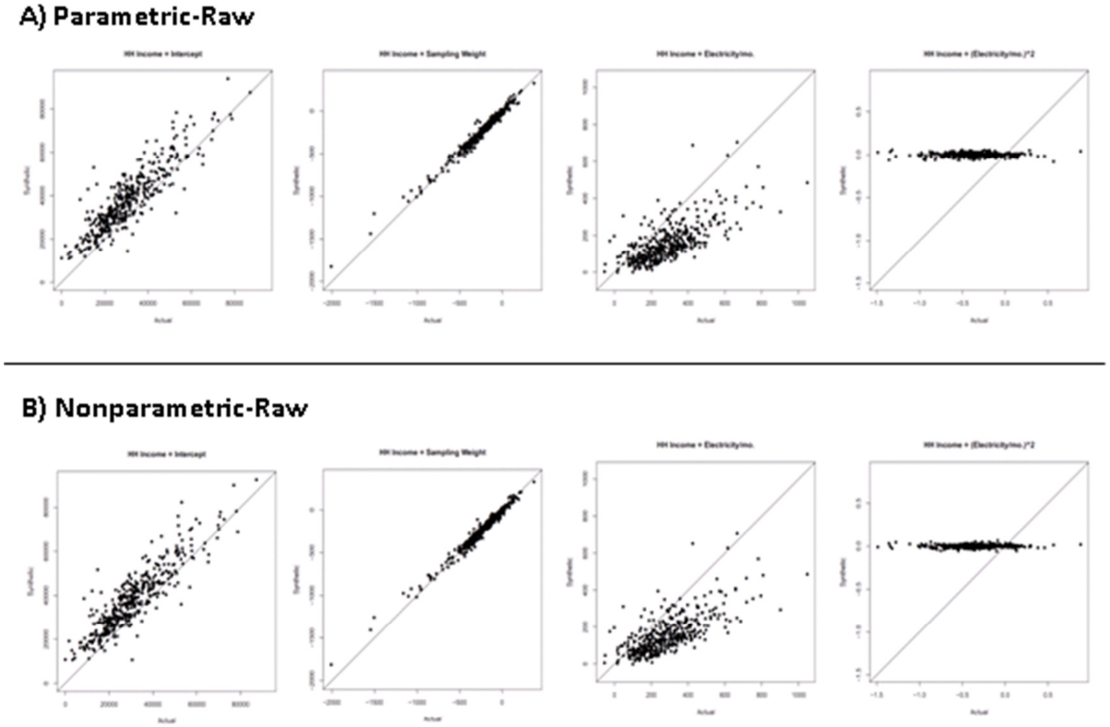
We now consider the effect of including a squared term in the synthetic regression model when the same term was omitted from the imputation model. In this scenario, the imputer's model is not in agreement or is "uncongenial" with the analyst's model of interest (Meng, 1994). Such disagreement should lead to attenuation of the squared variable term. We added a squared term for electricity cost to the household-level regression model. Scatter plots of actual and synthetic PUMA regression coefficients for main effects (left 3 plots) and the squared term (right-most plot) are shown in Figures 3.13 and 3.14 for transformed and nontransformed variables, respectively. Under both parametric and nonparametric approaches, the synthetic coefficient estimates for electricity squared are virtually zero, which is an expected result based on Meng's theory of congeniality. Hence it is worth emphasizing that the proposed nonparametric synthetic data method does not improve the analytic validity of higher-order terms that are omitted from the imputation model. In addition, the coefficient estimate for the main effect of electricity is essentially constant in the synthetic data for the transformed model. However, for the untransformed models (Figure 3.14), the validity of the synthetic electricity main effect term is much improved under either the parametric or nonparametric approaches; both approaches produce very similar synthetic coefficients. Thus it appears, that both the parametric and nonparametric data approaches, when applied to untransformed regression models, do a better job of defaulting to the main effects model when a higher-order term is included in the analyst's model, but not included in the imputer's model.

Figure 3.13 Scatter Plots of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients for Transformed Household-Level Main Effects and Squared Term.


A) Parametric-Transformed

B) Nonparametric-Transformed

Figure 3.14 Scatter Plots of Synthetic (y-axis) and Actual (x-axis) PUMA Regression Coefficients for Nontransformed Household-Level Main Effects and Squared Term.



A) Parametric-Raw

B) Nonparametric-Raw

4.3 Propensity Score Balance

Another indicator of the quality of the synthetic data is to assess the covariate balance between the synthetic and actual data. This is most easily performed using propensity scores (Rubin and Rosenbaum, 1983). Propensity scores are commonly used to identify imbalances in two or more groups (e.g., treatment and control groups) based on the distribution of a set of observed covariates. Biases caused by covariate imbalances may be adjusted by performing a weighted analysis with weights inversely proportional to the propensity scores (Ekholm and Laaksonen, 1991).

To assess the covariate balance between the synthetic and actual data sets, the actual data and a randomly selected synthetic data set are stacked vertically. Then an actual data indicator variable is regressed against all synthetic and actual variables using a logical regression model. The fitted model is used to obtain estimates of the propensity of a record belonging to the actual data. The propensity scores are then sorted and grouped into deciles and the proportions of synthetic and actual records are compared. If the synthetic and actual covariates are fully balanced, then the proportion of synthetic versus actual data should be approximately equal for each decile group. A chi-squared test with 9 degrees of freedom (if deciles are used) can be performed to assess the equivalence of the actual data proportions across the groups.

We use the propensity score balance method to assess the similarity of the synthetic and actual data in each PUMA for the parametric and nonparametric synthetic data generation methods. Table 3.7 shows summary statistics of the estimated probabilities of belonging to the actual data in each PUMA obtained from the household-level and person-level propensity models as well as test statistics for each parametric/nonparametric and transformed/nontransformed combination. The overall mean of estimated propensity scores was 0.13, which reflects the true proportion of actual data in each PUMA and the oversampling of synthetic data. Within each PUMA, the propensity scores were sorted and grouped into deciles and a chi-square statistic was computed. Small chi-square values indicate that the synthetic and actual data sets are balanced or statistically independent from each other, based on the set of covariates, while large values indicate poor covariate balance between the two data sets.

For the household-level data, the lowest mean chi-square values are observed for the nonparametric-nontransformed combination, followed by the parametric-transformed, nonparametric-transformed, and parametric-untransformed combination. For the person-level data, the lowest mean chi-square values are observed for the nonparametric-transformed combination, followed by the nonparametric-untransformed, parametric-transformed, and parametric-nontransformed. We interpret these results as supportive of the nonparametric method as it tends to produce synthetic data with a greater covariate balance relative to the parametric data method.

Table 3.7 Estimated Propensities of Belonging to the Actual Household-Level Data

PUMAs	**Households**			**Persons**		
Parametric-Transformed	Mean	Min	Max	Mean	Min	Max
Estimated probabilities $\hat{p}$	0.13	0.08	0.19	0.13	0.10	0.16
χ^2 statistic	63.06	31.14	207.95	455.27	250.57	862.92
P-value	0.03	0.00	0.16	0.00	0.00	0.00
Nonparametric-Transformed						
Estimated probabilities $\hat{p}$	0.13	0.04	0.25	0.13	0.10	0.17
χ^2 statistic	97.14	66.02	247.75	139.81	65.39	480.01
P-value	0.00	0.00	0.02	0.002	0.00	0.01
Parametric-Raw						
Estimated probabilities $\hat{p}$	0.13	0.05	0.21	0.13	0.04	0.19
χ^2 statistic	228.82	134.94	417.72	1175.01	810.07	1633.95
P-value	0.00	0.00	0.01	0.00	0.00	0.00
Nonparametric-Raw						
Estimated probabilities $\hat{p}$	0.13	0.08	0.20	0.13	0.10	0.17
χ^2 statistic	59.68	33.97	203.93	155.01	79.35	489.81
P-value	0.03	0.00	0.13	0.002	0.00	0.02

5 ACS-Based Simulation

This section evaluates the repeated sampling properties for small area inferences drawn from the synthetic data based on a simulation study. In this simulation, the 2005-2007 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. A total of 50,000 synthetic data sets are generated.

Two types of inferences can be obtained from the synthetic data sets: 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 most commonly encountered 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 confidence interval is obtained for each variable in each PUMA using the combining rule equations [1] and [2] in 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 estimates 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. $Q_{CIC} = 1$ 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 by,

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

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 from the 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 the synthetic interval completely covers the actual interval. Calculating the confidence interval overlap is only possible for conditional, not unconditional, 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 Confidence Interval Coverage

Tables 3.8 and 3.9 show the average confidence interval coverage (CIC) and confidence interval overlap (CIO) across all PUMAs for each household-level estimated mean computed at the PUMA- and State-level, respectively. For the transformation-based synthetic data estimates, the CIC is relatively high for basic (non-recoded) estimates ranging from 0.85-0.95 for the parametric data, and 0.88-0.99 for the nonparametric data; the corresponding range of CIC values for the recoded income variables is 0.52-0.89 and 0.77-0.84, respectively. 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 inferences for both the parametric and nonparametric results. In summary, the nonparametric synthetic data generation procedure produces univariate small area estimates with similar, and sometimes better, coverage properties as the parametric approach. The same general pattern holds true for the state-level confidence interval results shown in Table 3.9.

For the nontransformed-based synthetic data estimates, the confidence interval coverage is generally poor for both the parametric- and nonparametric-based approaches. One exception is the mean estimate of the household sampling weight, which yields mediocre coverage properties under the parametric approach (Conditional; CIC: 0.62, CIO: 0.51; Unconditional; CIC: 0.80), but exhibits a significant improvement under the nonparametric approach (Conditional; CIC: 0.99, CIO: 0.97; Unconditional; CIC: 0.99). This result suggests that the nonparametric approach has good coverage properties, especially when applied to nontransformed variables. However, the coverage properties for other variables are not as impressive. In fact, both the parametric and nonparametric approaches yield CIC and CIO values that are unimpressively low, ranging from 0.09-0.17 for conditional CIC values, 0.46-0.58 for conditional CIO values, and 0.08-0.19 for unconditional CIC values. There is no indication that the nonparametric approach outperforms the parametric approach for these untransformed variable estimates; both yield quite similar results. It is unclear why the coverage properties are quite good for the sampling weight estimate, but poor for all other household-level estimates.

Table 3.8 Simulation-Based Confidence Interval Results for Household-Level PUMA Means Based on Parametric/Nonparametric and Transformed/Nontransformed Data.

	Conditional Inference		Unconditional Inference	
Parametric-Transformed	CIC	CIO	CIC	CIC (Actual)
HH sampling weight	0.95	0.98	0.98	0.98
Electricity cost/mo.	0.86	0.87	0.90	0.98
HH income	0.90	0.91	0.94	0.98
Income > 50th pctile	0.89	0.92	0.94	0.98
Income > 75th pctile	0.71	0.72	0.80	0.98
Income > 90th pctile	0.52	0.61	0.62	0.97
Nonparametric-Transformed				
HH sampling weight	0.99	0.97	0.99	0.98
Electricity cost/mo.	0.88	0.88	0.92	0.98
HH income	0.93	0.91	0.96	0.98
Income > 50th pctile	0.77	0.78	0.81	0.98
Income > 75th pctile	0.78	0.78	0.85	0.98
Income > 90th pctile	0.84	0.80	0.90	0.97
Parametric-Raw				
HH sampling weight	0.62	0.51	0.80	0.98
Electricity cost/mo.	0.11	0.58	0.10	0.98
HH income	0.17	0.51	0.19	0.98
Income > 50th pctile	0.16	0.50	0.18	0.98
Income > 75th pctile	0.10	0.29	0.11	0.98
Income > 90th pctile	0.08	0.24	0.08	0.97
Nonparametric-Raw				
HH sampling weight	0.99	0.97	0.99	0.98
Electricity cost/mo.	0.12	0.58	0.15	0.98
HH income	0.17	0.50	0.14	0.98
Income > 50th pctile	0.09	0.55	0.10	0.98
Income > 75th pctile	0.15	0.53	0.18	0.98
Income > 90th pctile	0.13	0.46	0.15	0.97

Table 3.9 Simulation-Based Confidence Interval Results for Household-Level State Means Based on Parametric/Nonparametric and Transformed/Nontransformed Data.

	Conditional Inference		Unconditional Inference	
Parametric-Transformed	CIC	CIO	CIC	CIC (Actual)
HH sampling weight	0.64	0.99	0.74	0.99
Electricity cost/mo.	0.13	0.59	0.24	0.98
HH income	0.33	0.75	0.48	0.99
Income > 50th pctile	0.67	0.85	0.67	0.99
Income > 75th pctile	0.29	0.30	0.35	0.99
Income > 90th pctile	0.38	0.47	0.49	0.98
Nonparametric-Transformed				
HH sampling weight	0.89	0.89	0.92	0.99
Electricity cost/mo.	0.16	0.61	0.28	0.98
HH income	0.40	0.72	0.52	0.99
Income > 50th pctile	0.38	0.50	0.37	0.99
Income > 75th pctile	0.41	0.48	0.42	0.99
Income > 90th pctile	0.44	0.44	0.47	0.98
Parametric-Raw				
HH sampling weight	0.00	0.00	0.00	0.99
Electricity cost/mo.	0.00	0.50	0.00	0.98
HH income	0.00	0.50	0.00	0.99
Income > 50th pctile	0.36	0.37	0.40	0.99
Income > 75th pctile	0.00	0.00	0.00	0.99
Income > 90th pctile	0.00	0.00	0.00	0.98
Nonparametric-Raw				
HH sampling weight	0.98	0.89	0.99	0.99
Electricity cost/mo.	0.00	0.50	0.00	0.98
HH income	0.00	0.50	0.00	0.99
Income > 50th pctile	0.00	0.50	0.00	0.99
Income > 75th pctile	0.00	0.50	0.00	0.99
Income > 90th pctile	0.00	0.50	0.00	0.98

6 Application: Restricted ACS County-Level Data

In addition to the public-use microdata, restricted ACS microdata for years 2005-2009 were obtained from the Michigan Census Research Data Center and used to demonstrate the proposed synthetic data method. The restricted data contain identifiers for all counties in the United States. We restrict the data to the Northeast region which contains 217 counties, in contrast to the public-use microdata which contains 405 public-use microdata areas (PUMAs). Although 3 years of microdata were used in the public-use application, we use the restricted 5-year data set to facilitate the disclosure review and allow the publication of estimates for all counties. The same variables shown in Table 2.0 were synthesized in this application. The synthetic data estimates are based on $M = 10$ imputations.

Tables 3.10 and 3.11 show summary measures of actual and synthetic county means for transformed and non-transformed variables, respectively. In general, the synthetic means correspond relatively closely to the actual estimates, on average, with the parametric-transformed, nonparametric-transformed, and nonparametric-raw combinations all yielding very similar results. As in the public-use application, the actual and synthetic point estimates correspond relatively closely when applied to actual counties. This finding should give confidence to the synthetic data methodology, as the method is practically useful when applied to actual small areas, such as counties, as opposed to combined counties or PUMAs.

Table 3.10 Summary Measures of Actual and Synthetic County Means for Transformed Variables.

	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Parametric - Transformed				
Household variables				
Sampling weight	9.99	9.96	0.11	0.10
Electricity bill/mo.	118.89	118.28	1.25	1.04
Income	67983.89	67145.59	1067.29	747.62
Person variables				
Sampling weight	10.27	10.43	0.08	0.13
Age	40.89	41.48	0.25	0.28
Recodes				
Income > 50th pctile (%)	44.65	44.55	0.80	0.65
Income > 75th pctile (%)	19.34	21.02	0.59	0.43
Income > 90th pctile (%)	6.78	8.08	0.35	0.24
Nonparametric - Transformed				
Household variables				
Sampling weight	9.99	10.09	0.11	0.12
Electricity bill/mo.	118.89	117.15	1.25	1.26
Income	67983.89	67203.91	1067.29	1056.91
Person variables				
Sampling weight	10.27	10.36	0.08	0.14
Age	40.89	41.48	0.25	0.23
Recodes				
Income > 50th pctile (%)	44.65	45.63	0.80	0.64
Income > 75th pctile (%)	19.34	19.93	0.59	0.49
Income > 90th pctile (%)	6.78	7.15	0.35	0.32

Table 3.11 Summary Measures of Actual and Synthetic County Means for Non-Transformed Variables.

	Avg. Mean		**Avg. Standard Error of Mean**	
	Actual	Synthetic	Actual	Synthetic
Parametric - Raw				
Household variables				
Sampling weight	9.99	11.52	0.11	0.08
Electricity bill/mo.	118.89	118.09	1.25	1.26
Income	67983.89	67334.78	1067.28	1138.35
Person variables				
Sampling weight	10.27	12.09	0.08	0.11
Age	40.89	41.08	0.25	0.29
Recodes				
Income > 50th pctile (%)	44.65	52.66	0.80	0.72
Income > 75th pctile (%)	19.34	29.32	0.59	0.55
Income > 90th pctile (%)	6.78	11.45	0.35	0.29
Nonparametric - Raw				
Household variables				
Sampling weight	9.99	10.04	0.11	0.12
Electricity bill/mo.	118.89	118.36	1.25	1.17
Income	67983.89	67802.42	1067.29	1127.80
Person variables				
Sampling weight	10.27	10.40	0.08	0.15
Age	40.89	41.49	0.25	0.21
Recodes				
Income > 50th pctile (%)	44.65	46.31	0.79	0.72
Income > 75th pctile (%)	19.34	20.18	0.59	0.56
Income > 90th pctile (%)	6.78	7.21	0.35	0.34

7 Conclusions

In this chapter, we proposed and evaluated a continuous nonparametric simulation procedure for generating synthetic data for small geographic areas. The procedure is based on a hierarchical model which is appropriate for producing small area estimates, and can be easily implemented in large-scale applications. The method produces relatively high analytic validity for both simple univariate and multivariate estimates obtained from skewed and bimodal distributions. The analytic validity achieved by the nonparametric method is typically equivalent, or better, than the standard parametric method. The greatest improvements in analytic validity tend to be achieved when the nonparametic method is applied to non-normal and nontransformed variables (although the method can produce valid estimates for transformed data as well). This is a useful property of the method from a practical perspective as it does not require the imputer to transform the data in advance of the synthesization, which can be a time-consuming and rather subjective process.

Other practical advantages of the method include its versatility in terms of handling both skewed and bimodal distributions. Although the nonparametric procedure did not completely replicate the bimodal shape or upward concavity of the age distribution in the evaluation, it still seemed to produce more valid small area estimates, particularly for estimates of regression coefficients, than the parametric approach. In addition, the method yields relatively good analytic validity for estimating percentiles from recoded continuous variables. Although the method is intended for continuous variables, switching between the nonparametric and alternative parametric approaches for non-continuous variables (e.g., categorical) is possible and can be easily implemented in practical applications.

Some limitations of the method should also be noted. Although we refer to the method as a nonparametric one, the method itself is not completely nonparametric. The linear regression estimates obtained in Stage 1 still assume that the usual regression assumptions (e.g., normality of the error distribution) hold. In addition, the hierarchical Bayesian model assumes that the random effects are distributed as multivariate normal, which is an assumption we did not verify. A fully nonparametric data generation approach may have yielded greater analytic validity than the semi-parametric approach we considered here. Another limitation relates to the mixed repeated sampling properties of the method. In the simulation study, the nonparametric data generation method yielded good confidence interval coverage when applied to the transformed data; however, when applied to the nontransformed data the results were decidedly mixed, which may indicate an underlying problem with the method in repeated applications.

There are several possible extensions to this research. First, the proposed method could potentially be expanded into a fully nonparametric procedure by modeling the conditional densities (Stage 1) and random effects using nonparametric procedures. The method may also be combined with other nonparametric approaches (e.g., CART) to synthesize other types of variables (e.g., categorical) in a completely nonparametric synthetic framework. In addition, the method may be extended to handle item missing data prior to synthesization. This approach has been considered in single-level applications (Reiter, 2004), but never in a multilevel context when small area estimates are needed.

In conclusion, the proposed nonparametric synthetic data generation approach shows promise in the small area applications considered here. The method is easily implemented and can potentially be used in large-scale applications to produce public-use microdata for small geographic areas that are normally restricted to research data centers. The method addresses an important concern expressed by data users who are skeptical that the non-standard distributions and relationships in actual data files will be maintained and preserved in synthetic data files. As the demand for public-use microdata for small areas continues to grow, the synthetic data framework seems to be a promising option for releasing geographically-relevant data to users that are otherwise unable to obtain the data they need to pursue their research.

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 $p = (1,2,\ldots,P)$ 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,

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

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

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,

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

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

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

Chapter 4

Synthetic Data for Small Area Estimation in a Complex Sample Survey

1 Introduction

High quality survey data are often collected and used to monitor the health and well-being of populations. Such data are needed to establish baseline outcomes and monitor the progress of goals and objectives towards improving the health of the population. A prominent example of this strategy is the Healthy People initiative started in 2000 by the Department of Health and Human Services. The Healthy People initiative, started in 1990 by the Department of Human and Health Services, is a prominent example of using survey and other data sources to monitor and assess progress towards achieving hundreds of priority health objectives in the United States (USDHHS, 2010). Many key indicators for these objectives are obtained from leading public health surveillance systems, including vital statistics and population-based surveys, such as the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES), among others, which produce important national- and state-level statistics of interest to policy-makers and health professionals.

A limitation of these surveys is that they are not intended for the production of sub-state and other small area estimates. Small area estimates are of particular interest to county administrators and city planners, who may be interested in developing their own health initiatives in their local areas. Such estimates could also be used to inform the allocation of resources to support healthcare delivery systems and interventions at the local level. In addition, small areas, such as counties, or cities, may be used as test beds for innovative health programs that, if successful, could be implemented more broadly. Having cost-effective data monitoring systems in place to monitor and assess the effectiveness of these small area interventions would be invaluable, especially if existing and means-tested survey data could be used for the assessment.

Existing health survey data sources, such as NHIS and NHANES, have limitations that prevent them from being utilized for production of small area estimates and dissemination of small area microdata. First, neither survey was designed to produce reliable small area estimates. Only national estimates can be obtained from NHANES, while NHIS can produce national-, regional-, and state-level estimates. Finer levels of geographical identification are only accessible via Research Data Centers (RDCs). Second, the NHANES and NHIS surveys have complex sample designs, and most small areas of interest (e.g., counties, cities) contain no sampled cases. Hence, even if small area identifiers could be obtained for all small areas, there's no guarantee that any sampled cases will be available for analysis. Third, the sensitive nature of the survey content poses confidentiality concerns. In the context of small geographic areas, the possibility of reidentifying respondents in the survey data is non-trivial. Respondents living within sparse areas are susceptible to disclosure if they self-report unique and other identifying information during the survey interview. This is an important issue in the context of complex sample surveys, as an intruder would only need to know whether a particular area (or PSU) was sampled (and possibly a few unique personal characteristics) in order to narrow their search and successfully identify a respondent's record and survey responses.

1.1 Synthetic Data for Small Geographic Areas Based on Complex Sample Survey Data

The purpose of this chapter is to propose and evaluate an approach that overcomes many of the limitations associated with using population-based, complex sample surveys to disseminate microdata and produce estimates for small geographic areas. The basic idea is to generate synthetic data for sampled and non-sampled small areas. Synthetic data, originally proposed by Rubin (1993), replaces the observed data values with multiply-imputed, or synthetic, values. The conceptual idea behind the method 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 and released as public-use microdata. Valid inferences are obtained by analyzing each synthetic data set independently and combining the point estimates and their standard errors using standard combining rules (Raghunathan, Reiter, and Rubin, 2003).

The synthetic data framework offers many potential advantages in terms of disseminating microdata for small geographic areas and protecting data confidentiality based on complex sample survey data. Although the majority of synthetic data applications focus on replacing the observed values with synthetic values (Rodriguez, 2007; Abowd, Stinson, and Benedetto, 2006; Kinney and Reiter, 2008), it is also possible to generate and disseminate synthetic data for the unobserved cases in non-sampled areas based on an imputation model fitted using the observed cases. In addition, the imputation model can account for complex sample design features, which is the safest course of action in the specification of imputation models from a design-based perspective (Reiter, Raghunathan, and Kinney, 2006).

The synthetic data framework also offers several data protection benefits. For example, because the observed values are replaced with synthetic, yet plausible, values no actual data are released. The majority of synthetic data research has focused on synthesizing only a subset of survey variables that pose greater-than-average disclosure risks (Little, 1993; Kennickell, 1997; Liu and Little, 2002; Reiter, 2003, 2005). A more extreme approach is to synthesize all variables and release only synthetic data to the public. The former approach tends to yield greater analytic validity than the latter, but the latter tends to achieve greater data protection (Drechsler, Bender, and Raessler, 2008). We focus on the latter "fully synthetic" data approach as we believe it offers the greatest level of confidentiality protection for small area applications. A further benefit of the fully synthetic data approach is that it can easily be extended to handle non-sampled areas and cases. Generating synthetic data for non-sampled areas/units offers further data protection as it masks the sampled areas and makes it difficult for an intruder to distinguish between sampled and non-sampled areas. It also allows data users to study characteristics of small areas that were never sampled in the survey; hence, the utility of the survey data is potentially enhanced.

1.2 Organization of Chapter

This chapter proposes an extension of Rubin's synthetic data method for the purpose of generating fully-synthetic microdata sets for small geographic areas based on complex sample survey data. A hierarchical Bayesian model is proposed that accounts for multiple levels of geography and "borrows strength" across related areas using auxiliary information known for small and large geographical areas. A sequential multivariate regression procedure is used to approximate the joint distribution of the observed data, which is used to simulate synthetic values from the posterior predictive distribution (Raghunathan et al., 2001). The method is demonstrated on restricted data from the National Health Interview Survey (NHIS), an ongoing complex sample survey used to monitor trends in illness and disability and to track progress toward achieving national health objectives. The method is adapted to explicitly account for the stratification and clustering employed in the NHIS. Synthetic data is generated for several commonly used 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 are not addressed and we leave this to future work. Limitations of the model 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, Reiter, and Rubin (2003) and Reiter (2004). 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, I assume there are no item missing data in the observed survey data set, 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, 10, or 20) 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 $D_{syn} = (D^{(l)}; l = 1,2, \dots, M)$ 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 $P_{syn} = \left(P^{(l)}; l = 1,2, \ldots, M\right)$ under simple random sampling. The values of q and v computed on the M synthetic data sets are denoted by $\left(q^{(l)}, v^{(l)}, l = 1,2, \ldots, M\right)$.

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, Reiter, and Rubin (2003), is to treat $\left(q^{(l)}, v^{(l)}; l = 1,2, \ldots, M\right)$ 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, Reiter, and Rubin (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 the units that comprise the synthetic samples from each multiply-imputed synthetic population. Because of this additional sampling step, the between imputation variance already reflects the within imputation variability, which is not the case in the usual multiple imputation framework.

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) and Reiter (2005).

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 M or n_{syn}. Numerical routines can be used to calculate the integrals involved in the construction of T_M, yielding more precise variance estimates (Raghunathan, Reiter, and Rubin, 2003). A simpler variance approximation that is always positive is given in Reiter (2002).

3 Extension to Small Geographic Areas Based on Complex Sample Survey Data

I adopt a hierarchical Bayesian model to generate synthetic data for small geographic areas based on complex sample survey data. 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 (Yucel, 2008; Reiter, Raghunathan, and Kinney, 2006).

My approach 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 stage 1 is approximated and the between-area variation is modeled using auxiliary information. In the final stage, the unknown regression parameters are simulated for both sampled and non-sampled areas and used to draw synthetic microdata values from the posterior predictive distribution.

3.1 Stage 1: Approximation of Joint Density via Sequential Regression

For descriptive purposes, I introduce the following notation. I define "small areas" as primary sampling units (PSUs) (or counties), nested within strata (or states), which could also be nested within even larger areas (e.g., regions). In specific terms, 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 sampled PSU $c = (1,2, \dots, C_s)$ nested within stratum $s = (1,2, \dots, S)$. Let $Y_{cs} = \left(Y_{ics,p}; i = 1,2, \dots, n_{cs};\ p = 1,2, \dots, P\right)$ represent the $n_{cs} \times P$ matrix of survey variables collected from each survey respondent located in PSU c and stratum s. Let $X_{cs} = \left(X_{ics,j}; i = 1,2, \dots, n_{cs}, n_{cs} + 1, \dots, N_{cs};\ j = 1,2, .., J\right)$ represent the $N_{cs} \times J$ matrix of auxiliary or administrative variables known for every population member in a particular PSU and stratum. Although I 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 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 PSU-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 modeling 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 PSU-level regression parameters $\beta_{cs,p}$ are assumed to follow a multivariate normal distribution,

$$\beta_{cs,p} \sim MVN\left(\beta_{s,p} Z_{cs}, \Sigma_{s,p}\right) \tag{4}$$

where $Z_{c,s} = \left(Z_{s,k}; k = 1,2, \dots, K\right)$ is a $K \times 1$ matrix of PSU-level covariates, $\beta_{s,p}$ is a $(J + p) \times K$ matrix of unknown regression parameters, and $\Sigma_{s,p}$ is a $(J + p) \times (J + p)$ covariance matrix. PSU-level covariates are incorporated into the hierarchical model in order to “borrow strength” from related small areas. Prior distributions may be assigned to the unknown parameters $\beta_{s,p}$ and $\Sigma_{s,p}$, but for computational simplicity I assume that $\beta_{s,p}$ and $\Sigma_{s,p}$ are fixed at their respective maximum likelihood estimates (MLE), 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.

The estimated state-level parameters $\hat{\beta}_{s,p}$ (obtained from the maximum likelihood estimation step) are expressed as a $[(J+p)\times K]\times 1$ vector and approximated by a multivariate normal distribution,

$$\hat{\beta}_{s,p} \sim MVN\left(\beta_{s,p}, \hat{V}_{s,p}\right) \tag{5}$$

where $\beta_{s,p}$ is a $[(J+p)\times K]\times 1$ vector of unknown parameters and $\hat{V}_{cs,p}$ is the corresponding $[(J+p)\times K]\times[(J+p)\times K]$ estimated covariance matrix obtained from the Kronecker product of $I_K \otimes \Sigma_{s,p}$.The unknown stratum-level regression parameters $\beta_{s,p}$ are assumed to follow a multivariate normal distribution,

$$\beta_{s,p} \sim MVN\left(\beta_p Z_s, \Omega_p\right) \tag{6}$$

where $Z_s = \left(Z_{s,k}; t = 1,2,\ldots,T\right)$ is a $T\times 1$ matrix of stratum-level covariates, β_p is a $[(J+p)\times K]\times T$ matrix of unknown parameters, and Σ_p is a $[(J+p)\times K]\times [(J+p)\times K]$ covariance matrix. Again, we assume that the hyperparameters (in this case, $\beta_p$and Σ_p) are fixed at their maximum likelihood estimates by the EM algorithm. The details of the EM algorithm implementation can be found in Appendix 2.

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

$$\tilde{\beta}_{cs,p} \sim MVN\left[\left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_{s,p}^{-1}\right)^{-1}\left(\hat{V}_{cs,p}^{-1}\hat{\beta}_{cs,p} + \hat{\Sigma}_{s,p}^{-1}\hat{\beta}_{s,p}Z_{cs}\right), \left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_{s,p}^{-1}\right)^{-1}\right] \tag{7}$$

$$\tilde{\beta}_{s,p} \sim MVN\left[\left(\hat{V}_{s,p}^{-1} + \hat{\Omega}_{p}^{-1}\right)^{-1}\left(\hat{V}_{s,p}^{-1}\hat{\beta}_{s,p} + \hat{\Omega}_{p}^{-1}\hat{\beta}_{p}Z_{s}\right), \left(\hat{V}_{s,p}^{-1} + \hat{\Omega}_{p}^{-1}\right)^{-1}\right] \tag{8}$$

where $\tilde{\beta}_{cs,p}$ and $\tilde{\beta}_{cs,p}$ are simulated vectors of values for the unknown parameters $\beta_{cs,p}$ and $\beta_{s,p}$, respectively.

For the nonsampled PSUs it is not possible to fit sequential regression models and obtain direct estimates of the regression parameters in Stage 1. We therefore must rely on a purely model-based approach to obtain values of $\tilde{\beta}_{cs,p}$ and $\tilde{\beta}_{s,p}$ for the nonsampled areas. Specifically, for this purpose we use the model equations [4] and [6] from above, which are repeated below for convenience,

$$\beta_{cs,p} \sim MVN\left(\beta_{s,p} Z_{cs}, \Sigma_{s,p}\right) \tag{4}$$

$$\beta_{s,p} \sim MVN\left(\beta_p Z_s, \Omega_p\right) \tag{6}$$

It should be noted that the PSU- and stratum-level auxiliary variables, denoted by Z_{cs} and Z_s, respectively, must be known for all nonsampled areas. The implementation steps are described as follows,

1. Draw a $[(J+p) \times K] \times 1$vector of values $\tilde{\beta}_{s,p}$ from a multivariate normal distribution with location parameter $\beta_p Z_s$ and scale parameter Ω_p, where β_p and Ω_p are replaced with their maximum likelihood estimates $\hat{\beta}_p$and $\widehat{\Omega}_p$, respectively, which were already obtained from the second EM implementation for all sampled cases.
2. Vectorize the drawn values of $\tilde{\beta}_{s,p}$ to obtain a $(J+p) \times K$ matrix, i.e., $vec\left(\tilde{\beta}_{s,p}\right)$.
3. Draw a $(J+p) \times 1$ vector of values $\tilde{\beta}_{cs,p}$ from a multivariate normal distribution with location parameter $\beta_{s,p} Z_{cs}$ and scale parameter $\Sigma_{s,p}$, where $\beta_{s,p}$ and $\Sigma_{s,p}$ are replaced with their maximum likelihood estimates $\hat{\beta}_{s,p}$ and $\hat{\Sigma}_{s,p}$, respectively,

which were already obtained from the first EM implementation for all sampled cases within stratum s.

4. Once $\tilde{\beta}_{cs,p}$ has been drawn for the sampled and non-sampled small areas, the actual synthetic values can be simulated from the posterior predictive distribution (Stage 3) using the instructions described in the next section.

3.3 Stage 3: Simulating from the Posterior Predictive Distribution

The ultimate objective is to generate synthetic populations for each sampled and non-sampled small area using an appropriate posterior predictive distribution. Simulating a synthetic variable $\tilde{Y}_{cs} = \left(\tilde{Y}_{lcs,p}; l = 1,2,\ldots,N_{cs}; p = 1,2,\ldots,P\right)$ for observed (or unobserved) variable Y_{cs} for synthetic population unit $l = (1,2,\ldots,N_{cs})$ is achieved by drawing, in sequential fashion, from the posterior predictive distributions $f\left(\tilde{Y}_{cs,1}|X_{cs},\tilde{\beta}_{cs,1}\right)$, $f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1},X_{cs},\tilde{\beta}_{cs,1}\right)$, …, $f\left(\tilde{Y}_{cs,P}|\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\ldots,\tilde{Y}_{cs,P-1},X_{cs},\tilde{\beta}_{cs,1}\right)$. 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 $\sigma^2_{cs,1}$, respectively, where $\sigma^2_{cs,1}$ may be drawn from an appropriate posterior predictive distribution $f\left(\tilde{\sigma}^2_{cs,1}|Y_{cs,1},X_{cs},\sigma^2_{cs,1}\right)$, or fixed at the maximum likelihood estimate $\hat{\sigma}^2_{cs,1}$ (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 $N\left[\left(X_{cs},\tilde{Y}_{cs,1}\right)\tilde{\beta}_{cs,2},\sigma^2_{cs,2}\right]$, and so on up to $\tilde{Y}_{cs,P} \sim N\left[\left(X_{cs},\tilde{Y}_{cs,1},\tilde{Y}_{cs,2},\ldots,\tilde{Y}_{cs,P-1}\right)\tilde{\beta}_{cs,P},\sigma^2_{cs,P}\right]$. Alternatively, if the variable under synthesis $Y_{cs,p}$ is binary, then $\tilde{Y}_{cs,p}$ is drawn from a binomial distribution

$Bin[1, \hat{p}\{(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,p-1})\tilde{\beta}_{cs,P}\}]$, where $\hat{p}\{(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,p-1})\tilde{\beta}_{cs,P}\}$ is the predicted probability computed from the inverse-logit of $\{(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,p-1})\tilde{\beta}_{cs,P}\}$. For polytomous variables, the same procedure is used to obtain posterior probabilities for each categorical response, which are used to generate the synthetic values from a multinomial distribution. The iterative simulation process continues until all synthetic variables $(\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \ldots, \tilde{Y}_{cs,P})$ 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 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 small areas. Inferences for a variety of estimands can be obtained using the combining rules in Section 2.2.

4 Application: National Health Interview Survey (Restricted Microdata)

In this section, I demonstrate and evaluate the above procedure on a subset of restricted-use microdata from the 2003-2005 National Health Interview Survey (NHIS). The NHIS is an ongoing, cross-sectional national survey and is the principal source of information on the health of the civilian non-institutionalized population of the United States. NHIS provides annual estimates on a variety of topics, including health status and disability, healthcare access and utilization, and illness and disease. The survey data are used to evaluate various Federal health programs and to track progress toward achieving national health objectives.

The NHIS employs a complex sample design en route to interviewing a sample of civilian non-institutionalized persons living in the United States. The sample consists of a multistage area probability design consisting of 358 PSUs sampled each year (during the 2003-2005 period) drawn from approximately 1,900 geographically defined PSUs that cover the 50 states and the District of Columbia. A PSU consists of a county, small group of contiguous counties, or a metropolitan statistical area. The NHIS sample is drawn from each state and the District of Columbia. Obtaining state-level estimates with acceptable precision for each state is not possible using the annual NHIS files. The National Center for Health Statistics recommends that users combine multiple years of data in order to produce state-level estimates.

The NHIS is a face-to-face survey that collects data on several units of the household, including the household itself, all persons living in the household, families, a sampled adult, and a sampled child. For this application we restrict the data to sampled adults ages 18 and older. A total of 93,606 sampled adults (age 18 and older) completed interviews between the years 2003-2005 (2003: n=30,852; 2004: n=31,326; 2005: n=31,428). Full details of the NHIS methodology can be found elsewhere (Pleis, 2010).

For this application, we define the smallest areas of interest as PSUs (for the sampled areas) and counties (for the nonsampled areas). (The use of counties instead of PSU's for the nonsampled areas was necessary as the population frame of PSUs was not available to us.) Because the NHIS sample design consists of state-level stratification, we use state-level identifiers to complete the nested hierarchy. The state-level covariates include: number of metropolitan areas, number of micropolitan areas, region, and population. No county-level covariates were used in this application. The selected variables for this application (listed in Table 4.0) include two continuous variables: body mass index and age; and five binary variables: smoking status, moderate activity, sex, ever receiving a hypertension diagnosis, and self-reported health status. These variables were selected based on their common usage in analyses of NCHS data and their recommended use for this project by statisticians from the National Center for Health Statistics.

For continuous variables, design-based estimates of regression parameters (Stage 1) were obtained by fitting normal linear models within each PSU 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. All regression models in Stage 1 accounted for the sampling weights via pseudo-maximum likelihood estimation using the R survey package. To ensure the stability of the design-based regression estimates, a minimum sample size rule of 15 times the number of predictors in the model (i.e., $15 \cdot p$) was applied within each PSU. If a PSU did not meet this threshold, then a Mahalanobis distance metric was used to pool "statistically similar" PSUs together, based on the variables listed in Table 4.0, until the sample size threshold was met.

$M = 10$ fully synthetic data sets are generated for each sampled and non-sampled "small area" (i.e., PSU and county, respectively). For each small area a synthetic sample of 500 cases is generated. To help 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 each previous implementation. All estimates based on the observed data are weighted.

Table 4.0. List of NHIS Variables Used in Synthetic Data Application. Variables Shown in the Order of Synthesis.

Variable	Type	Range/Categories	Transformation
Body mass index (BMI)	continuous	9.15	log
Age	continuous	18 - 84	--
Smoker	binary	recoded; yes,no	--
Moderate activity	binary	recoded: yes,no	--
Sex	binary	male,female	--
Hypertension diagnosis	binary	yes,no	--
Self-reported health status	binary	recoded; fair/poor, excellent/very good/good	--

4.1 Validity of Univariate Estimates

Figures 4.1-4.3 contain back-to-back histograms depicting the overall distributions of each NHIS variable variable. The actual distribution is shown in red and the synthetic distribution in blue. All variables are presented on the untransformed scale. Figure 4.1 shows the full synthetic data for sampled and non-sampled areas, Figure 4.2 shows the synthetic data for the sampled areas only, and Figure 4.3 shows the synthetic data for only the non-sampled areas. A few general remarks can be made about these figures. First, the synthetic distributions for the sampled and non-sampled areas reflect the shape of the actual distributions reasonably well in for most variables. Second, the rarer characteristics (hypertension, fair/poor health rating) tend to be slightly overestimated in the synthetic data. This could be due to the fact that "statistically similar" PSUs were combined during the direct estimation stage (Stage 1) to yield more precise estimates, which may lead to overestimation of the characteristic relative to the actual PSU of interest. Third, the synthetic distributions for the non-sampled areas appear to be slightly more variable for the continuous distributions compared to the distributions for the sampled areas. This makes sense as the synthetic data for the non-sampled areas were generated from a purely model-based perspective with no actual data from these areas to inform their synthesization. And fourth, the overall shape of the combined synthetic data distribution (sampled plus nonsampled; Figure 4.1) resembles the synthetic distribution for the sampled areas (Figure 4.2) fairly well. This is a reassuring result as it means that the combined synthetic data file, if released to the public, could be used in lieu of the synthetic data for the sampled areas only to produce valid distributional properties. It also suggests that the sampled and non-sampled synthetic data are indistinguishable for the most part, which suggests that an intruder could have additional difficulty in determining which cases are associated with the sampled areas.

Figure 4.1. Back-to-Back Histograms of Actual (Red) and Combined Synthetic (Blue) Distributions for NHIS Variables.

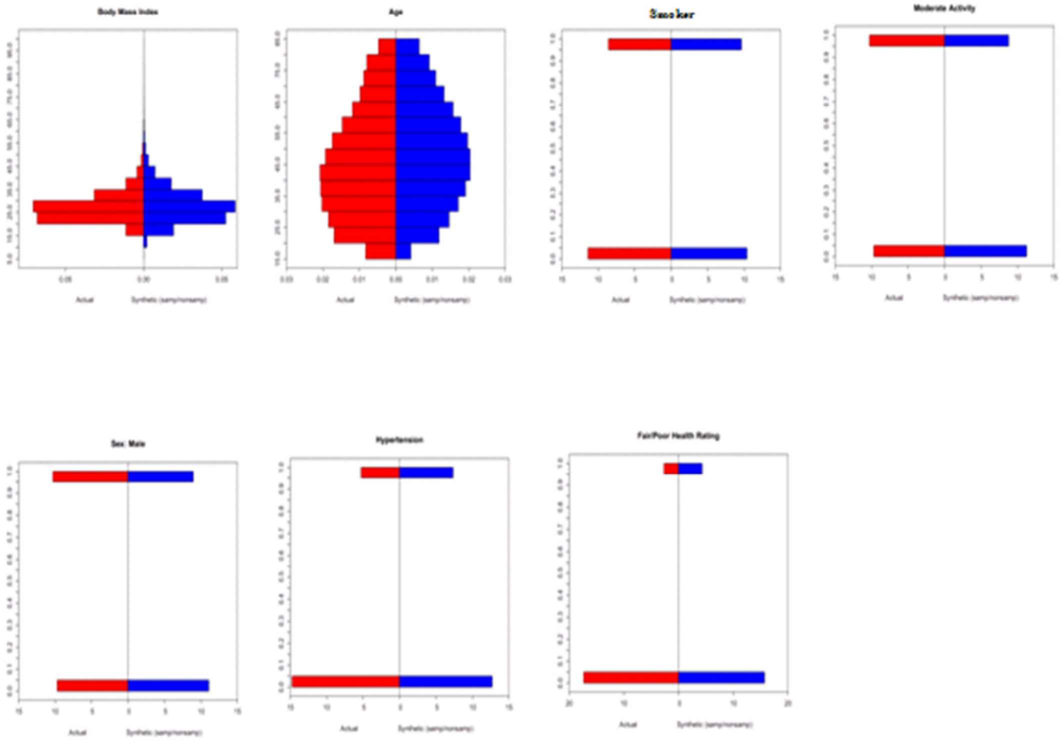


Figure 4.2. Back-to-Back Histograms of Actual (Red) and Sampled Synthetic (Blue) Distributions for NHIS Variables.

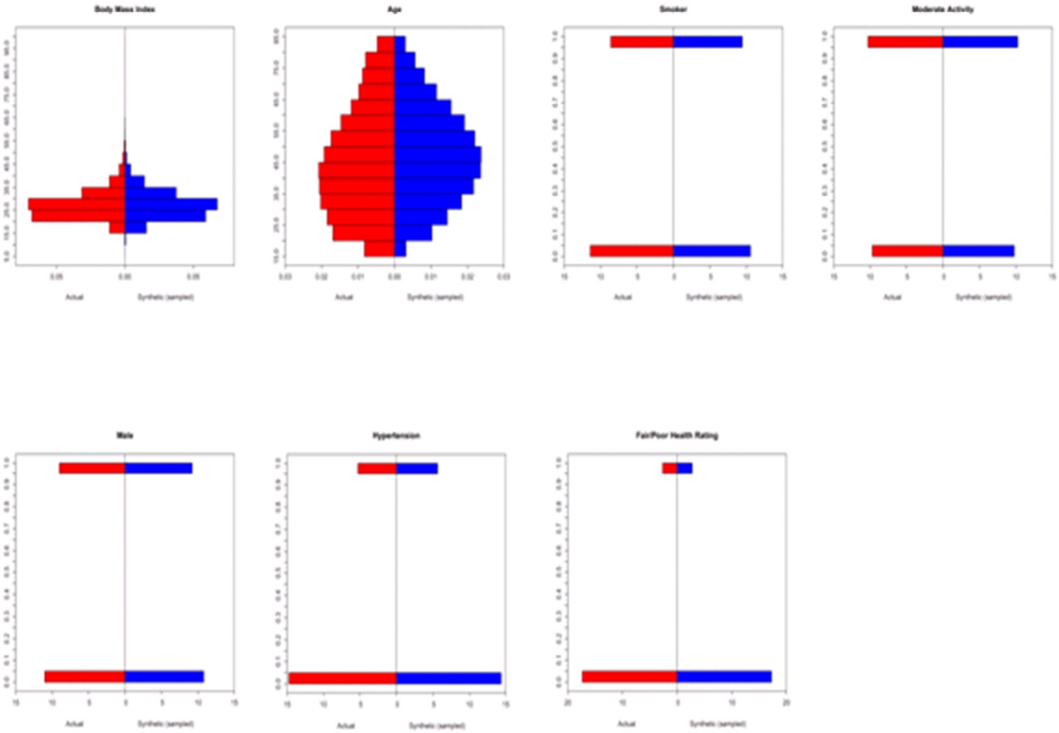


Figure 4.3. Back-to-Back Histograms of Actual (Red) and Nonsampled Synthetic (Blue) Distributions for NHIS Variables.

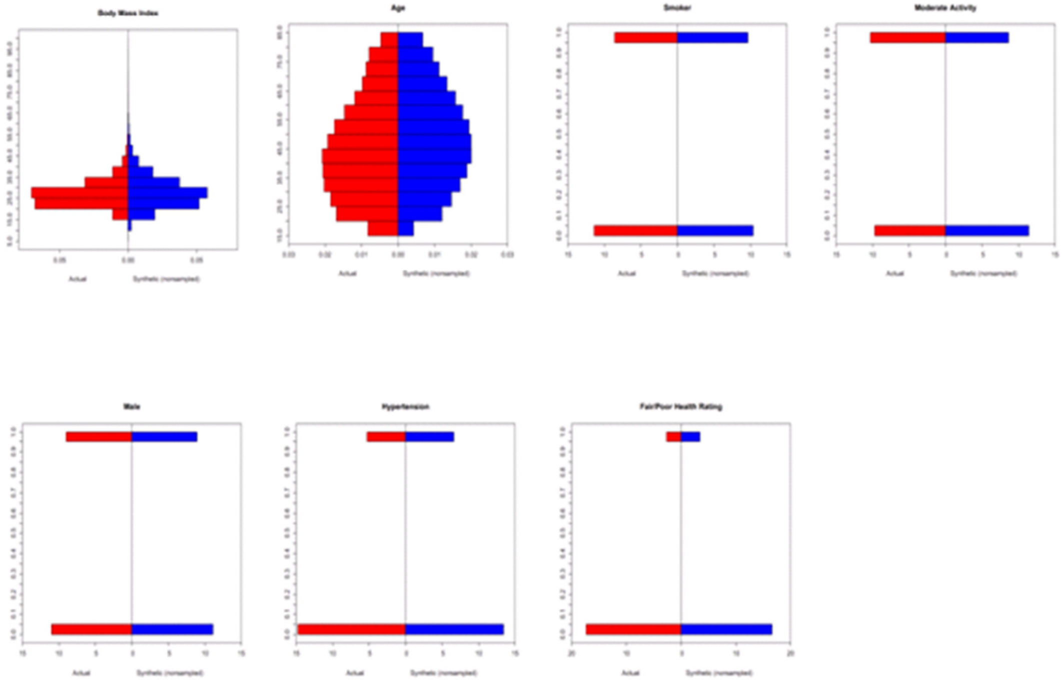


Table 4.1 provides summary measures of actual- and synthetic- means obtained at the PSU-level (or county-level). Columns 2-4 show the average PSU/county mean obtained from the actual data, synthetic (sampled areas), and synthetic (nonsampled areas), respectively. Columns 5-7 show the corresponding average standard errors for the PSU/county means obtained from the synthetic and actual data. The last two columns contain the intercept and slope values obtained from regressing the actual PSU/county means against the corresponding synthetic means. Intercept values close to 0 and slope values close to 1 indicate strong correspondence between the synthetic and actual means.

Table 4.2 contains the same summary measures computed for the stratum-level estimates.

The results are generally positive. For the most part the summary measures of the synthetic estimates correspond well to those of the actual estimates, on average. The nonsampled synthetic standard errors are quite large. In all cases, the average standard errors of the synthetic estimates are larger than the corresponding actual estimates, as expected. The average standard errors for the non-sampled area estimates tends to be about 5-6 times as large as the actual data standard errors, whereas the average standard errors for the sampled area estimates are only slightly larger than the corresponding actual standard errors.

Table 4.1 Average PSU/County Means Obtained from Synthetic and Actual NHIS (2003-2005) Data. Actual estimates are weighted.

	Avg. Means of PSU/County Means			Avg. Standard Errors of PSU/County Means			Regression of Actual and Synthetic PSU Means	
	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Intercept	Slope
BMI	27.29	27.25	27.41	0.45	0.73	1.45	-4.21	1.16
Age	46.60	47.45	47.84	1.35	1.15	4.27	-20.99	1.43
Smoker	46.13	46.43	47.11	3.84	3.65	12.05	-0.07	1.14
Moderate activity	51.54	51.72	51.75	3.62	3.82	21.04	-0.03	1.06
Sex: Male	45.94	45.89	46.24	3.87	3.71	9.72	-0.06	1.14
Hypertension diagnosis	28.22	27.94	29.86	3.45	3.74	10.58	-0.04	1.16
“Fair or Poor” health	13.95	13.14	15.13	2.58	2.16	8.84	0.00	1.04

Table 4.2 Average Stratum/State Means Obtained from Aggregate Synthetic and Actual NHIS (2003-2005) Data. Actual estimates are weighted.

	Avg. Means of Strata Means			Avg. Standard Errors of Strata Means		
	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)
BMI	27.06	27.23	27.27	0.21	0.18	0.96
Age	45.98	47.40	48.49	0.84	0.50	2.85
Smoker	45.04	46.74	50.36	1.85	1.81	8.39
Moderate activity	53.30	52.12	52.11	3.98	1.57	16.90
Sex: Male	46.21	45.88	46.47	1.06	1.70	6.56
Hypertension diagnosis	26.32	27.64	30.50	1.54	1.74	6.90
“Fair or Poor” health	12.48	12.89	15.72	1.18	1.04	6.50

The variability in the synthetic estimates across sampled PSUs is depicted via the scatter plot in Figure 4.4. The synthetic means (y-axis) are contrasted against the actual means (x-axis) for all sampled PSUs. Ideally, each point will lie exactly on the 45-degree line if the synthetic and actual estimates correspond perfectly. In general, most of the points lie about the 45-degree line indicating good correspondence between the synthetic and actual small area estimates. For example, the estimates of the proportion of moderate activity tend to be tightly clustered around the 45-degree line, indicating strong correspondence between the synthetic and actual data. However, some estimates tend to depart from the equalizing line. In particular, the synthetic data tends to overestimate age in PSUs where the average age is low relative to other PSUs. This could be due to the fact that age was simulated from a normal distribution, even though the distribution tends to be slightly right-skewed. A transformation or alternative simulation approach might yield stronger better results. Given the fully-synthetic nature of this application, we would expect the standard errors of the synthetic PSU means to be larger than the actual standard errors. Figure 4.5 shows scatter plots of the synthetic and actual standard errors for each variable of interest. As expected, the synthetic data yield larger standard errors of the means, on average.

Figure 4.4 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PSU Means for NHIS Variables

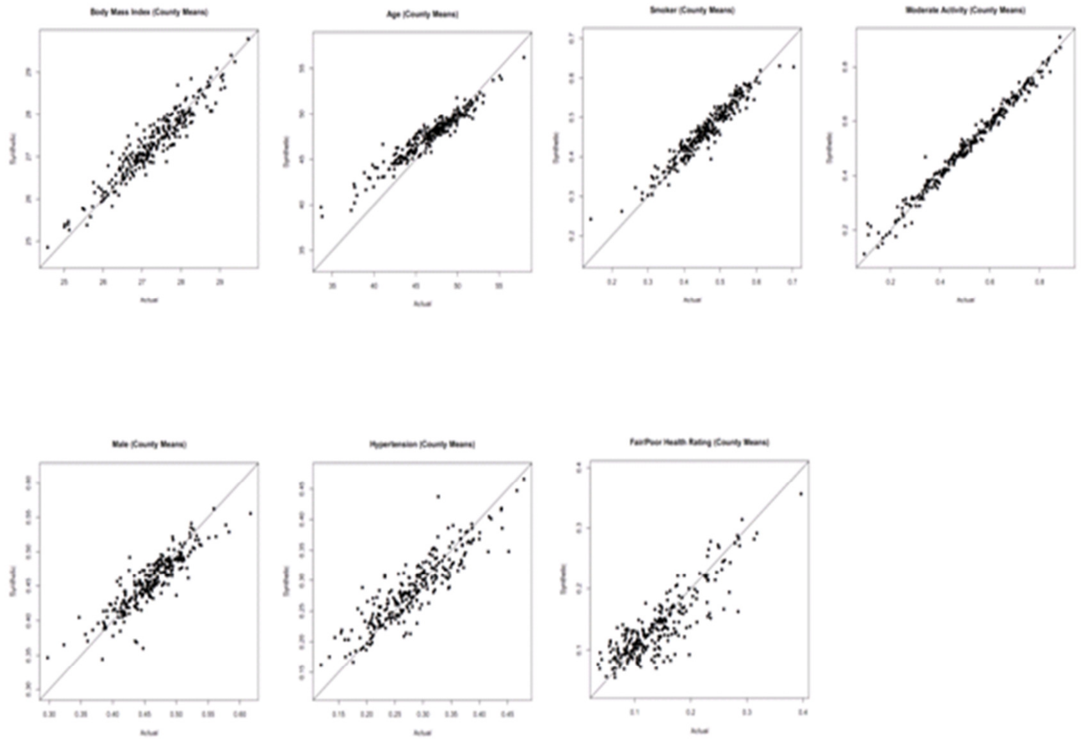
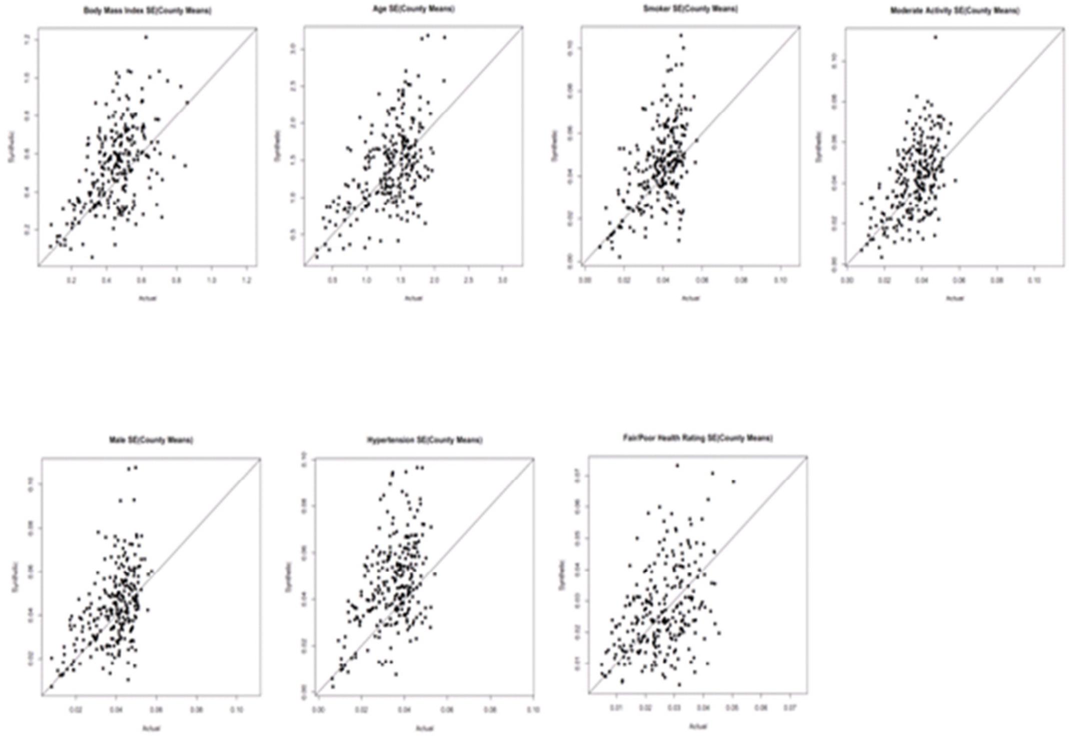


Figure 4.5 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) Standard Errors of PSU Means for NHIS Variables.



4.2 Cross-Validation Study of Non-Sampled Small Area Estimates

Because no observed data exists for the non-sampled areas, it is not possible to directly assess the validity of the small area estimates obtained from these areas. To overcome this limitation we perform a cross-validation study by randomly removing a sampled area from the observed data and treating it as if it were a non-sampled area during the synthetic data generation process. We randomly selected 63 sampled PSUs for this cross-validation study. Each of the selected PSUs was dropped from the observed data one at a time and all three stages of the synthetic data generation process were performed to obtain synthetic data for the dropped (or "unsampled") area.

Figure 4.6 contains scatter plots of estimates for all 63 cross-validated PSUs. The synthetic estimates are displayed on the y-axis and actual estimates on the x-axis. The synthetic point estimates tend to lie about the 45-degree line. This is a reassuring finding. About half of the estimates lie above and below the line, indicating lack of bias for synthetic non-sampled area estimates. However, the precision of the estimates is rather small. The points are dispersed widely indicating large variability in the synthetic estimates. Figure 4.7 shows scatter plots of the standard errors of the PSU means. The fact that nearly all of the points lie above the 45-degree line reaffirms that the synthetic estimates contain a large amount of variability. Given that the synthetic data for these areas was generated strictly from a model-based perspective, it is not surprising that the estimates exhibit a significant amount of variability.

Figure 4.6 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) PSU Means for Cross-Validation Study

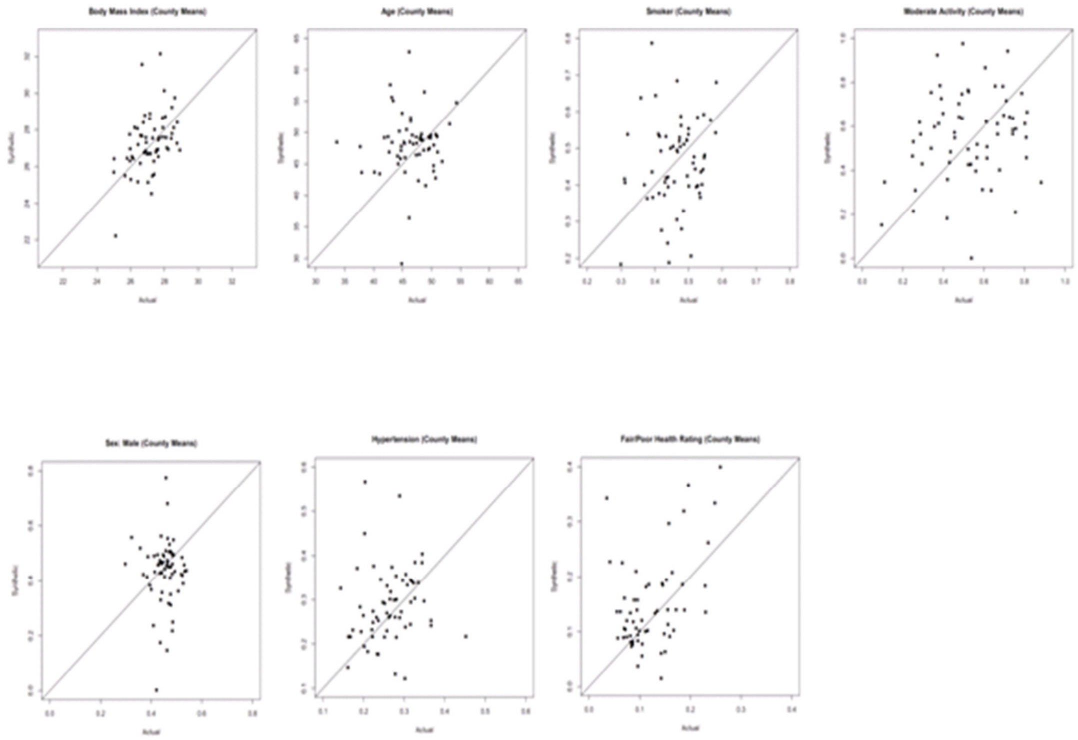
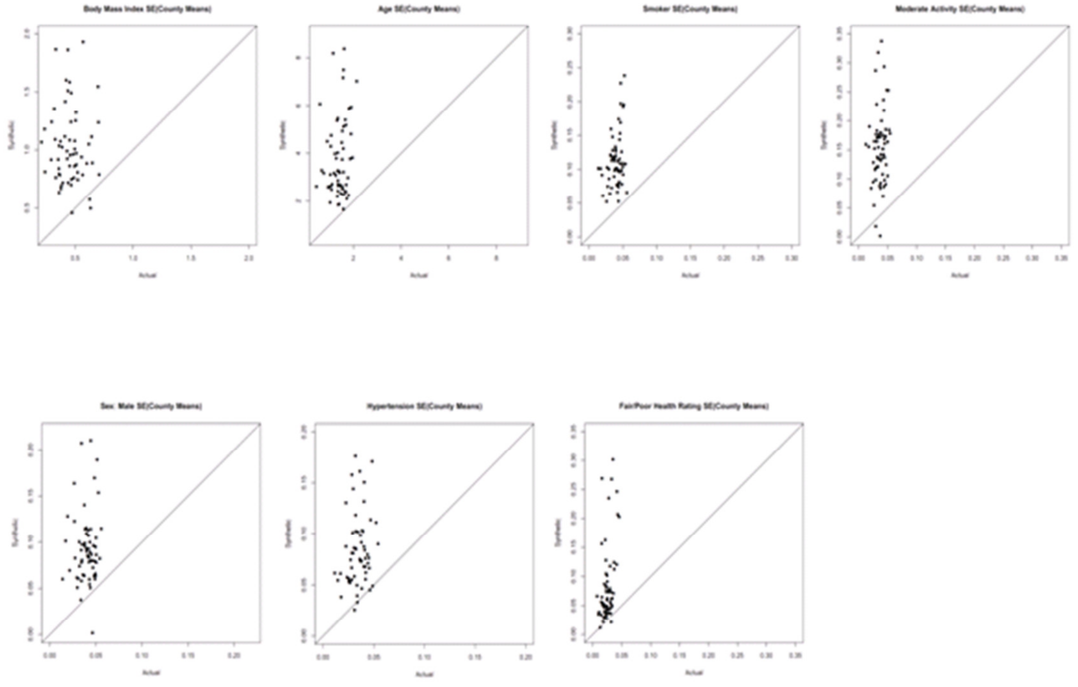


Figure 4.7 Scatter Plot of Synthetic (y-axis) and Actual (x-axis) Standard Errors of PSU Means for Cross-Validation Study



4.3 Validity of Multivariate Estimates

The next set of analyses assesses the analytic validity of synthetic multivariate estimates obtained from two multiple regression models. Tables 4.3 and 4.4 show coefficient estimates (and their standard errors) for two regression models (linear and logistic) fit at the PSU/county- and strata-level, respectively. The dependent variable for the linear model is log(BMI) and for the logistic model is hypertension diagnosis.

The results are reassuring. The synthetic coefficient estimates (for both sampled and nonsampled areas) are quite similar and correspond well with the actual estimates, on average. All synthetic estimates lie within one of their respective standard errors from the corresponding actual estimate, on average. This pattern holds true for both PSU- and stratum-level coefficient estimates. The standard errors of the sampled synthetic point estimates are comparable to the corresponding actual standard errors. The nonsampled synthetic standard errors tend to be between 2-4 times larger than the actual standard errors.

Table 4.3 PSU-Level Linear and Logistic Regression Coefficients and Standard Errors Obtained from Actual and Synthetic Data Sets.

	Avg. Regression Coefficients			Avg. Standard Errors of Regression Coefficients		
Linear regression of BMI (log) on	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)
Intercept	3.26	3.27	3.28	0.05	0.06	0.18
Age	0.10	0.11	0.11	0.04	0.04	0.10
Smoker	-0.00	-0.00	-0.00	0.00	0.00	0.00
Moderate activity	-0.01	-0.02	-0.02	0.03	0.03	0.09
Male	-0.01	-0.01	-0.00	0.03	0.04	0.11
Hypertension	0.04	0.03	0.02	0.03	0.03	0.10
Fair/poor health	0.05	0.06	0.06	0.05	0.04	0.13
	Avg. Regression Coefficients (Odds Ratios)			Avg. Standard Errors of Regression Coefficients		
Logistic regression of Hypertension on	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)
Intercept	-14.77	-14.71	-14.99	3.99	4.01	12.27
BMI	3.12	3.10	3.26	1.13	1.13	3.49
Age	0.07	0.07	0.07	0.14	0.15	0.06
Smoker	0.05	0.08	0.20	0.44	0.43	1.42
Moderate activity	-0.03	-0.06	-0.22	0.46	0.48	1.96
Male	0.04	0.10	0.06	0.43	0.46	1.73
Fair/poor health	0.91	0.86	0.75	0.60	0.48	1.67

Table 4.4 Strata-Level Linear and Logistic Regression Coefficients and Standard Errors Obtained from Actual and Synthetic Data Sets.

	Avg. Regression Coefficients			Avg. Standard Errors of Regression Coefficients		
Linear regression of BMI (log) on	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)
Intercept	3.24	3.26	3.28	0.02	0.04	0.13
Age	0.11	0.11	0.11	0.01	0.02	0.08
Smoker	-0.00	-0.00	-0.00	0.00	0.00	0.00
Moderate activity	-0.01	-0.02	-0.02	0.01	0.02	0.07
Male	-0.01	-0.01	-0.01	0.01	0.02	0.08
Hypertension	0.04	0.04	0.04	0.01	0.02	0.07
Fair/poor health	0.05	0.06	0.06	0.02	0.03	0.09
	Avg. Regression Coefficients (Odds Ratios)			Avg. Standard Errors of Regression Coefficients		
Logistic regression of Hypertension on	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)	Actual	Synthetic (sampled areas)	Synthetic (nonsampled areas)
Intercept	-14.18	-13.71	-10.90	1.00	2.27	6.32
BMI	2.99	2.89	2.34	0.28	0.65	1.75
Age	0.06	0.06	0.05	0.00	0.01	0.02
Smoker	0.10	0.11	0.13	0.10	0.25	0.57
Moderate activity	-0.01	-0.04	-0.06	0.11	0.28	0.69
Male	-0.02	0.07	0.01	0.11	0.28	0.80
Fair/poor health	0.91	0.84	0.74	0.14	0.30	0.64

4.4 Propensity Score Balance

Another indicator of the quality of the synthetic data is to assess the covariate balance between the synthetic and actual data. This is most easily performed using propensity scores (Rubin and Rosenbaum, 1983). Propensity scores are commonly used to identify imbalances in in two or more groups (e.g., treatment and control groups) based on the distribution of a set of observed covariates. Biases caused by covariate imbalances may be adjusted by performing a weighted analysis with weights inversely proportional to the propensity scores (Ekholm and Laaksonen, 1991).

To assess the covariate balance between the synthetic and actual data sets, a randomly selected (sampled) synthetic data set and the actual data are stacked vertically. Then an actual data indicator variable is regressed against all synthetic and actual variables using a logistic regression model. The fitted model is used to obtain estimates of the propensity of a record belonging to the actual data. The propensity scores are then sorted and classified into deciles and the proportions of synthetic and actual records are compared. If the synthetic and actual covariates are fully balanced, then the proportion of synthetic versus actual data should be the same for each decile group. A chi-squared test with 9 degrees of freedom (if deciles are used) can be performed to assess the equivalence of the actual data proportions across the groups.

We use the propensity score balance method to assess the similarity of the synthetic and actual data in each PSU. Table 4.5 shows summary statistics of the estimated probabilities of belonging to the actual data in each PSU, as well as associated test statistics. The overall mean estimated propensity score was 0.30, which reflects the true proportion of actual data in each PSU. Within each PSU, the propensity scores were sorted and grouped into deciles and a chi-square statistic was computed. Small chi-square values indicate that the synthetic and actual data sets are balanced or statistically independent from each other, based on the set of covariates, while large values indicate poor covariate balance between the two data sets. The mean chi-square p-value was 0.23. The average p-value is not statistically significant. This suggests that the synthetic data is statistically balanced with the actual data based on the selected covariates. This is another reassuring finding, indicating strong correspondence of the distributional properties between both the synthetic and actual data sources.

Table 4.5 Estimated Propensities of Belonging to the Observed Data

	Mean	Min	Max
Estimated probabilities $\hat{p}$	0.30	0.18	0.48
χ^2 statistic	14.80	7.92	42.90
P-value	0.23	0.01	0.57

5 NHIS-Based Simulation

This section evaluates the repeated sampling properties for small area inferences drawn from the synthetic data based on a simulation application. In this simulation, the 2003-2005 NHIS data is treated as a population from which subsamples are drawn. 500 stratified random subsamples are drawn from each PSU with replacement. Each subsample accounts for approximately 30% of the total sample in each PSU. Each NHIS subsample is used as the basis for constructing a synthetic population from which 100 synthetic samples are drawn. A total of 50,000 synthetic data sets are generated.

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 most commonly encountered 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 NHIS subsamples. For each set of 10 synthetic samples, a synthetic estimate and associated confidence interval is obtained for each variable in each PSU using the combining rule equations [1] and [2] in Section 2.2. To obtain unconditional inferences, 100 sets of 10 synthetic samples are randomly selected with replacement *across* each of the 100 NHIS subsamples and estimates are obtained again 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 $\left[L_{\hat{q}_M,syn}, U_{\hat{q}_M,syn}\right]$ 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. $Q_{CIC} = 1$ 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 by,

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

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 from the 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 the synthetic interval completely covers the actual interval. Calculating the confidence interval overlap is only possible for conditional, not unconditional, 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

Tables 4.6 and 4.7 show the average confidence interval coverage (CIC) and confidence interval overlap (CIO) for means obtained from sampled PSUs and strata, respectively. The conditional CIC is quite high for the PSU-level estimates ranging from 0.91-0.99. The stratum-level conditional CIC values are high and range from 0.94-0.99. All of the unconditional CIC values correspond closely to their true CIC values. All of these results indicate that the repeated sampling properties of the synthetic data method perform well when applied to the sampled PSU and stratum areas.

Table 4.6 Simulation-Based Confidence Interval Results for Sampled PSU-Level Means.

Sampled PSUs	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
BMI	0.99	0.99	0.99	0.97
Age	0.91	0.92	0.99	0.98
Smoker	0.99	0.98	0.99	0.98
Moderate activity	0.99	0.99	0.99	0.98
Male	0.99	0.98	0.99	0.98
Hypertension	0.99	0.97	0.99	0.97
Fair/poor health status	0.99	0.92	0.99	0.97

Table 4.7 Simulation-Based Confidence Interval Results for Sampled Stratum-Level Means.

Sampled Strata	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
BMI	0.99	0.95	0.98	0.98
Age	0.96	0.88	0.96	0.98
Smoker	0.99	0.94	0.97	0.99
Moderate activity	0.99	0.94	0.96	0.98
Male	0.99	0.95	0.98	0.98
Hypertension	0.98	0.92	0.97	0.98
Fair/poor health status	0.94	0.88	0.95	0.98

Tables 4.8 and 4.9 show the average confidence interval coverage (CIC) and confidence interval overlap (CIO) for means obtained from the nonsampled counties and strata, respectively. For the nonsampled counties, all CIC and CIO values equal 0.99 and correspond perfectly with the actual CIC values (also equal to 0.99). These results suggest that the estimated means for the nonsampled counties tend to be valid from a repeated sampling perspective. With regard to the nonsampled stratum-level estimates, the CIC and CIO values are high, but not quite as high as the county-level values. The range of conditional CIC values is 0.89-0.99, with the lowest value corresponding to the "fair or poor" health status variable. However, in general, the confidence interval coverage and overlap is good for both sampled and nonsampled areas.

Table 4.8 Simulation-Based Confidence Interval Results for Nonsampled County-Level Means.

Nonsampled Counties	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
BMI	0.99	0.99	0.99	0.99
Age	0.99	0.99	0.99	0.99
Smoker	0.99	0.99	0.99	0.99
Moderate activity	0.99	0.99	0.99	0.99
Male	0.99	0.99	0.99	0.99
Hypertension	0.99	0.99	0.99	0.99
Fair/poor health status	0.99	0.99	0.99	0.99

Table 4.9 Simulation-Based Confidence Interval Results for Nonsampled Stratum-Level Means.

Nonsampled Strata	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
BMI	0.97	0.97	0.98	0.99
Age	0.98	0.97	0.99	0.99
Smoker	0.98	0.97	0.99	0.99
Moderate activity	0.98	0.97	0.99	0.99
Male	0.99	0.99	0.99	0.96
Hypertension	0.94	0.92	0.99	0.99
Fair/poor health status	0.89	0.88	0.90	0.99

5.2 Validity of Multivariate Estimates

Multivariate simulation results are shown in Tables 4.10 and 4.11 for sampled and nonsampled areas, respectively. This table shows average CIC and CIO values for regression coefficient estimates obtained within each PSU (or county) and stratum. For the sampled PSUs and strata (Table 4.10), the conditional CIC and CIO values are high and range from 0.98-0.99 and 0.94-0.99, respectively, indicating good analytic validity for these multivariate estimands with PSUs and strata. The unconditional CIC values equal 0.99, which either meets or exceeds the true CIC values obtained from the actual data. For the nonsampled counties and strata, the confidence interval coverage and overlap is similarly high for all coefficient estimates, ranging from 0.98-0.99. The simulation evidence suggests that the synthetic data method produces estimates that are valid from a repeated sampling perspective.

Table 4.10 Simulation-Based Confidence Interval Results for Sampled PSU- and Stratum-Level Regression Coefficients

Sampled; PSUs	Conditional Inference		Unconditional Inference	
Covariates	CIC	CIO	CIC	CIC (Actual)
Regression of BMI(log) on				
Intercept	0.99	0.98	0.99	0.97
Age	0.99	0.98	0.99	0.97
Smoker	0.99	0.98	0.99	0.98
Moderate activity	0.99	0.98	0.99	0.97
Male	0.99	0.98	0.99	0.98
Hypertension	0.99	0.99	0.99	0.98
Fair/poor health	0.99	0.94	0.99	0.96
Sampled; Strata	Conditional Inference		Unconditional Inference	
Covariates	CIC	CIO	CIC	CIC (Actual)
Regression of Hypertension on				
Intercept	0.99	0.99	0.99	0.97
BMI	0.99	0.98	0.99	0.97
Age	0.99	0.98	0.99	0.97
Smoker	0.99	0.99	0.99	0.98
Moderate activity	0.99	0.98	0.99	0.98
Male	0.99	0.99	0.99	0.98
Fair/poor health	0.98	0.95	0.99	0.97

Table 4.11 Simulation-Based Confidence Interval Results for Nonsampled County- and Stratum-Level Regression Coefficients

Nonsampled; Counties	Conditional Inference		Unconditional Inference	
Covariates	CIC	CIO	CIC	CIC (Actual)
Regression of BMI(log) on				
Intercept	0.99	0.99	0.99	0.99
Age	0.99	0.99	0.99	0.99
Smoker	0.99	0.99	0.99	0.99
Moderate activity	0.99	0.99	0.99	0.99
Male	0.99	0.99	0.99	0.99
Hypertension	0.99	0.99	0.99	0.99
Fair/poor health	0.99	0.98	0.99	0.99
Nonsampled; Strata	Conditional Inference		Unconditional Inference	
Covariates	CIC	CIO	CIC	CIC (Actual)
Regression of Hypertension on				
Intercept	0.99	0.99	0.99	0.99
BMI	0.99	0.99	0.99	0.99
Age	0.99	0.99	0.99	0.99
Smoker	0.99	0.99	0.99	0.99
Moderate activity	0.99	0.99	0.99	0.99
Male	0.99	0.99	0.99	0.99
Fair/poor health	0.99	0.99	0.99	0.99

6 Conclusions

In this chapter, I demonstrated a synthetic data methodology that produces microdata for small geographic areas. The method accounts for the complex sample design by incorporating the clustering and stratification identifiers into a Bayesian hierarchical model. The sampling weights are incorporated into the model at the design stage (Stage 1). We evaluated the method using restricted county-level data from the National Health Interview Survey. Based on analytic and simulation studies, the analytic validity of the synthetic small area estimates (univariate and multivariate) is high due to their strong correspondence with the actual estimates. Aggregating the PSU-level microdata to the stratum-level also yields similarly high validity for large-area estimates.

An intriguing feature of the method is the ability to use the model to generate synthetic data for nonsampled small areas (e.g., PSUs, counties). This feature increases the utility of the survey data, as it allows data users (e.g., students, community planners, local organizations) to generate estimates for small areas that may be more relevant to them. It also provides a bit of confidentiality protection as an intruder may have difficulty determining which small areas in the combined synthetic data set were part of the actual sampled. Based on cross-validation and simulation studies, the nonsampled area estimates are valid and comparable to the actual estimates; however, they do tend to possess a large amount of variability because the nonsampled area synthetic data is generated from a completely model-based procedure.

This study has limitations that should be mentioned. This study demonstrated the method on a basic set of continuous and binary variables. Other variable types, such as count, multinomial, and semi-continuous should also be considered as they form the basis for many important variables that are collected in survey data. Another limitation is that the method is based on a fully-parametric framework. Thus, any variable that does not follow a standard distribution (e.g., skewed, bimodal) must be transformed, or else the synthetic method must be modified to handle these non-standard variable types. This limitation was evident for the age variable, which was slightly right-skewed, and resulted in PSU-level age estimates that were slightly overestimated.

The validity of the synthetic data estimates could potentially be improved by adding small area-level covariates to the hierarchical imputation model. In preliminary runs, it was decided not to incorporate PSU/county-level covariates into the model due to the lack of covariate balance between sampled and nonsampled areas based on the set of chosen covariates. Three county-level covariates were tested, including poverty rate, median household income, and population size. The distributions among sampled and nonsampled areas were quite similar, but median household income and population size exhibited distributional differences between sampled and nonsampled areas. When the latter two variables were incorporated into the hierarchical model, the resulting synthetic data estimates conflicted substantially between sampled and nonsampled areas. The sampled estimates were similar to the actual estimates, but the nonsampled estimates showed serious departures from the actual and synthetic estimates for sampled areas. Given that the nonsampled estimates are based on synthetic data that are generated from a purely model-driven framework, we concluded that the nonsampled estimates were highly sensitive to the choice of county-level covariates and could conflict with the sampled estimates if there is little overlap in the county-level covariates between sampled and nonsampled areas. Therefore, a broader choice of area-level covariates and careful examination of distributional differences among those covariates would be wise if implementing this method in practice.

In addition to nonparametric methods and complex variable types, future research may consider how to extend the proposed method to handle additional levels of geography. Prospective data users may be interested in analyzing data for sub-county areas (cities/towns, districts, neighborhoods). The hierarchical Bayesian framework allows for several levels of geography to be incorporated into the model, but with each new level brings additional computational complexity. A nice feature about the model considered here is that the method is easily implemented and does not require complex

MCMC routines. Incorporating additional levels of geography may be beneficial from a data utility perspective, but might also reduce the simplicity of the method.

Despite the potential for future improvements, the method is promising and could easily be adopted by large-scale survey projects, including the National Health Interview Survey, to release more geographically-relevant data to the public. Such efforts could potentially help meet the growing demand for microdata in small geographic areas.

Appendix 1 EM Algorithm for Estimating Stratum-Level Bayesian Hyperparameters

The EM algorithm is used to estimate the unknown population parameters $\beta_{s,p}$and $\Sigma_{s,p}$from the following setup,

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

$$\beta_{cs,p} \sim MVN\left(\beta_{s,p} Z_{cs}, \Sigma_{s,p}\right)$$

where $p = (1,2, \ldots, P)$ 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,

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

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

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

$$\hat{\beta}_{s,p} = \left[\sum\nolimits_{c=1}^{C_s}\left(\beta^{*}_{cs,p}Z_{cs}\right)\right]\left[\sum\nolimits_{c=1}^{C_s}\left(Z_{cs}Z^{T}_{cs}\right)\right]^{-1} \text{, and}$$

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

After convergence the maximum likelihood estimates are incorporated into the posterior distribution of $\beta_{cs,p}$ shown in equation [7].

Appendix 2 EM Algorithm for Estimating Overall Bayesian Hyperparameters

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

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

$$\beta_{s,p} \sim MVN\left(\beta_p Z_s, \Omega_p\right)$$

where $p = (1,2,\ldots,P)$ 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 hyperparameter estimates $\hat{\beta}_s$ and $\hat{V}_s$ were obtained via the EM algorithm.

The E step consists of solving the following expectations,

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

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

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

$$\hat{\beta}_p = \beta^*_{s,p} Z_s (Z_s Z_s^T)^{-1} \text{ , and}$$

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

After convergence the maximum likelihood estimates are incorporated into the posterior distribution of $\beta_{s,p}$ shown in equation [8].

Chapter 5

Conclusions and Discussion

1 Summary of Dissertation

1.1 Chapter Overview

Statistical agencies are constantly participating in a push-and-pull match between their respondents whose confidentiality they are sworn to protect and from consumers of their data who are demanding greater access to more detailed geographical information in public-use data sets. Agencies attempt to increase the utility of their data by providing data users with options related to accessing restricted geographical information (e.g., RDC), but many of these options are too restrictive for some users and therefore may reduce the utility of the collected data. At a time when survey budgets are either stagnant or in decline, it is critically important that agencies demonstrate the usefulness and and promote the utility of their data in order to maintain or increase their funding levels.

This dissertation addresses the data confidentiality and data utility dilemma by utilizing the synthetic data framework and extending it for the purpose of producing generating public-use microdata sets for small geographic areas. Under this framework, it is possible to release fully-synthetic datasets for each small area of interest. The released synthetic data contain no observed values and therefore data confidentiality is preserved. From a data utility perspective, valid small area inferences can be obtained for a variety of descriptive and analytic statistics, meeting the needs of the majority of data users. Although the proposed synthetic data framework may not eliminate the need for research data centers, it offers potential data users a less burdensome option for beginning their analysis.

In this dissertation, I develop three separate methods of generating synthetic data specifically for small area estimation. Each method is designed to handle specific practical issues that may facilitate the use and acceptance of synthetic data in the public domain. Chapter 2 develops a parametric framework for generating synthetic data for small geographic areas. The method uses a hierarchical Bayesian model to account for the multi-level structure of the data. The procedure may be considered a hierarchical extension to the sequential regression multiple imputation framework proposed by Raghunathan et al., (2001). The method is easily implemented and can handle a variety of variable types and parametric distributions. The method was evaluated using public-use and restricted data obtained from the American Community Survey. For both data sources, the small area (i.e., PUMA, county) estimates obtained from the generated synthetic data yielded high analytic validity for basic univariate and multivariate estimands. This finding is reassuring to statistical agencies and potential data users, and lends support to the external validity of the method.

Chapter 3 extends the basic framework proposed in Chapter 2 by implementing a nonparametric simulation procedure for continuous variables. Specifically, the procedure replaces the parametric simulation procedure described in Chapter 2 by generating synthetic values that are a function of the predicted values (based on the hierarchical model) and deviations of predicted and actual values computed within each small area. An approximate Bayesian Bootstrap procedure is used to draw the deviations used in the simulation procedure. Evaluation results indicate that the nonparametric simulation method can produce more valid small area estimates than the parametric synthetic data approach for univariate and multivariate statistics obtained from right-skewed and bimodal distributions. Moreover, the nonparametric extension does not require any attempt to transform the data to normality prior to synthesis. In fact, in many cases the nonparametric method yields more valid estimates when applied to nontransformed variables, than when applied to transformed variables.

The third study, discussed in Chapter 4, extends the framework of Chapter 2 to handle complex sample designs. Specifically, the hierarchical model explicitly accounts for clustering and stratification. Auxiliary information collected at the PSU- and stratum-levels can be incorporated into the model to "borrow strength" across related areas and increase precision. A nice feature of the method is that it can produce synthetic data for nonsampled small areas, greatly increasing the utility of the synthetic data. A pleasant byproduct of this feature is that it assists in masking the sampled areas; thus, making it potentially more difficult for an intruder to distinguish between sampled and nonsampled areas in the synthetic data. A successful practical implementation of this method was applied on the National Health Interview Survey, a large complex ongoing cross-sectional survey. Valid inferences and high confidence interval coverage were obtained for several descriptive and analytic statistics, for both sampled and nonsampled small areas. Although the nonsampled synthetic data yielded less precise estimates than those obtained from the sampled synthetic data, the nonsampled estimates are still valid and unbiased as demonstrated in a cross-validation study.

1.2 Future Research

There are several extensions to the methods developed in this dissertation that one could pursue. For example, we did not quantify the disclosure risk associated with the synthetic data generation methods. Rather the main focus of our evaluations was on the analytic validity of the resulting estimates. Because we adopted a fully-synthetic framework, we argue that the resulting synthetic data can no longer be interpreted as having originated from a given individual, which leads to no grounds for evaluating the risk of being re-identified. Quantifying disclosure risks in fully-synthetic data is a topic that has been virtually untouched in the literature, and seems worthwhile for future work.

Although we only considered fully-synthetic applications, the method can easily be extended to partially-synthetic applications, where only a small subset of variables (or records) are synthesized. This appears to be the most common use of synthetic data in real-world applications (Rodriguez, 2007; Abowd et al., 2006). However, these applications have only considered single-level data sources where obtaining small area inferences was not the primary focus.

Another extension to the hierarchical synthetic data generation approach considered here is a two-stage approach that handles both item missing data and full synthesization in a systematic fashion. Reiter (2004) developed a similar procedure in a non-small area context for partially synthetic data applications. Developing an extended approach specifically for small geographic areas in a fully-synthetic data context is feasible.

As we learned in the evaluation studies presented in this dissertation, the standard errors of the synthetic small area estimates tend to be larger than the corresponding standard errors obtained from the actual data. From a data user's perspective this is an undesirable characteristic as the less precise data may decrease the signal to noise ratio and hide systematic effects that may exist in the observed data set. The goal of the imputer should therefore be to generate synthetic data that is highly efficient and precise. Incorporating strong area-level auxiliary information into the imputation model is one approach to this end. This could be achieved by incorporating administrative data, census data, or survey data into the imputation model. For example, small area estimates produced by Federal statistical programs (e.g., SAIPE) could potentially be incorporated as area-level covariates in the imputation model to improve the efficiency of the synthetic small area estimates. Incorporating such information into the synthesis model offers additional protection against model failure. Model failure can also be improved by considering other modeling approaches that specifically address bimodal distributions, or skewness, more effectively than the models used here. Adapting mixture models to the imputation process may yield improvements for these types of distributions.

The proposed methods may also be extended to handle more complex distributional forms, including multinomial, poisson, and semi-continuous variables. These variables are highly prevalent in practice. Although we applied the methods to polytomous and count variables, we simply used a series of logistic regressions and Gaussian distributions, respectively, to handle these variable types. Extending the hierarchical synthetic data method to explicitly handle Poisson and multinomial distributions is an area for future work.

Lastly, it is possible to extend the hierarchical Bayesian model to incorporate additional levels of geography. In this project, we only considered PSUs, PUMAs, Counties, States, and Strata. Prospective data users may be interested in analyzing data for sub-county areas (cities/towns, districts, neighborhoods). The hierarchical Bayesian framework allows for several levels of geography to be incorporated into the model, but with each new level brings additional computational complexity. A nice feature of the model considered here is that the method is easily implemented and does not require complex MCMC routines. Incorporating additional levels of geography may be beneficial from a data utility perspective, but might also reduce the simplicity of the method.

References

Abowd, J.M., Stinson, M., and 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.

Auchincloss, A.H., Roux, A.V., Brown, D. Erdmann, C.A., and Bertoni, A.G. (2008). "Neighborhood Resources for Physical Activity and Healthy Foods and their Association with Insulin Resistance." *Epidemiology*, 19(1), 146-157.

Bell, W., Basel, W., Cruse, C., Dalzell, L., Maples, J., O'Hara, B., and Powers, D. (2007). "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties." Technical Report, U.S. Bureau of the Census. http://www.census.gov/did/www/saipe/publications/files/report.pdf

Caiola, G., and Reiter, J.P. (2010). "Random Forests for Generating Partially Synthetic, Categorical Data." *Transactions on Data Privacy*, 3(1), 27-42.

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

Datta, G.S., Fay, R.E., and 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., and Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B*, 39(1), 1-38.

Diez Roux, A.V. (2004). "Estimating Neighborhood Health Effects: The Challenges of Causal Inference in a Complex World." *Social Science and Medicine*, 58(10, 1953-60.

Drechsler, J., Bender, S., and Raessler, S. (2008). "Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel." *Transactions on Data Privacy*, 105-130.

Drechsler, J., and Reiter, J.P. (in press) "An Empirical Evaluation of Easily Implemented, Nonparametric Methods for Generating Synthetic Data Sets." *Computational Statistics and Data Analysis.*

Ekholm, A., and Laaksonen, S. (1991). "Weighting via Response Modeling in the Finnish Household Budget Survey." *Journal of Official Statistics*, 7, 325-337.

Fay, R.E., and 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(366), 269-277.

Fisher, K.J., Li, M.Y., and Cleveland, M. (2004). "Neighborhood-Level Influences on Physical Activity Among Older Adults: A Multilevel Analysis." *Journal of Aging and Physical Activity*, 12(1), 45-63.

Fisher, R., and Turner, J. (2004). "Small Area Estimation of Health Insurance Coverage from the Current Population Survey's Social and Economic Supplement and the Survey of Income and Program Participation" *Presented at the American Statistical Association Meetings,* Toronto, Canada.

Gelfand, A.E., and Smith, A.F.M. (1990). "Sampling Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85, 398-409.

Geman, S., and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

He, Y., and Raghunathan, T.E. (2006). "Tukey's gh Distribution for Multiple Imputation." *The American Statistician*, 60(3), 251-256.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., and 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 W. Alvey and B. Jamerson, eds., Record Linkage Techniques, 248-267. Washington D.C.: National Academy Press.

Kinney, S.K. and Reiter, J.P. (2008). "Making Public Use, Synthetic Files of the Longitudinal Business Database." *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference Proceedings*, Istanbul, Turkey.

Lindley, D.V., and Smith, A.F.M. (1972). "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society, Series B*, 34(1), 1-41.

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

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

Liu, F. and Little, R.J.A. (2002). "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata." In *ASA Proceedings of the Joint Statistical Meetings*, 2133-2138.

Malec, D., Sedranks, J., Moriarity, C.L., and Leclere, F.B. (1997). "Small Area Inference for Binary Variables in the National Health Interview Survey." *Journal of the American Statistical Association*, 92(439), 815-826.

Meng, X.L. (1994). "Multiple Imputation Inference with Uncongenial Sources of Input (with discussion)." *Statistical Science*, 9, 538-573.

Mujahid, M.S., Diez Roux, A.V., Morenoff, J.D., Raghunathan, T.E., Cooper, R.S., Ni, H., Shea, S. (2008). "Neighborhood Characteristics and Hypertension." *Epidemiology*, 19(4), 590-598.

Platek, R., Rao, J.N.K., Saerndal, C.E., and Singh, M.P. (1987). *Small Area Statistics*. Wiley, New York.

Pleis, J.R., Ward, B.W., and Lucas, J.W. (2010). "Summary Health Statistics for U.S. Adults: National Health Interview Survey, 2009. National Center for Health Statistics. *Vital Health Statistics*, 10 (249).

Raghunathan, T.E., and 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., and Solenberger, P. (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology*, 27(1), 85-95.

Raghunathan, T.E,. Reiter, J.P., and Rubin, D.B. (2003). "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E. (2008). "Diagnostic Tools for Assessing the Validity of Synthetic Data Inferences." Unpublished manuscript.

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.

Reiss, S.P. (1984). "Practical Data Swapping: The First Steps." *ACM Transactions on Database Systems*, 9, 20-37.

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). "Using CART to Generate Partially Synthetic Public Use Microdata." *Journal of Official Statistics*, 21, 441-462.

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,. and 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., and Raghunathan, T.E. (2007). "The Multiple Adaptations of Multiple Imputation." *Journal of the American Statistical Association*, 102, 1462-1471.

Reiter, J.P. (2009). "Using Multiple Imputation to Integrate and Disseminate Confidential Microdata." *International Statistical Review*, 77, 179-195.

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

Rosenbaum, P.R., and Rubin, D.B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70, 41-55.

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., and Schenker, N. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association*, 81, 366-74.

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., and Gardiner, C. (2005). “The Case for Small Area Microdata.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1), 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

U.S. Census Bureau (2011). American FactFinder. http://factfinder.census.gov/

U.S. Department of Health and Human Services (USDHHS; 2010). “Healthy People 2020: Topics and Objectives.” http://www.healthypeople.gov/2020/

Winkler, W. (2004). “Re-identification Methods for Masked Microdata.” In *Privacy in Statistical Databases.*, Eds. J. Domingo-Ferrer and V. Torra, 216-230.

Woodcock, S.D., and Benedetto, G. (2009). “Distribution-Preserving Statistical Disclosure Limitation.” *Computational Statistics and Data Analysis*, 53, 4228-4242.

Yucel, R.M. (2008). “Multiple Imputation Inference for Multivariate Multilevel Continuous Data with Ignorable Non-Response.” *Philosophical Transactions of the Royal Society A*, 366(1874), 2389-2403.