**WP. 30**
ENGLISH ONLY

**UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION**

**STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**
(Tarragona, Spain, 26-28 October 2011)

Topic (v): Privacy for new types of microdata: sequence data and mobility data

# Synthetic Data for Small Area Estimation in the U.S. Federal Statistical System

Prepared by Joseph W. Sakshaug, University of Michigan, U.S.A.

# Synthetic Data for Small Area Estimation in the U.S. Federal Statistical System

Joseph W. Sakshaug*

* Program in Survey Methodology, University of Michigan, Ann Arbor, MI 48104, USA, joesaks@umich.edu

**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 data from small areas but are prevented from releasing detailed geographical identifiers in public-use data sets due to disclosure concerns. Alternative data dissemination methods used in practice include releasing summary/aggregate tables, suppressing detailed geographic information in public-use data sets, and accessing restricted data via Research Data Centers. This research considers a new method for disseminating public-use microdata that contains more geographical details than are currently being released. Specifically, the method replaces the observed survey values with imputed, or synthetic, values generated from a posterior predictive distribution. A hierarchical Bayesian model is used to preserve the small area inferences and simulate the synthetic data. Confidentiality protection is enhanced because no actual values are released. I demonstrate the method using data from a prominent federal statistical survey: the 2005-2009 American Community Survey. The analytic validity of the synthetic data is assessed by comparing the synthetic small area estimates to those obtained from the actual data.

## 1 Introduction

Demands for greater access to microdata for counties, municipalities, neighborhoods, and other small geographic areas is ever increasing (Tranmer et al., 2005). Analysts require such data to answer important research questions that affect policy decisions at local levels. Statistical agencies regularly collect data from small areas, but are prevented from releasing detailed geographic identifiers due to the risk of disclosing respondent identities and potentially sensitive information.

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

This article investigates a fourth approach that statistical agencies may implement to release more detailed geographical information in public-use data sets. The approach builds on the statistical disclosure control method, originally proposed by Rubin (1993), in which multiple synthetic populations (conditional on the observed data) are generated and samples drawn from each synthetic population, which

comprise the public-use data files, are released. Valid inferences on a variety of estimands are obtained by analyzing each data file separately and combining the results using methods described in 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 geographical identifiers are almost always suppressed in public-use microdata sets. Significant theoretical and practical research on model-based small area estimation has led to a greater understanding of how small area data can be summarized (and potentially simulated) by statistical models (Platek et al., 1987; Rao, 2003). The current study utilizes a Bayesian hierarchical model to "borrow strength" across related areas and to increase the efficiency of the resulting small-area estimates. The use of Bayesian hierarchical models for multilevel imputation, and, particularly, for synthetic data applications, is sparse (Yucel, 2008; Reiter, Raghunathan, and Kinney, 2006; Yu, 2008).

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 misspecified 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 as inferences are less sensitive to the specification of the imputation model, but it may not provide the same level of disclosure protection relative to fully-synthetic data because the observed sample units and/or the majority of their data values are released to the public (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. 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. Hence, maintaining high levels of 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 creating fully-synthetic data sets is that an arbitrarily large sample size may be drawn from the synthetic population, facilitating analysis for data users who would otherwise have to exclude or apply complicated indirect estimation procedures to areas with sparse (or nil) sample sizes. Synthetic sample sizes may be deliberately chosen to facilitate the use of direct estimation

methods and routine statistical procedures, easing the burden of analysis for data users.

In this article, I propose an extension to existing synthetic data procedures for the purpose of creating synthetic, public-use microdata sets for small geographic areas from which valid small area inferences may be obtained. A Bayesian hierarchical model is developed that accounts for the hierarchical structure of the geographical areas and "borrows strength" across related geographic areas. A sequential multivariate regression procedure is used to approximate the joint distribution of the observed data and to simulate synthetic values from the resulting posterior predictive distribution (Raghunathan et al., 2001). I demonstrate how statistical agencies may generate fully-synthetic data for small geographic areas on a subset of data from the U.S. American Community Survey. Synthetic data is generated for several commonly used household- and person-level variables and their analytic validity is evaluated by comparing small area inferences obtained from the synthetic data with those obtained from the observed data. I do not empirically evaluate the disclosure risk properties of the proposed synthetic data approach and leave this to future work.

## 2 Review of Fully Synthetic Data

The general framework for creating and analyzing fully synthetic data sets is described in (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 $X = (X_i; i = 1, 2, \ldots, N)$ representing the design or geographical variables available on all *N* units in the population, and $Y = (Y_i; i = 1, 2, \ldots, N)$ representing the survey variables of interest, observed only for the sampled units. Let $Y_{obs} = (Y_i; i = 1, 2, \ldots, n)$ be the observed portion of *Y* corresponding to sampled units and $Y_{mis} = (Y_i, i = n+1, n+2, \ldots, N)$ be the unobserved portion of *Y* corresponding to the nonsampled units. The observed data set is $D = [X, Y_{obs}]$. For simplicity, 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 based on the observed data *D* in two steps. First, multiple synthetic populations are generated by simulating $\left(Y_{mis}^{(i)}; i = 1, 2, \ldots M\right)$ for the nonsampled units using independent draws from the Bayesian posterior predictive distribution, $(Y_{mis}|X, Y_{obs})$, i.e., conditional on the observed data *D*. Alternatively, one can generate synthetic values of *Y* for all *N* units based on the posterior predictive distribution of "future" or "super" populations $(Y_f|D)$, conditional on the observed data. This procedure ensures that the synthetic populations contain no real values of *Y*, thereby avoiding the release of any observed value of *Y*. Second, a random sample (e.g., simple random sample) of size $n_{syn}$ is drawn from each of *M* synthetic populations. These sampled units comprise the public-use data sets that are released to, and analyzed by, data users.

From these publicly-released synthetic data sets, data users can make inferences about a scalar population quantity $Q = Q(X,Y)$, such as the population mean of *Y* or the population regression coefficients of *Y* on *X*. In each synthetic data set, the user estimates *Q* with some point estimator *q* and an associated measure of uncertainty *v*. Let $(q^{(i)}, v^{(i)}; i = 1,2,\ldots,M)$ be the values of *q* and *v* computed on the *M* synthetic data sets. I assume that these quantities are estimated based on a simple random sampling design. Under assumptions described in (Raghunathan, Reiter, and Rubin, 2003), the data user can obtain valid inferences for scalar *Q* by combining the $q^{(i)}$ and $v^{(i)}$ using the following quantities:

$$\bar{q}_M = \sum_{i=1}^{M} q^{(i)}/M \quad (1)$$

$$b_M = \sum_{i=1}^{M} (q^{(i)} - \bar{q}_M)^2/(M-1) \quad (2)$$

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

where $q_M$ is used to estimate *Q*, and

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

is used to approximate the variance of $\bar{q}_M$. A disadvantage of $T_M$ is that it can be negative. Negative values generally can be avoided by making *M* and $n_{syn}$ large. A more precise variance estimator that is always positive is outlined in (Raghunathan, Reiter, and Rubin, 2003). Inferences for scalar *Q* are based on a normal distribution when $T_M > 0$, *n*, *M*, and $n_{syn}$ are large. For moderate *M*, inferences can be based on t-distributions (Reiter, 2002).

## 3 Creation of Synthetic Data Sets for Small Area Estimation

I adopt a Bayesian approach, using a hierarchical imputation model, to generate synthetic data for small area estimation. Hierarchical models have been used in several applications of small area estimation (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 imputation of missing data in multilevel data structures (Yucel, 2008; Reiter, Raghunathan, and Kinney, 2006).

My approach involves three stages. In the first stage, incremental regression models are fit using the observed data within each small area to approximate the joint conditional density of the set of variables to be synthesized. In the second stage, the joint sampling distribution of regression parameters is approximated and the between-area variation is modeled by incorporating state-level covariates. In the final stage, the regression parameters are simulated and are used to draw synthetic microdata values from the posterior predictive distribution.

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

### 3.1 Stage 1: Direct Estimates

For descriptive purposes, I introduce the following notation. I define small areas as *counties*, nested within *states*, which could be nested within an even larger area (e.g., region). In specific terms, suppose that a sample of size *n* is drawn from a finite population of size *N*. Let $n_{sc}$ and $N_{sc}$ denote the respective sample and population sizes for county $c = 1,2,\ldots,C_s$ within state $s = 1,2,\ldots,S$. Let $Y_{sc} = (Y_{icsp}; i = 1,2,\ldots,n_{sc}; p = 1,2,\ldots,P)$ represent the $n_{sc} \times P$ matrix of survey variables collected from each survey respondent located in county *c* and state *s*. Let $X_{sc} = (X_{icsj}; i = 1,2,\ldots,n_{sc}, n_{sc}+1,\ldots,N_{sc}; j = 1,2,\ldots,J)$ represent the $N_{sc} \times J$ 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 $Y_{sc}$ only, it is straightforward to synthesize the auxiliary variables $X_{sc}$ as well.

A desirable property of synthetic data is that the multivariate relationships between the observed variables are maintained in the synthetic data, i.e., the joint distribution of variables is preserved. The first task is to specify the joint conditional distribution of the observed county-level variables to be synthesized $(Y_{sc1}, Y_{sc2}, \ldots, Y_{scP}|X_{scj})$, where the synthetic values are drawn from a corresponding posterior predictive distribution. Specifying and simulating from the joint conditional distribution can be difficult for complex data structures involving large numbers of variables representing a variety of distributional forms. Alternatively, one can approximate the joint density as a product of conditional densities (Raghunathan et al., 2001). Drawing synthetic variables from the joint posterior density $(Y_{sc1}, Y_{sc2}, \ldots, Y_{scP}|X_{scj})$ can be achieved by sampling from $(Y_{sc1}|X_{scj})$, $(Y_{sc2}|Y_{sc1}, X_{scj}), \ldots, (Y_{scP}|Y_{sc1}, \ldots, Y_{sc,P-1}, X_{scj})$. In practice, a sequence of generalized linear models are fit on the observed county-level data where the variable to be synthesized comprises the outcome variable and any auxiliary variables or previously fitted variables are used as predictors, e.g., $Y_{icsp+1} = (X_{ics}, Y_{ics1,\ldots,p})\beta_{sp} + \varepsilon_{icsp}$. 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, if needed, to satisfy modeling assumptions. After fitting each conditional density, estimates of the regression parameters $\hat{\beta}_{cs}$, the corresponding covariance matrix $\hat{V}_{cs}$, and the residual variance $\hat{\sigma}^2_{cs}$ are obtained and incorporated into the hierarchical structure described below in Section 3.2.

### 3.2 Stage 2: Sampling Distribution and Between-Area Model

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

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

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

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

where $Z_s = (Z_{sk}; k = 1, 2, \ldots, K)$ is a $K \times 1$ matrix of state-level covariates, $\beta$ is a $(J+p) \times K$ matrix of population regression coefficients, and $\Sigma$ is a $(J+p) \times (J+p)$ covariance matrix. State-level covariates are incorporated into the hierarchical model in order to "borrow strength" from related areas. Prior distributions may be assigned to the unknown parameters $\beta$ and $\Sigma$, but for ease of presentation, I assume that $\beta$ and $\Sigma$ 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).

### 3.3 Stage 3: Synthetic Data Generation for Small Areas

The ultimate objective is to generate synthetic populations within a small area using an appropriate posterior distribution. To this end, one can simulate the unknown regression parameters $\beta_{cs}$ specified in the hierarchical model described in Section 3.2. Based on standard theory of the normal hierarchical model (Lindley and Smith, 1972), the distribution of the population regression parameters is,

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

where $\tilde{\beta}_{as}$ is a simulated vector of values for the vector of regression parameters $\beta_{as}$. Simulating a synthetic variable $\tilde{Y}_{as}$ for observed variable $Y_{as}$ from the posterior predictive distribution can then be achieved by drawing $\tilde{Y}_{as}$ from a parametric distribution with location and scale parameters $X_{as}\tilde{\beta}_{as}$ and $\sigma^2_{as}$, respectively, where $\sigma^2_{as}$ may be drawn from an appropriate posterior predictive distribution $(\sigma^2_{as}|Y_{as}, X_{as})$; alternatively, the maximum likelihood estimate $\hat{\sigma}^2$ obtained from Section 3.1 may be used. For example, to simulate a normally distributed variable $Y_{as,1}$ one can draw $\tilde{Y}_{as,1}$ from the distribution $N(X_{as}\tilde{\beta}_{as}, \hat{\sigma}^2)$. Generating a second (normally distributed) synthetic variable $\tilde{Y}_{as,2}$ from the posterior predictive distribution $(Y_{as,2}|Y_{as,1}, X_{as})$ is achieved by drawing $\tilde{Y}_{as,2}$ from $N(X'_{as}\tilde{\beta}_{as}, \hat{\sigma}^2)$, where $X'_{as} = (X_{as}, \tilde{Y}_{as,1})$. If the second synthetic variable is binary, then $\tilde{Y}_{as,2}$ is drawn from $Bin(1, \hat{p}(X'_{as}\tilde{\beta}_{as}))$, where $\hat{p}(X'_{as}\tilde{\beta}_{as})$ is the predicted probability computed from the inverse-logit of $X'_{as}\tilde{\beta}_{as}$. For polytomous variables, the same procedure is adapted to obtain posterior probabilities for each categorical response and the synthetic values are sampled from a multinomial distribution. This iterative process continues until all synthetic variables $(\tilde{Y}_{as,1}, \tilde{Y}_{as,2}, \ldots, \tilde{Y}_{as,P})$ are generated. This procedure is repeated *M* times to create multiple replicates of synthetic variables $(\tilde{Y}^{(i)}_{as,1}, \tilde{Y}^{(i)}_{as,2}, \ldots, \tilde{Y}^{(i)}_{as,P}; i = 1,2, \ldots, M)$. In addition, the entire process 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 each small area. Inferences for a variety of small-area estimands $Q_{as}$ and large-area estimands $Q_a$ or $Q$ can be obtained using the combining rules in Section 2.

## 4 Evaluation of Synthetic Data for Small Area Inferences

In this section, I illustrate the above procedure on a subset of restricted microdata from the U.S. American Community Survey (ACS). I generate fully-synthetic data sets for relatively small geographic areas (i.e., counties) and evaluate the analytic validity of the resulting estimates. The data consist of seven household-level variables and seven person-level variables measured on 846,832 households and 2,093,525 persons during years 2005-2009. The variables, shown in Table 1, were chosen by researchers at the U.S. Census Bureau for this project. We treat small areas as counties identified in the restricted ACS microdata. All such areas are non-overlapping and are nested within a state. I restrict the sample to the Northeast region, which contains 9 states and 217 counties.

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

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

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

The first survey variable to be synthesized is household size. Creating a household size variable facilitates the generation of synthetic person-level variables in a later step. Because no administrative or other conditioning variables $X_{ss}$ are available for this application, household size is simulated using a Bayesian Poisson-Gamma model conditional on the observed household size variable with unknown hyperparameters estimated using maximum likelihood estimation: specifically, the Newton-Raphson algorithm. The remaining household-level variables are synthesized using the hierarchical modeling procedure described in Section 3. The sample selection weights (both household and person) are included among the set of variables to be synthesized. State-level covariates $Z_s$, including population size (log-transformed), number of metropolitan, and number of micropolitan areas, are incorporated into the hierarchical model.

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

Once the household variables were synthesized, the synthetic household data sets were converted to person-level data sets and the person-level variables were synthesized conditional on the household-level variables. Taylor series linearization (Binder, 1993) was used to obtain design-based regression estimates, accounting for the clustering of persons within households. To reduce the ordering effect of synthesizing the household variables first, I implemented 5 conditioning cycles where each synthetic variable was conditioned on the full set of household- and person-level variables from the previous implementations.

All analyses were conducted at the Michigan Census Research Data Center at the Institute for Social Research in Ann Arbor, Michigan. The Census Bureau's Disclosure Review Board approved the output presented here.

### 4.1 Univariate Inferences for Small Areas

I evaluate the analytic validity of the synthetic data by comparing county estimates obtained from the synthetic data with those obtained from the observed data for all 217 counties. First, I compute basic univariate estimates, namely, overall means (or proportions), subgroup means, standard deviations, and standard errors for each county; multivariate estimates are evaluated in Section 4.2.

Table 2 presents the overall mean of the county means and standard errors obtained from the synthetic and observed data. Scatter plots of synthetic and observed county means are shown in Appendices 1-2. The fifth column contains the regression intercept and slope of the observed point estimates regressed against the synthetic point estimates for all 217 counties. A slope equal to (or close to) 1 indicates a strong linear correspondence between the synthetic and observed estimates. On average, most of the synthetic county means are generally within two standard errors of the observed county means and the estimated slopes are relatively close to the desired value of 1. One exception is the Age variable, which is overestimated by the synthetic data. The observed age variable has a bimodal distribution, which is not ideally simulated with a Gaussian distribution; this is a limitation of the Bayesian parametric framework. Nonparametric strategies may

improve the accuracy of the synthetic data. Some of the binary variables (e.g., black, hispanic, rent, poverty) are overestimated by the synthetic data due to pooling of neighboring areas. For example, the prevalence of blacks in many counties did not meet the minimum sample size criterion and had to be pooled with neighboring counties to obtain stable estimates of regression parameters. Other instances where the synthetic data overestimates small area statistics occur for percentile estimates. For the estimated percentage of households with incomes greater than the $50^{th}$ percentile, the synthetic data estimate corresponds quite well to the actual estimate. However, as the percentiles increase, the accuracy of the synthetic data drops. Simulation results (not shown) yielded relatively high confidence interval coverage for the estimated synthetic means, producing coverage rates ranging from 0.86 to 0.99). Aggregating the synthetic data to the state- and region-levels yielded estimates with similar correspondence to the observed aggregated data (not shown), indicating that this method may be useful for producing valid estimates across multiple levels of geography.

Table 2. Mean of synthetic and observed county means/proportions and standard errors and regression slope of actual means on the synthetic means for all 217 counties.

| | **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 > 75th pctile,% | 19.34 | 21.49 | 37.34 | 38.69 | 0.59 | 0.43 | -0.00 | 0.91 |
| Income > 90th pctile,% | 6.78 | 8.38 | 22.96 | 24.58 | 0.35 | 0.24 | 0.56 | 0.74 |
| Income (Mortgage=1) | 84667.0 | 86992.6 | 69019.2 | 58960.1 | 1536.0 | 1195.3 | 5460.0 | 0.91 |
| Income (Own=1) | 61076.6 | 60456.9 | 76053.1 | 45083.6 | 2132.8 | 1232.7 | 1717.0 | 0.98 |
| Income (Rent=1) | 38844.5 | 36921.9 | 37759.4 | 32527.3 | 1436.0 | 1166.5 | 3480.0 | 0.99 |
| *Person variables* | | | | | | | | |
| Sampling weight | 10.27 | 10.67 | 7.59 | 8.02 | 0.08 | 0.14 | -0.09 | 0.97 |
| Gender (%) | 48.63 | 48.63 | 49.97 | 49.97 | 0.53 | 0.44 | 0.04 | 0.91 |
| Education (%) | | | | | | | | |
| < 12 years | 31.48 | 31.67 | 46.31 | 46.31 | 0.49 | 0.39 | 0.09 | 0.71 |
| 12 years | 28.34 | 27.74 | 44.40 | 44.06 | 0.48 | 0.57 | 0.01 | 0.97 |
| 13-15 years | 20.33 | 20.25 | 40.11 | 40.04 | 0.43 | 0.50 | 0.01 | 0.96 |
| 16+ years | 19.85 | 20.35 | 38.72 | 39.14 | 0.40 | 0.51 | -0.01 | 1.00 |
| Hispanic (%) | 3.85 | 4.23 | 15.72 | 16.99 | 0.14 | 0.26 | -0.00 | 1.00 |
| Age | 40.89 | 41.16 | 22.98 | 30.34 | 0.25 | 0.27 | 22.02 | 0.46 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Race (%) | | | | | | | | |
| White | 92.21 | 91.34 | 22.17 | 24.08 | 0.20 | 0.36 | 0.01 | 1.00 |
| Black | 3.55 | 4.01 | 14.54 | 16.26 | 0.13 | 0.26 | -0.01 | 1.00 |
| Other | 4.24 | 4.65 | 14.54 | 18.61 | 0.16 | 0.27 | -0.00 | 1.00 |
| Poverty (%) | 8.65 | 9.04 | 27.54 | 28.13 | 0.30 | 0.53 | -0.00 | 1.00 |
| Poverty (White=1; %) | 7.93 | 8.19 | 26.41 | 26.84 | 0.30 | 0.51 | -0.00 | 1.00 |
| Poverty (Black=1; %) | 20.48 | 21.30 | 36.86 | 37.03 | 4.62 | 3.52 | -0.01 | 1.01 |
| Poverty (Other=1; %) | 16.62 | 17.84 | 35.37 | 36.07 | 2.96 | 4.38 | 0.01 | 0.87 |
| Poverty (Hispanic=1; %) | 19.92 | 21.11 | 37.08 | 37.96 | 3.52 | 5.54 | -0.01 | 0.98 |

## 4.2 Multivariate Inferences for Small Areas

Next I evaluate the analytic validity of the synthetic data for multivariate estimates. Table 3 presents overall summary results of two multiple regression models fitted within each county. The first model regresses household income (cube root) on the remaining household-level variables, and the second model regresses a binary variable indicating college graduation (college graduate vs. less than high school/high school graduate/some college) on all other person-level variables. The summary measures shown in Table 3 consist of overall means of the estimated regression coefficients and corresponding standard errors obtained from each county. Scatter plots of synthetic and observed regression coefficients are presented in Appendices 3-4. On average, the synthetic point estimates correspond relatively well with the observed point estimates. The synthetic point estimates lie within about two standard errors of the observed point estimates, on average. Many of the synthetic standard errors are similar in magnitude to the observed standard errors, on average; however, the loss of information in the synthetic data is apparent for the estimates that yield relatively larger synthetic standard errors (Hispanic, race). Simulation results (not shown) for the estimated regression coefficients yielded confidence interval coverage rates that ranged from 0.93 to 0.99. Relatively strong correspondence was found between the synthetic and observed coefficient estimates when the synthetic data were aggregated to higher levels of geography (e.g., states, region).

Table 3. Mean of synthetic and observed county regression coefficients and standard errors and regression slope of actual coefficients on the synthetic coefficients for all 217 counties.

| | **Y=Household income (linear)** | |
|---|---|---|
| Household-level covariates | Actual Beta (SE) | Synthetic Beta (SE) |
| Intercept | 24.34 (1.11) | 24.26 (1.09) |
| Household size | 1.52 (0.14) | 1.44 (0.14) |
| Sampling weight | -0.04 (0.24) | -0.05 (0.26) |
| Total bedrooms | 1.15 (0.19) | 1.23 (0.18) |

| | | |
|---|---|---|
| Electricity bill/mo. | 0.99 (0.18) | 1.04 (0.17) |
| Total rooms | 1.25 (0.14) | 1.26 (0.13) |
| Tenure | | |
| Mortgage/loan | Ref | Ref |
| Own free & clear | -3.47 (0.37) | -3.05 (0.34) |
| Rent | -6.01 (0.44) | -6.84 (0.47) |
| | **Y=College graduate (logistic)** | |
| Person-level covariates | Actual Beta (SE) | Synthetic Beta (SE) |
| Intercept | -2.27 (0.12) | -2.17 (0.13) |
| Sampling weight | 0.03 (0.05) | 0.03 (0.05) |
| Gender: Male | -0.06 (0.06) | -0.06 (0.05) |
| Hispanic | -0.70 (0.34) | -0.66 (0.67) |
| Age | 0.02 (0.001) | 0.02 (0.05) |
| Race | | |
| White | Ref | Ref |
| Black | -1.06 (0.36) | -0.65 (0.80) |
| Other | 0.23 (0.24) | 0.33 (0.36) |
| Poverty | -1.26 (0.17) | -1.26 (0.28) |

## 5 Conclusions

This study addresses an important data dissemination issue facing statistical agencies, which is how to meet the growing demand for high quality, public-use microdata for small geographic areas while protecting data confidentiality and respondent privacy. These competing aims are likely to receive even more attention in the future as research into small area effects and societal sensitivity towards privacy continues to grow.

This paper proposes a fully-synthetic data approach that utilizes a hierarchical model for creation of microdata for small geographic areas. The resulting data sets could conceivably be released to the public, along with additional data products that contain finer levels of detail than those currently being released. The methodology is flexible, easy to implement, and can be straightforwardly adapted to a variety of Federal statistical surveys and other data sources representing various geographical structures and variable types. This approach has been applied to other large Federal survey data sets, including the National Health Interview Survey and additional evaluations are underway.

Results of the empirical ACS evaluation suggest that valid small area inferences can be obtained from fully-synthetic data for basic descriptive and multivariate estimands. However, there is room for improvement as there was not always strong correspondence between the synthetic and actual estimates. More flexible modelling

approaches, such as those utilizing nonparametric imputation models, could be used to improve the quality of the synthetic data and the resulting small area estimates, particularly for variables that do not follow strict parametric distributions.

One issue that was not addressed in this paper is the level of disclosure protection offered by the synthetic data for small areas. Although there is evidence that fully-synthetic data offers better protection against disclosure than partially-synthetic data (Drechsler, Bender, and Raessler, 2008), this may not be true for small geographic areas or sparse subpopulations. Further research is needed to determine whether fully-synthetic data offers adequate levels of disclosure protection to be suitable for public release in a small area context.

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

## References

Abowd, J.M., Stinson, M., Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. Int. Stat. Rev. 51, 279-292

Datta, G.S., Fay, R.E., Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. In: Proceedings of the Bureau of the Census 1991 Annual Research Conference, pp. 63-79. U.S. Bureau of the Census, Washington, DC

Drechsler, J., Bender, S., Raessler, S. (2008). Comparing fully and partially synthetic data sets for statistical disclosure control in the German IAB establishment panel. Trans. Data Priv 1(3), 105-130

Kennickell, A.B. (1997). Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In: Alvey, W., Jamerson, B. (eds.) Record Linkage Techniques 1997. pp. 248-267. National Academy Press, Washington DC

Kinney, S.K., Reiter, J.P. (2008). Making public use, synthetic files of the Longitudinal Business Database. In: Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference Proceedings. Istanbul, Turkey

Little, R.J.A. (1993). Statistical analysis of masked data. J. Off. Stat. 9, 407-426

Liu, F., Little, R.J.A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In: Proceedings of the Joint Statistical Meetings, pp. 2133-2138. American Statistical Association, Blacksburg, VA

Lindley, D.V., Smith, A.F.M. (1972). Bayes estimates for the linear model. J. Roy. Stat. Soc. B 34(1), 1-41

Malec, D., Sedransk, J., Moriarity, C.L., LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. J. Amer. Stat. Assoc. 92(439), 815-826

Platek, R., Rao, J.N.K., Sarndal, C.E., Singh, M.P. (1987). Small area statistics. Wiley, New York

Raghunathan, T.E., Reiter, J.P., Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. J. Off. Stat. 19, 1-16

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv. Methodol. 27, 85-95

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. Surv. Methodol. 25, 175-186

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

Reiter, J.P. (2005). Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. J. Roy. Stat. Soc. A 168, 185-205

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. Surv. Methodol. 30, 235-242

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. J. Off. Stat. 18, 531-544

Fay, R.E. III., Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to Census data. J. Amer. Stat. Assoc. 74(366), 269-277

Reiter, J.P., Raghunathan, T.E., Kinney, S. (2006). The importance of modeling the sampling design in multiple imputation for missing data. Surv. Methodol. 32, 143-150

Reiter, J.P. (2003). Inference for partially synthetic public use microdata sets. Surv. Methodol. 29, 181-188

Rubin, D.B. (1993). Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. J. Off. Stat. 9, 461-468

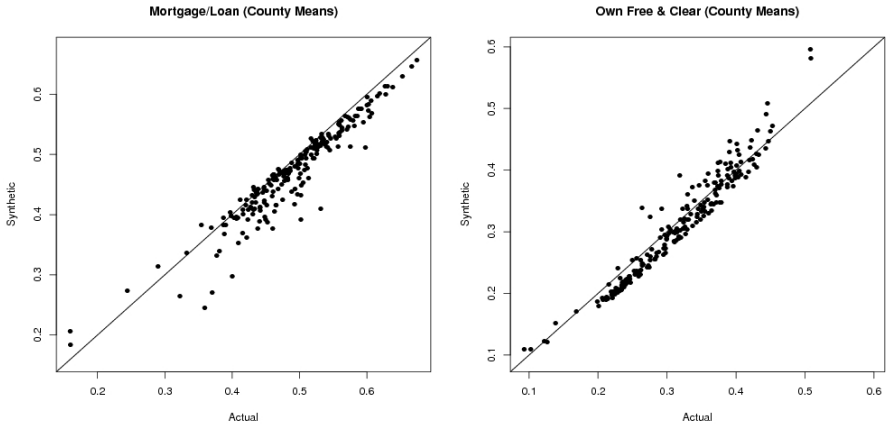
Rodriguez, R. (2007). Synthetic data disclosure control for American Community Survey group quarters. In: Proceedings of the Joint Statistical Meetings, pp. 1439-1450. American Statistical Association, Salt Lake City, UT

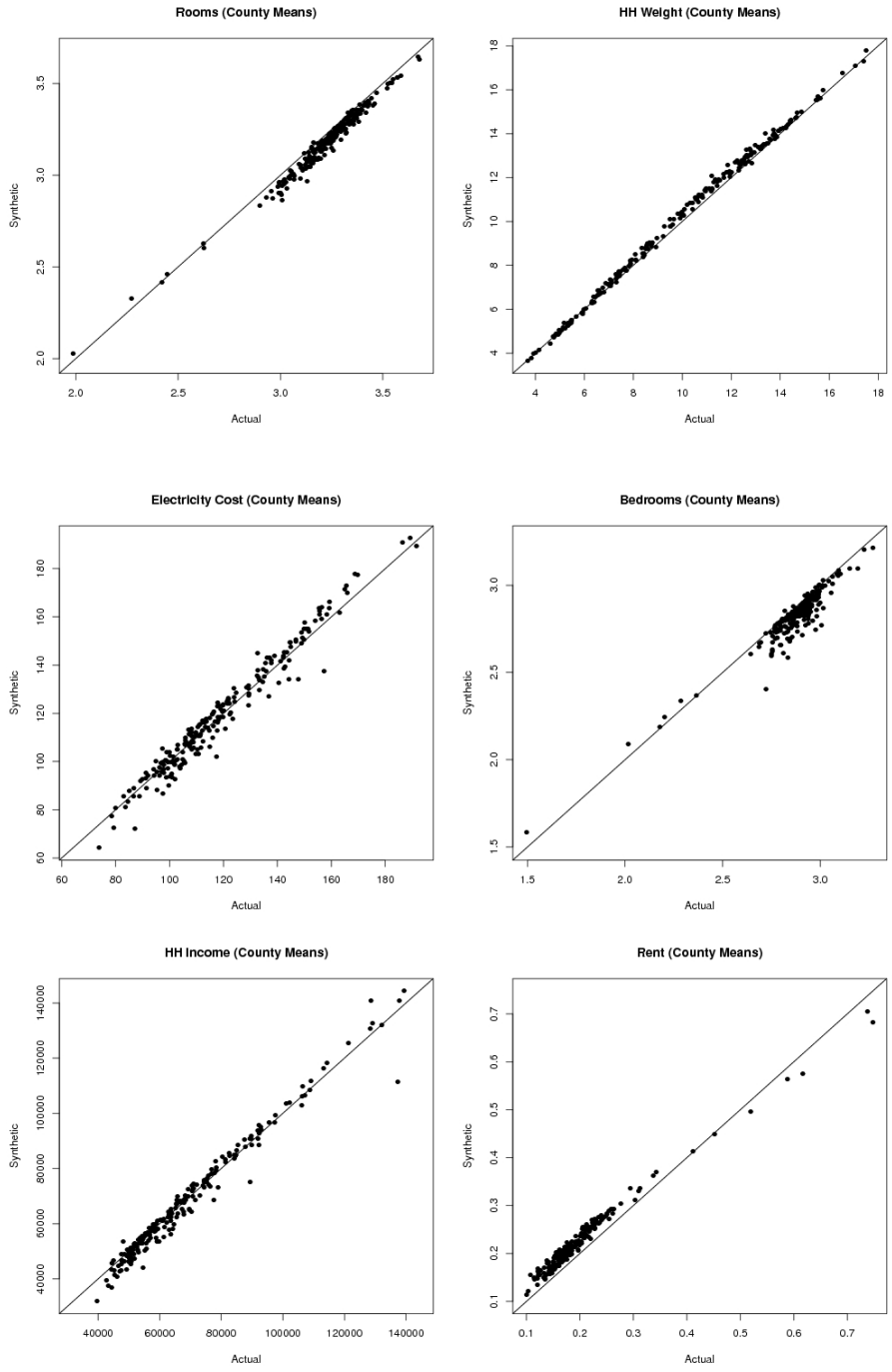
Tranmer, M., Pickles, A., Fieldhouse, E., Elliot, M., Dale, A., Brown, M., Martin, D., Steel., D., Gardiner, C. (2005). The case for small area microdata. J. Roy. Stat. Soc. A 168, 29-49

Yu, Mandi. (2008). Disclosure Risk Assessments and Control. Doctoral Dissertation, University of Michigan

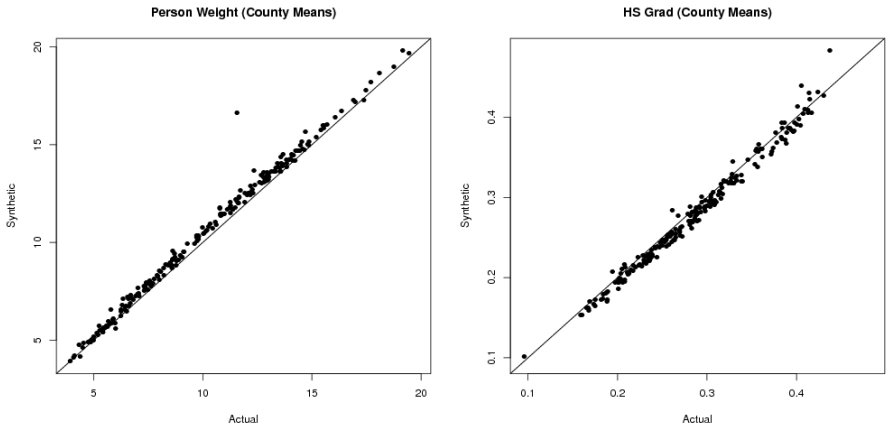
Yucel, R.M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. Phil. Trans. R. Soc. A 366(1874), 2389-2403

## Appendix 1 Scatter plot of synthetic and observed county means for household-level variables

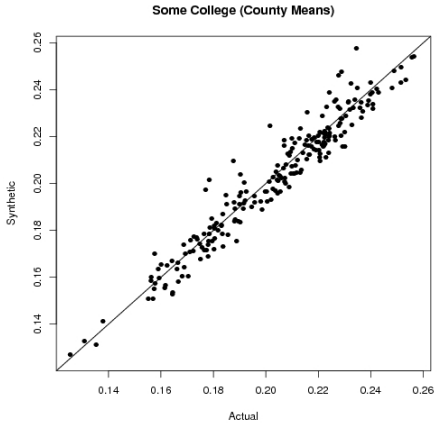


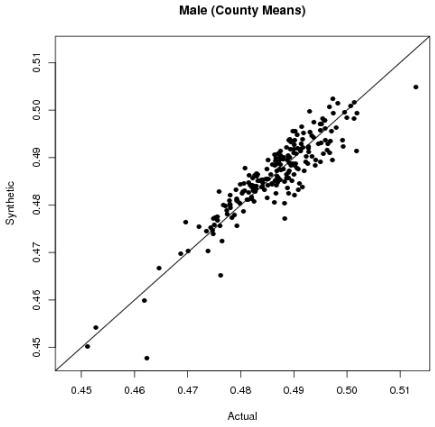


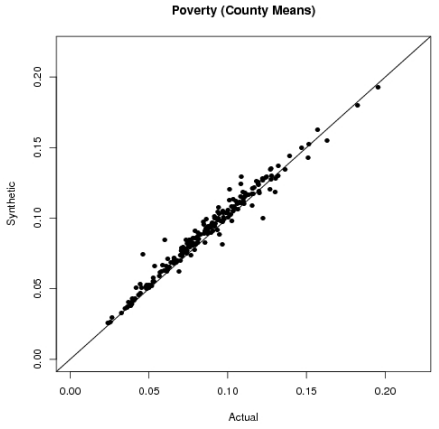

**Appendix 2 Scatter plot of synthetic and observed county means for person-level variables**

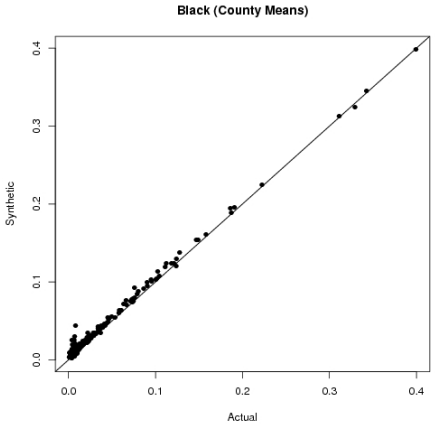

Person Weight (County Means)
Synthetic
Actual
5
10
15
20

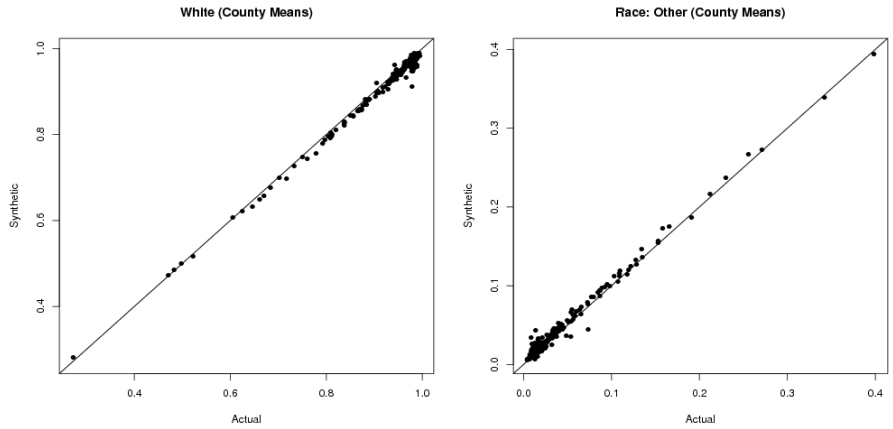

HS Grad (County Means)
Synthetic
Actual
0.1
0.2
0.3
0.4


Some College (County Means)
Synthetic
Actual
0.14
0.16
0.18
0.20
0.22
0.24
0.26

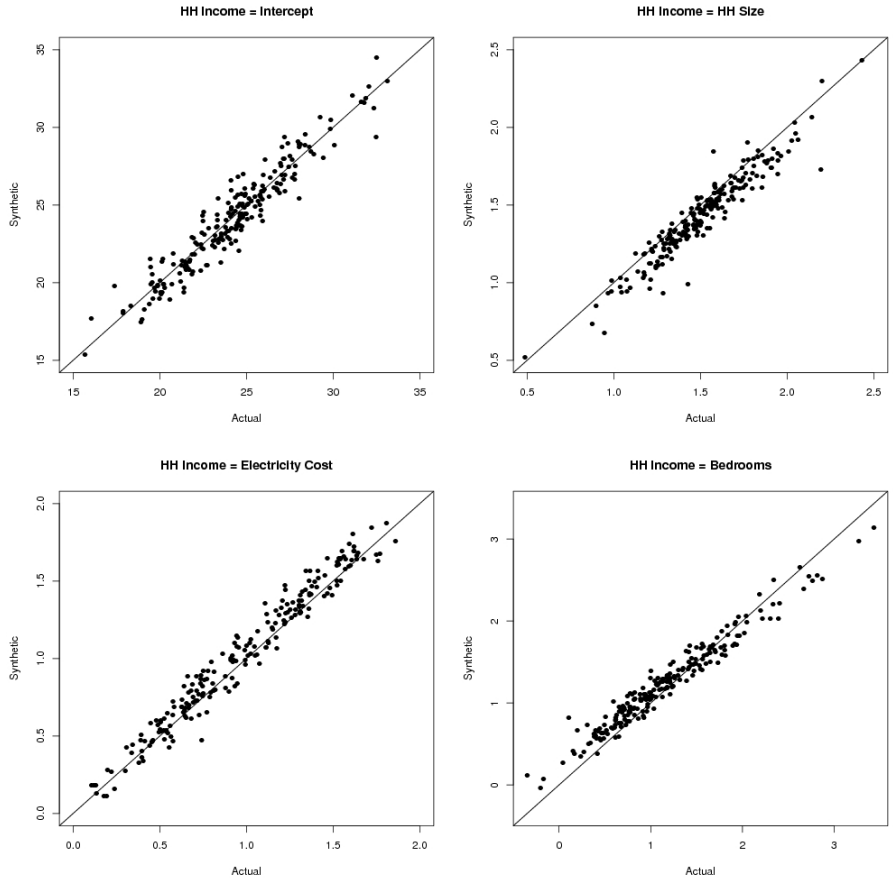

Male (County Means)
Synthetic
Actual
0.45
0.46
0.47
0.48
0.49
0.50
0.51

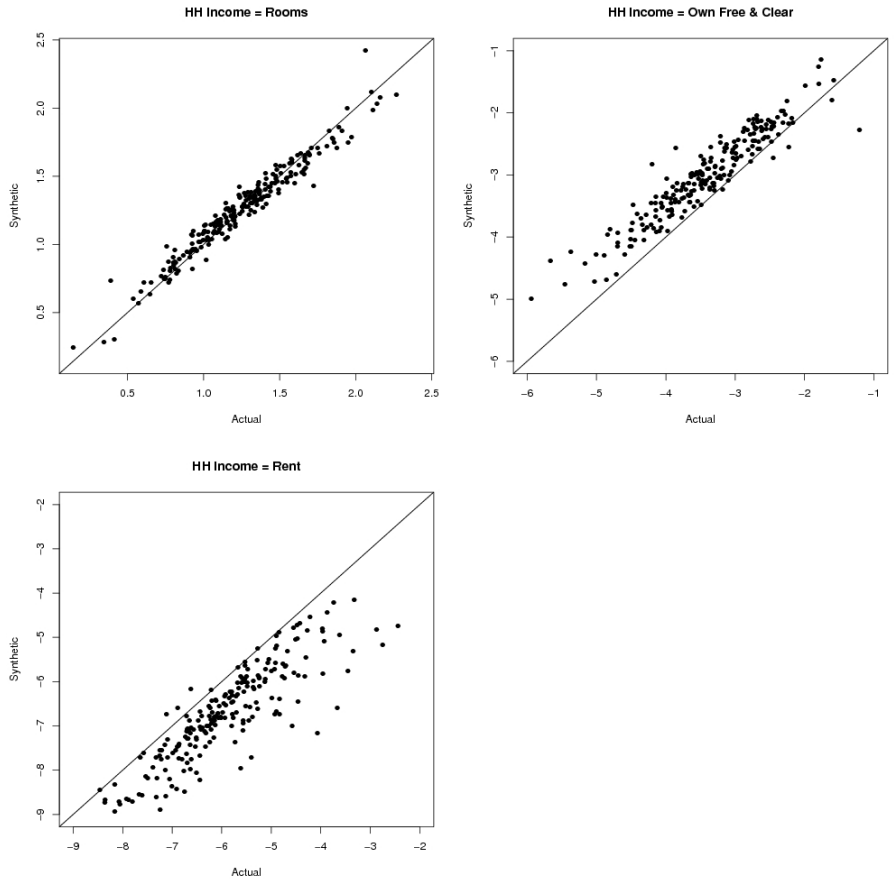

Poverty (County Means)
Synthetic
Actual
0.00
0.05
0.10
0.15
0.20


Black (County Means)
Synthetic
Actual
0.0
0.1
0.2
0.3
0.4

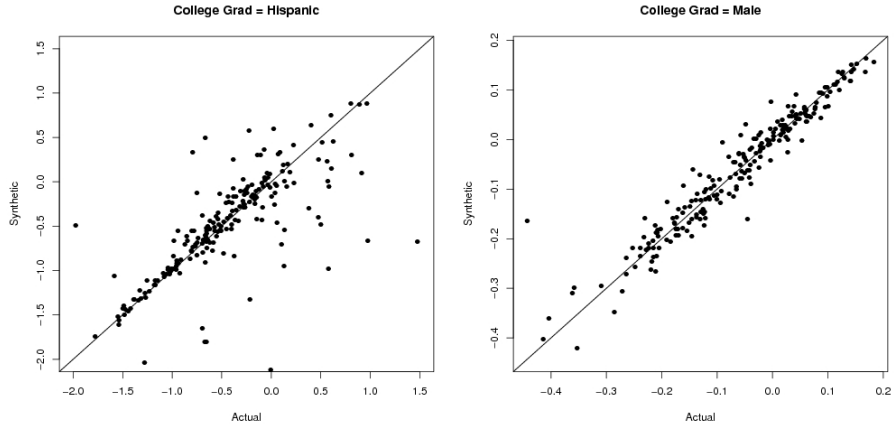


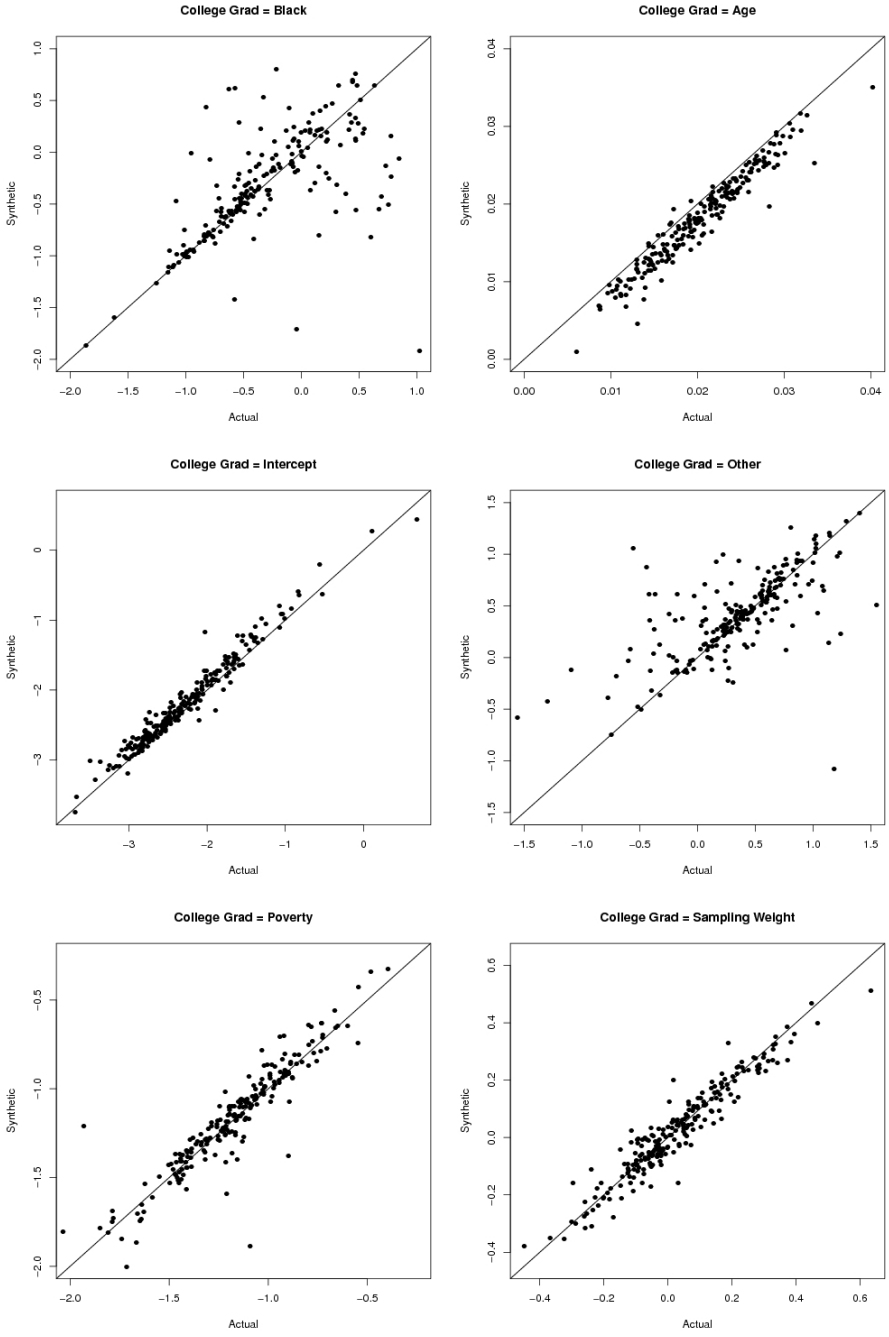

**Appendix 3 Scatter plot of synthetic and observed county linear regression coefficients of household income (cube root) on household-level variables**






## Appendix 4 Scatter plot of synthetic and observed county logistic regression coefficients of college graduation on household-level variables




College Grad = Black
Synthetic
Actual
College Grad = Age
Synthetic
Actual
College Grad = Intercept
Synthetic
Actual
College Grad = Other
Synthetic
Actual
College Grad = Poverty
Synthetic
Actual
College Grad = Sampling Weight
Synthetic
Actual