J. R. Statist. Soc. A (2005)
168, *Part* 1, *pp.* 185–205

Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study

Jerome P. Reiter

Duke University, Durham, USA

[Received May 2002. Revised September 2003]

Summary. The paper presents an illustration and empirical study of releasing multiply imputed, fully synthetic public use microdata. Simulations based on data from the US Current Population Survey are used to evaluate the potential validity of inferences based on fully synthetic data for a variety of descriptive and analytic estimands, to assess the degree of protection of confidentiality that is afforded by fully synthetic data and to illustrate the specification of synthetic data imputation models. Benefits and limitations of releasing fully synthetic data sets are discussed.

Keywords: Confidentiality; Disclosure; Microdata; Multiple imputation; Synthetic data

1. Introduction

Statistical agencies that release public use microdata confront a dilemma. On one hand, the agencies must guard the confidentiality of survey respondents' data. On the other hand, they seek to maximize the utility of the released data for analyses. Agencies address this dilemma in various ways. These include recoding variables into coarse categories, such as releasing only 5-year intervals for age, reporting exact values only below certain thresholds, e.g. reporting all incomes above 100 000 as '100 000 or more', swapping some units' data values with other units' data values and adding random noise to data values. However, these methods can compromise estimation by distorting relationships between variables in the data set. Furthermore, they complicate analyses for users. For example, to analyse properly data that include additive random noise, users should apply measurement error models (Fuller, 1993).

An alternative approach was proposed by Rubin (1993): release multiply imputed, synthetic data sets. In this approach, the agency

(a) randomly and independently samples units from the sampling frame for each synthetic data set and
(b) imputes unknown data values for units in the synthetic samples by using models fitted with the original survey data.

This fully synthetic approach is compelling for two main reasons. First, it can preserve confidentiality, since the identification of units and their sensitive data can be difficult when the released data are not actual collected values. Second, with appropriate imputation and estimation methods based on the concepts of multiple imputation (Rubin, 1987), the approach can allow data users to make valid inferences for various estimands by using standard, complete-data statistical methods and software.

Address for correspondence: Jerome P. Reiter, Institute of Statistics and Decision Sciences, Box 90251, Duke University, Durham, NC 27708, USA.
E-mail: jerry@stat.duke.edu

 0964–1998/05/168185

The synthetic data approach recently has attracted the interest of statisticians in both academia and statistical agencies. Raghunathan *et al.* (2003) developed methods for combining point and variance estimates from multiple synthetic data sets to obtain valid inferences. Their combining rules differ from the rules for multiple imputation of missing data that were developed by Rubin (1987). Reiter (2002) showed that these methods can yield valid inferences in a variety of survey design settings. He also discussed the selection of the number and sample size of synthetic data sets. Kennickell (1997), Abowd and Woodcock (2001) and Liu and Little (2002) investigated a variant of the approach: release the units that are originally surveyed but replace some of these units' data with multiple imputations. Inferential methods that are appropriate for such partially synthetic data sets have been developed by Reiter (2003). Other discussions and variants of synthetic data approaches include those in Little (1993), Fienberg *et al.* (1998), Dandekar *et al.* (2002a, b), Franconi and Stander (2002, 2003), Polettini *et al.* (2002), Polettini (2003) and General Accounting Office (2001).

Despite this growing interest, few empirical studies of the benefits and limitations of implementing Rubin's original proposal have appeared in published journals. Raghunathan *et al.* (2003) conducted a simulation with data from the US Consumer Expenditure Survey, but they generated data from a bootstrap of observed records rather than sampling new units from the sampling frame. The bootstrap releases observed units' complete records, which may compromise the protection of confidentiality. Kennickell (1997) and Abowd and Woodcock (2001) conducted simulations with genuine data, but their methods release units that are originally surveyed rather than simulate entirely new data sets.

This paper presents an empirical study of releasing fully synthetic microdata. Simulations based on data from the US Current Population Survey are used

(a) to evaluate the potential validity of inferences based on fully synthetic data for a variety of descriptive and analytic estimands,
(b) to assess the degree of protection of confidentiality that is afforded by fully synthetic data and
(c) to illustrate the specification of synthetic data imputation models.

The remainder of the paper is organized as follows. Section 2 presents the methods that are proposed by Raghunathan *et al.* (2003) for generating and obtaining inferences from fully synthetic data sets. Section 3 discusses the protection of confidentiality and the utility of data for the fully synthetic data approach, including comparisons with alternative disclosure control techniques. Section 4 and Section 5 describe respectively the data set and the methods of generating synthetic data that are used in the empirical study. Section 6 contains the results of the study, illustrating the benefits and limitations of the synthetic data approach. Finally, Section 7 remarks on implementing the fully synthetic approach in practice.

2. Synthetic data generation and inferential methods

To describe the construction of and inferences from multiply imputed, synthetic data sets, we use notation that is similar to that of Raghunathan *et al.* (2003). Let $z_j = 1$ if unit j is selected in the original survey, and $z_j = 0$ otherwise. Let $\mathbf{z} = (z_1, \ldots, z_N)$, where N is the number of units in the population. Let Y_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $z_j = 1$, let Y_{nobs} be the $(N - n) \times p$ matrix of unobserved survey data for the units with $z_j = 0$, and let $Y = (Y_{\text{obs}}, Y_{\text{nobs}})$. For simplicity, it is assumed that all sampled units fully respond to the survey. Let X be the $N \times d$ matrix of design variables for all N units in the population, e.g. stratum or cluster indicators or size measures. It is assumed that such design information is

known at least approximately, e.g. from census records or the sampling frames. When it is not known, X can be treated as part of Y and imputed.

The agency that releases synthetic data, henceforth abbreviated as the *imputer*, constructs synthetic data sets based on the observed data, $(X, Y_{\text{obs}}, \mathbf{z})$, in a two-part process. First, the imputer imputes values of Y for the $N-n$ unobserved units to obtain a completed data population, $(X, Y_{\text{com},i})$. For reasons that are discussed in Rubin (1987) and Raghunathan *et al.* (2003), imputations should be generated from the Bayesian posterior predictive distribution of $(Y|X, Y_{\text{obs}}, \mathbf{z})$. The imputer also may choose to impute values of Y for all N units so that the completed data contain no real values of Y, thereby avoiding the release of any respondent's actual value of Y. Second, the imputer samples n_{syn} units randomly from the completed data population $(X, Y_{\text{com},i})$, using a simple random sample. These sampled units are released as public use data, so that the released data set $(X, Y_{\text{syn},i})$ contains the values of Y only for units in the synthetic sample. This entire process is repeated independently $i = 1, \ldots, m$ times to obtain m different synthetic data sets, which are released to the public. In practice, it is not necessary to generate completed data populations for constructing the $Y_{\text{syn},i}$. The imputer needs only to generate values of Y for units in the synthetic samples.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the *analyst*, seeks inferences about some estimand $Q = Q(X, Y)$, where the notation $Q(X, Y)$ means that the estimand Q is a function of (X, Y). For example, Q could be the population mean of Y or the population regression coefficients of Y on X. In each synthetic data set, the analyst estimates Q with some estimator q and the variance of q with some estimator v. It is assumed that the analyst specifies q and v by acting as if the synthetic data were collected data from a simple random sample of (X, Y).

For $i = 1, \ldots, m$, let q_i and v_i be respectively the values of q and v in synthetic data set i. Under assumptions that are described in Raghunathan *et al.* (2003), the analyst can obtain valid inferences for scalar Q by combining the q_i and v_i. Specifically, the following quantities are needed for inferences:

$$\bar{q}_m = \sum_{i=1}^{m} q_i/m, \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m-1), \tag{2}$$

$$\bar{v}_m = \sum_{i=1}^{m} v_i/m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_{\text{s}} = \left(1 + \frac{1}{m}\right) b_m - \bar{v}_m \tag{4}$$

to estimate the variance of $\bar{q}_m$. The $b_m - \bar{v}_m$ is an approximately unbiased estimator of the variance of $q_{\text{obs}} = Q(X, Y_{\text{obs}}, \mathbf{z})$, and the $(1/m)b_m$ adjusts for using only a finite number of synthetic data sets. Although it is possible for $T_{\text{s}} < 0$, negative values generally can be avoided by making m and n_{syn} large. A more complicated variance estimator that is always positive is described in Raghunathan *et al.* (2003).

When $T_{\text{s}} > 0$, and n, n_{syn} and m are large, inferences for scalar Q can be based on a normal distribution, so that a synthetic 95% confidence interval for Q is

$$\bar{q}_m \pm 1.96\sqrt{T_{\text{s}}}. \tag{5}$$

Inferential methods for multivariate Q are not presented here.

3. Protection of confidentiality and utility of data

Statistical agencies that release altered rather than original data—e.g. by recoding variables, swapping data or adding noise—sacrifice some utility of the data in exchange for greater protection of confidentiality. Releasing fully synthetic data also involves such trades. This section spells out confidentiality–utility trades for fully synthetic data, with comparisons with alternative disclosure control techniques.

3.1. Protection of confidentiality

Disclosure control methods are typically evaluated on two measures of disclosure risk: the risk of reidentification and of predictive disclosure (Willenborg and de Waal, 2001). Reidentifications occur when analysts identify sampled units in the released data and therefore learn their sensitive values. To reduce the risk of reidentification agencies sometimes alter values of key identifiers such as age, sex and race. Predictive disclosures occur when analysts use the altered data to estimate closely unknown sensitive values. An example of reducing the risk of predictive disclosure is adding random noise to sensitive values, where the noise is drawn randomly with sufficiently large variance.

When fully synthetic data are released, the risk of reidentification is practically non-existent. Almost none of the released, synthetic units are in the original sample, having been randomly selected from the sampling frame. Their values of survey data are simulated, so that no genuine sensitive values are disclosed for these units. Furthermore, the synthetic records cannot be matched meaningfully to records in other data sets, such as administrative records, because the values of released survey variables are simulated rather than actual and, therefore, not identical to those in administrative databases.

In contrast, there are risks of reidentification for other disclosure control methods. For example, even when ages are collapsed in 5-year categories, analysts may be able to reidentify records by examining rare combinations of other characteristics; or analysts with access to administrative databases may be able to match on several unaltered variables. With data swapping, typically some records are not altered, so there are positive probabilities of reidentifications, as well as the potential for using administrative records to obtain matches for perturbed records. Similar risks apply when data are protected with additive noise. For example, Yancey *et al.* (2002) used record linkage techniques to match perturbed records to administrative records, obtaining substantial probabilities of true matches even after noise has been added to identifying variables.

The gravest risks of reidentification are often for units with values that are in the tails of distributions. Recoding or swapping key identifiers can be less effective than fully synthetic data at protecting such values. To illustrate this, we consider a scenario like one described by Duncan *et al.* (2001). Suppose that a statistical agency samples household incomes from a particular neighbourhood. Further, suppose that the analyst knows that a particular household is in the sample and has the largest income in the neighbourhood. When the actual value of that income is released in the data set, the analyst learns the exact value of that household's income, even if key identifiers are recoded, swapped or suppressed. With fully synthetic data, imputations can be made so that no actual values of income are released, thereby protecting that household's and other households' incomes.

Although safe from the risk of reidentification, releasing fully synthetic data is subject to predictive disclosure risk. The seriousness of this risk depends on the properties of the imputation models that are used to simulate data. For example, when imputations are made from a regression model with a very small variance, analysts can estimate outcomes precisely; or,

when imputations are made by using bootstrap methods, real values are released, which may lead to disclosures. Imputers can reduce this risk by using less precise imputation models and avoiding bootstraps. Willenborg and de Waal (2001), pages 44–45, also pointed out that predictive disclosures for units that are not in the original sample—almost all units in fully synthetic data sets—may not constitute an unlawful breach of confidentiality, since no promise of confidentiality is made to units that are not in the original sample. Alternative disclosure control techniques face similar predictive disclosure risks (Willenborg and de Waal, 2001).

3.2. Utility of data

The previous arguments suggest that the fully synthetic approach is at least as effective, and can be more effective, at reducing risks of disclosure than other disclosure control techniques. We now consider the utility of fully synthetic data. Here, the approach is not so dominant: there are advantages and disadvantages of the approach relative to other methods.

A primary advantage of the fully synthetic approach is that, when data are simulated from posterior predictive distributions that reflect the distributions of the observed data, frequency valid inferences can be obtained from the multiple synthetic data sets (Raghunathan *et al.*, 2003). Furthermore, these inferences can be determined by combining standard likelihood-based or randomization-based estimates; the analyst need not learn new statistical methods or software programs. In contrast, perturbed data should be analysed by using the likelihood-based methods that were described by Little (1993) or the measurement error methods that were described by Fuller (1993), which are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

Fully synthetic data sets can have other positive utility features, as discussed by Raghunathan *et al.* (2003). First, synthetic data sets can be sampled by schemes other than the typically complex design that is used to collect the original data, so that analysts can ignore the design for inferences and instead perform analyses based on simple random samples. Second, imputation models can incorporate adjustments for non-sampling error and can borrow strength from other sources of data, thereby resulting in inferences that can be even more accurate than those based on the original data. Third, because all units are simulated, geographic identifiers can be included in the data sets, making estimation easier for small areas. These substantial gains in the utility of data are not realized with other disclosure methods.

There is a cost to these benefits: the validity of synthetic data inferences depends critically on the validity of the models that are used to generate the synthetic data. This is because the synthetic data reflect only those relationships that are included in the imputation models. When the models fail to reflect accurately certain relationships, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions that are built into the models will be passed on to the users' analyses. This dependence is a potentially serious limitation to releasing fully synthetic data. Practically, it means that some analyses cannot be performed accurately, and that the imputer needs to release information that helps analysts to decide whether or not the synthetic data are reliable for their analyses.

Of course, other disclosure control techniques have limitations on the utility of data as well. For example, when ages are recoded in 5-year categories, the recoded data cannot be reliably used to model differences across individual ages; when high incomes are reported only as 'above 100000', the data provide no detail on the upper tail of income. Swapping data or adding noise typically results in biased estimates of joint distributions and, therefore, some estimands that rely on joint distributions. Additionally, as mentioned previously, obtaining valid inferences from

perturbed data requires special software and non-standard statistical methods. Such limitations could render the released data practically useless for some analysts.

Clearly there are confidentiality–utility trade-offs to releasing fully synthetic data. As noted by Rubin (1993) in his original proposal, the pay-offs to releasing synthetic data are potentially so large that the approach is worth considering. The main question is whether imputations can provide adequate data utility. Although answers to this question are context specific, one useful way to assess this question at least partially is to implement and to evaluate empirically the approach on genuine data.

4. Data for empirical study

The empirical study is based on public release data from the March 2000 US Current Population Survey. The data are comprised of 51016 households for a total of 133710 people. There are household indicators, so that each person can be placed in a single household. There are no geographic identifiers in the data.

Synthetic data sets are generated for the 10 variables that are displayed in Table 1. These variables were selected and provided by statisticians at the US Bureau of the Census. Although these 10 variables are only a subset of the variables in the Current Population Survey, imputing them presents several generic modelling challenges. In particular, the data set includes individual level and household level variables, both numerical and categorical variables, numerical variables with clearly non-Gaussian distributions and numerical variables with large percentages of values equal to 0.

Marital status M has seven types: $M = 1$ for married civilians with both spouses present at the home; $M = 2$ for married people in the armed forces with both spouses present at the home; $M = 3$ for married people with one spouse not present at the home; $M = 4$ for widowers; $M = 5$ for divorced people; $M = 6$ for separated people; $M = 7$ for people who never have been married. All 30484 children, defined as people under age 15 years, have $M = 7$.

For people age 15 years or over, highest attained education level E increases from 31 to 46 in correspondence with years of schooling. As examples, $E = 31$ represents highest educational attainments of less than first grade, $E = 35$ represents highest educational attainments of ninth grade, $E = 39$ represents a high school degree, $E = 43$ represents a Bachelor's degree, $E = 44$ represents a Master's degree, $E = 45$ represents a professional school degree and $E = 46$ represents a doctoral degree. All children are in a child-only category defined as $E = 0$.

Table 1. Description of the variables that are used in the empirical study

Variable	*Label*	*Range*
Sex	X	Male, female
Race	R	White, black, American Indian, Asian
Marital status	M	7 categories, coded 1–7
Highest attained education level	E	16 categories, coded 31–46
Age (years)	G	0–90
Child support payments ($)	C	0, 1–23917
Social security payments ($)	S	0, 1–50000
Household alimony payments ($)	A	0, 1–54008
Household property taxes ($)	P	0, 1–99997
Household income ($)	I	−21011–768742

All children receive zero child support payments C and zero social security payments S, i.e. their values of $C = S = 0$. Because the data come from a public use data file, the monetary variables and age G are restricted to lie below certain maximum values. Such top coding would not be necessary when releasing synthetic data.

Marginally, there are ample numbers of people in each sex, race, marital status and education category. Many cross-classifications have few or 0 people, especially those involving minorities with $M \notin \{1, 7\}$. Out of the 133 710 people, there are 18 389 who receive social security payments and 2508 who receive child support payments. Only 206 households receive positive alimony payments, whereas 33 076 have positive property taxes. There are 132 households with negative incomes, 582 with zero income and the remainder with positive income. The negative incomes are legitimate values: some households actually report paying out more money than they took in over the year. The distributions of positive values for all monetary variables are right skewed.

5. Generation of synthetic data sets

In the study, the 133 710 people are the target population. Samples of observed data, (X, Y_{obs}), are collected by randomly selecting $n = 10\,000$ households and using their actual data. This sample size yields a decent probability that subpopulations are represented adequately, while providing ample variability in survey estimates. Observed data are sampled independently 500 times.

For each of these 500 sets of observed data, $m = 100$ synthetic data sets, $(X, Y_{\text{syn},i})$ for $i = 1, \ldots, 100$, are generated, each with $n_{\text{syn}} = 10\,000$ households selected by independent simple random sampling from the target population. Setting $m = 100$ reduces the chance that many values of T_s are less than 0 (Reiter, 2002). The effect on inferences of using $m = 10$ is discussed in Section 6. Setting $n_{\text{syn}} = 10\,000$ keeps the sampling fraction of synthetic data reasonably small while allowing sufficient synthetic sample size to examine the accuracy of inferences for subpopulation estimands.

Each synthetic data set includes the actual values of sampled peoples' ages, races and sexes. People are placed in their actual households, and the head of the household is flagged. This assumes that the demographic variables in this target population are available to the imputer, e.g. from a census. Because the survey variables are simulated, releasing real values of demographic variables and placing people in households should not result in disclosures, provided that the imputation models sufficiently protect the survey variables.

5.1. Methods for imputing survey variables

When generating synthetic data, the goal is to reproduce as much of the structure in the target population as possible, without compromising confidentiality. This goal primarily drives the construction of the imputation models that are used here. All models are selected by a process of 'trial and error'. For simulation, parsimonious models are selected to ensure that coefficients can be estimated in nearly any random sample of observed data. Hence, the models that are used here may not maximize the utility of data for these variables.

The imputation models are specified by using a sequence of conditional regressions. Each successive regression may include demographic variables and survey variables from the preceding regressions as predictors, but not variables from future regressions. The order of the conditional regressions is E–M–A–C–S–I–P. Other orderings can be used; this one is chosen because it facilitates model specification. Education and marital status are fitted first because it is conceptually logical to model alimony and child support payments as functions of education

and marital status rather than the other way around. Additionally, fitting them first is computationally convenient: this reduces the number of parameters in the multinomial regression models that are used for imputations. The variables A, C and S are imputed before I because these three variables are components of income. In general, when specifying the sequence of regressions, imputers should select one that leads to convenient modelling of the variables that are in the data set.

The regressions are fitted using only people aged 15 years or older. In addition to the variables in Table 1, variables that are used in the models include household size (labelled HS), an indicator for household heads (labelled HH) and the number of people who are under age 18 years (labelled HY, for household youths). For E, M and S, values are drawn first for household heads (HH = 1), and then for other household members (HH = 0). This facilitates the modelling of within-household relationships for these variables, allowing the models for non-heads to include predictors that are associated with values for household heads. For C, values for household heads and other household members are imputed from one combined model, which includes an indicator for household heads. There are not sufficient numbers of people with $C > 0$ to estimate accurately parameters in two separate models. Values of person level variables for children are fixed by definition.

The basic forms of the models are outlined in Table 2, in which predictors that are modelled as continuous are in bold, and those somehow categorized are in regular type. The labels for variables are identical to those used in Table 1. Exact specifications of the models, as well as rationales for these specifications, can be found in Appendix A.

Using these imputation models, synthetic data sets are then drawn from posterior predictive distributions for each variable in sequence. This involves

(a) drawing values of the parameters from their posterior distributions, or approximations to those distributions, given the observed data, and
(b) generating synthetic values of the variables given the drawn values of the parameters, relevant demographic data and relevant imputed data for previously generated variables in that synthetic data set.

Flat prior distributions are assumed on all the parameters. The methods that were used to draw parameters and synthetic values are described in Appendix B.

Table 2. Imputation models for survey variables

Variable	*Type of model*	*Predictors*
E	Multinomial logit	X, R, G
M	Multinomial logit	X, R, G, HS, HY, E
A	Logistic regression (0 or greater than 0)	X, E, M
	Bayesian bootstrap (when greater than 0)	None
C	Logistic regression (0 or greater than 0)	$X, R, \boldsymbol{G}, \boldsymbol{G}^2$, HH, HY, E, M
	Linear regression (when greater than 0)	$X, R, \boldsymbol{G}, \boldsymbol{G}^2$, HH, HY, E, M
S	Logistic regression (0 or greater than 0)	X, R, G, E, M
	Linear regression (when greater than 0)	X, R, G, E, M
I	Multinomial logit (income categories)	$X, R, \boldsymbol{G}, \boldsymbol{G}^2$, HS, E, M, A, C, S
	Linear regression (some categories)	X, R, G, HS, E, M, A, C, S
	Bayesian bootstrap (some categories)	None
P	Logistic regression (0 or greater than 0)	$X, R, \boldsymbol{G}, \boldsymbol{G}^2$, HY, E, M, I
	Linear regression (when greater than 0)	$X, R, \boldsymbol{G}, \boldsymbol{G}^2$, **HY**, E, M, I

5.2. General issues with imputing data

This section discusses issues with imputation that could occur generally when implementing the fully synthetic approach.

The sample size affects the specification of the imputation models. Some cells in cross-tabulations of categorical variables have low or zero counts in the observed data samples. Such cells are collapsed in the imputation models, effectively setting to zero some interactions and main effects when generating synthetic data. Since the synthetic data include the same information as the imputation models, these interactions or main effects are estimated as zero when analysing the synthetic data. Imputers can avoid forcing interactions and main effects to zero by using informative prior distributions for the parameters that are associated with these terms. Such priors can be constructed from other sources, e.g. previous survey data or administrative records.

Some estimated parameters in the models have small estimated variances. This happens most often in the multinomial regressions, particularly when people in less populated demographic subgroups have similar multinomial outcomes. Imputations for such groups tend to be similar across units and close to the estimated expected values of the models. This is not a problem for inferences, but it could potentially raise concerns about confidentiality. When there are parameters with small estimated variances, imputers can check for predictive disclosures and, if necessary, use coarser imputation models.

It is difficult to model precisely within-household relationships, and imputations may generate unlikely household compositions. For example, the marital statuses of household members are strongly related to the marital status of the household head. In fact, they are partly deterministic: if $M = 1$ for the household head, then $M = 1$ for at least one other person in the household. The imputation models that are used here do not guarantee this determinism, although it is extremely rare to impute households that have only one person with $M = 1$. It is also possible to generate unlikely spousal arrangements by chance. For example, a household with a person who is 52 years old, a person 55 years old and a person 78 years old might be imputed $M_{52} = 1$, $M_{55} = 7$ and $M_{78} = 1$. To minimize the chance of releasing households with odd marital status arrangements, imputers can perform automated data edits that fix oddities.

Imputations from bootstraps might cause concern about confidentiality, since actual data values are released. For example, suppose that intruders know that a particular person is in the original survey, and they know the percentile of that person's income. The intruders may be able to determine that person's exact income by finding the corresponding percentile of the bootstrapped incomes. In the absence of such detailed information on particular people, bootstrapped values should be sufficiently confidential, since the values are not attached to their actual units. Otherwise, imputations can be drawn from parametric or other nonparametric methods rather than bootstraps.

When the drawn values of parameters are far from the posterior modes, imputed data can be out of line with the observed data and the synthetic data from other imputations. The effects of such rogue synthetic imputations are somewhat attenuated when a large number of synthetic data sets are released. For smaller m, they can have a strong affect on $\bar{q}_m$ and b_m, leading to inaccurate estimates. To diagnose this problem, imputers can perform sensibility checks on each of the synthetic data sets. For example, they can examine histograms of all the synthetic point estimates q_i for common estimands. Rogue synthetic data sets then can be discarded and replaced with new, independent replications. Another strategy, which was employed by Abowd and Woodcock (2001), is to constrain the drawn value of each parameter to lie within 3 standard deviations of its distribution's posterior mode.

6. Simulation results

This section examines the properties of inferences based on the imputations that were described in Section 5. The estimands include population and subpopulation means and percentages, as well as coefficients from linear and logistic regressions. It is important to note that the properties of these inferences are specific to this empirical study; however, the results do provide general insight into the benefits and limitations of releasing fully synthetic data.

6.1. Inferences for descriptive estimands

Table 3 summarizes the properties of inferences for 32 descriptive estimands. In Table 3, the estimand $\bar{D}$ is the average difference in social security payments for household heads and other household members. $\bar{D}$ is calculated by using only households with HS > 1 and people older than age 54 years. The estimand β_A is the coefficient of A in the ordinary least squares regression of C on A for household heads. The third column of Table 3 displays the median of each $\bar{q}_{100}$ across the 500 simulation runs. The fourth column displays the absolute differences between these medians and their corresponding Q as percentages of the Q. Standard errors of the $\bar{q}_{100}$ are sufficiently small to indicate that large differences reflect biases more than simulation variance. The last column displays the percentages of the 500 synthetic, 95% confidence intervals that cover their corresponding Q.

The synthetic variance estimates T_s are mostly positive: for the 32 estimands, only about 200 of a possible 16000 variance estimates are negative. All except 20 of these negative values are associated with the categories for marital status of black women. Negative values of T_s are set equal to $\bar{u}_m$ for constructing confidence intervals.

For most estimands, the absolute percentage differences are less than 10%, indicating that the synthetic data point estimates are typically close to their corresponding Q. Although this does not provide a decisive general argument for releasing synthetic data, it certainly is encouraging that these imputations result in reasonable estimates for a wide range of descriptive estimands.

There are some estimands for which the synthetic data do not yield close point estimates. The β_A has practically 100% error. This is because the synthetic data generation models do not include a correlation between C and A, except for residual correlation due to common predictors in the models for imputing C and A. This highlights a drawback of releasing fully synthetic data: when relationships are not specified in the imputation models, they cannot be recovered in the synthetic data. Of course, the estimand β_A is not especially meaningful; few analysts would fit a linear regression when both variables have large density spikes at zero. In fact, of the 500 confidence intervals for β_A based on the observed data, only 20% cover the true value $\beta_A = 0.18$.

Large differences also exist for the synthetic estimators of the joint probabilities of having at least a Bachelor's degree and being an American Indian. There are 128 American Indians with $E > 42$ in the target population, so the difference between the population joint probabilities (0.05%) and the synthetic joint probabilities (0.10%) corresponds to about 64 additional people. This results because, in the interest of parsimony, the imputation models for education assume similar relationships between education and other variables across races. This smoothing assumption does not hold exactly in the population, although the resulting error in the number of American Indians with $E > 42$ is small in absolute numbers.

Moderately large percentage errors exist for $\bar{D}$ and some marital status categories for black women. These errors result from the models' inabilities to capture precisely within-household relationships for these variables. This problem does not affect analyses that are based only on household heads.

Table 3. Inferences from simulations for descriptive estimands

Estimand	Q	*Median of* $\bar{q}_{100}$	$\left\lvert \frac{Q - median}{Q} \right\rvert$	*Coverage of synthetic 95% confidence intervals*
Person level estimands				
$\bar{C}$	74	77	0.04	94.0
$\overline{SS}$	1201	1216	0.01	99.0
$\bar{E}$ (household head)	39.8	39.8	0.00	73.0
% of people over age 55 years with $SS > 0$	60	60	0.00	98.6
% of people with $C > 0$	1.88	1.91	0.02	95.8
$\bar{D}$	1054	744	0.30	91.2
β_A	0.18	0.01	0.94	0
Household level estimands				
$\bar{I}$	52632	51610	0.02	38.4
$\bar{P}$	1008	1080	0.07	7.6
$\bar{A}$	41	45	0.10	96.6
% of households with $A > 0$	0.40	0.43	0.08	99.2
% of households with $P > 0$	65	65	0.00	96.0
% of households with $I < 14000$	15.4	13.3	0.14	0.6
% of households with $I > 100000$	11.7	11.2	0.04	58.0
% of households with $I > 200000$	2.1	2.1	0.00	75.6
Subgroup estimands				
% of household heads who are divorced and have $A > 0$	0.30	0.32	0.07	98.4
% of people who are divorced and have $CS > 0$	0.77	0.79	0.04	95.2
% of people with $E \geqslant 43$ and who are				
White men	7.4	7.4	0.00	95.6
Black men	0.43	0.46	0.07	93.4
Asian men	0.48	0.46	0.02	95.6
American Indian men	0.05	0.10	1.00	98.6
White women	6.9	6.8	0.01	93.0
Black women	0.58	0.53	0.08	80.4
Asian women	0.48	0.43	0.10	87.8
American Indian women	0.05	0.09	0.80	99.0
% of black females (including children) who are				
Single	58.23	59.27	0.02	56.0
Married	28.72	26.00	0.09	6.4
Married with spouse not at home	1.11	1.34	0.21	56.8
Married with spouse in armed forces	0.16	0.14	0.12	67.0
Divorced	6.57	7.38	0.12	49.8
Separated	2.86	3.28	0.14	57.0
Widowed	2.34	2.51	0.10	67.2

The results for the synthetic 95% confidence intervals are mixed. Encouragingly, half of the intervals have coverage rates that are over 90%. Discouragingly, 11 estimands' intervals have coverage rates that are between 50% and 90%, and five estimands' intervals have coverage rates that are less than 50%. These low rates result because the biases of these $\bar{q}_{100}$, even when small as percentages of Q, are relatively large compared with their standard errors. Similar problems can be expected in other implementations of the fully synthetic data approach, since biases are likely to exist for some estimands.

Using a smaller m does not affect the biases in the synthetic point estimators, but it does increase their variances. This can increase the coverage rates of synthetic confidence intervals when the biases of the $\bar{q}_m$ become swamped by the increased standard errors. A repetition of the original simulation design using $m = 10$ instead of $m = 100$ shows only small improvements in confidence coverage. Using smaller m increases the number of negative variance estimates to over 1600. As before, these negative variances are concentrated in the marital status categories. However, almost all estimators have some negative variance estimates, which is not so with $m = 100$.

6.2. Inferences for analytic estimands

This section considers inferences for the coefficients of two regression models, both fitted by using only data on the heads of households. The first is a linear regression of $\log(I)$ on a quadratic function of age and on indicator variables for sex, race, marital status and education. Gaussian errors are assumed in the model, and only households with $I > 0$ are used in parameter estimation. The second model is a logistic regression used to predict whether or not a household has positive property taxes. The predictors in the logistic regression include a quadratic function of age, a linear function of income and indicator variables for sex, race, marital status, education and whether there are youths in the household. These analytical models differ from the imputation models for I and P, which are displayed in Appendix A.

Table 4 summarizes the synthetic data inferences for the coefficients of these two models. In addition to the percentage differences and synthetic 95% confidence interval coverage rates, Table 4 displays the medians across the 500 simulation runs of the coefficients' t-statistics, based on both the observed and the synthetic data. Similar values of the observed and synthetic t-statistics result in similar conclusions about statistical significance, a desirable property for analytical inferences from synthetic data. Standard errors of the t-statistics across the 500 simulation runs are small, so large differences in the observed and synthetic medians are not the result of simulation variance. Medians of the q_{obs} are close to their corresponding Q, and observed data confidence intervals are almost all within simulation error of 95% coverage.

For the linear regression, the synthetic point estimates are typically within about 10% of their corresponding population quantities. Two exceptions include the coefficients of the indicator variables for American Indians and Asian Americans. These result because part of the imputation model for income—the multinomial regression that is used to predict categories of income—collapses all non-whites into one category. Most of the confidence intervals have good coverage rates, although some have poor coverage rates due to biases in the point estimates. Most median t-statistics for the synthetic data are similar to those for the observed data, again except for the indicator variables for American Indians and Asian Americans.

Unlike the promising results for the linear regression, the results for the logistic regression are discouraging. Several coefficients have large biases, leading to 0% confidence coverage. The synthetic t-statistics are, for many coefficients, quite different from the observed t-statistics. For example, analysts would reach different conclusions about the statistical significance for the indicator variables for sex and educational attainment below high school. Substantial percentages of the variance estimates T_s are negative. In fact, some coefficients have more than 40% negative T_s.

The differences in the synthetic data and observed data inferences are caused by uncongeniality (Meng, 1994) in the analytical and imputation models. The analytical model fits P on a continuous variable for income and indicator variables for education, whereas the imputation model fits P on indicator variables for income and a continuous variable for education.

Table 4. Inferences from simulations for analytic estimands

Predictor	Q	*Median of* $\bar{q}_{100}$	$\left\|\frac{Q-\text{median}}{Q}\right\|$	*Coverage of synthetic 95% confidence interval*	*Median of* $q_{\text{obs}}/\sqrt{u_{\text{obs}}}$	*Median of* $\bar{q}_{100}/\sqrt{T_{\text{s}}}$
Linear regression of log(I)						
Intercept	9.38	9.46	0.01	78.0	130	136
Age	0.048	0.043	0.11	50.6	17.2	16.8
Age^2	−0.00052	−0.00046	0.11	40.2	−19.2	−19.0
Sex (male, 0; female, 1)	−0.135	−0.132	0.03	97.0	−8.1	−7.3
Indicator for $M=2$	−0.064	−0.066	0.03	85.2	−0.4	−0.6
Indicator for $M=3$	−0.555	−0.553	0.00	95.0	−8.7	−7.6
Indicator for $M=4$	−0.558	−0.520	0.07	69.6	−17.6	−19.2
Indicator for $M=5$	−0.591	−0.569	0.04	84.6	−24.9	−23.2
Indicator for $M=6$	−0.797	−0.755	0.05	86.6	−16.9	−14.9
Indicator for $M=7$	−0.576	−0.576	0.00	97.2	−24.4	−21.4
Indicator for $R=2$	−0.182	−0.169	0.07	96.8	−7.0	−5.6
Indicator for $R=3$	−0.275	−0.117	0.57	96.4	−3.8	−0.6
Indicator for $R=4$	0.078	0.035	0.56	89.6	1.7	0.8
Indicator for $E>38$	0.59	0.60	0.02	94.0	28.0	26.2
Logistic regression of $P>0$						
Intercept	−3.404	−4.266	0.25	0.0	−16.5	−44.9
Age	0.132	0.135	0.02	88.6	16.5	37.0
Age^2	−0.00088	−0.00086	0.02	94.2	−11.6	−25.0
Sex (male, 0; female, 1)	−0.041	−0.157	2.79	0.0	−0.8	−7.0
Income	0.000014	0.000019	0.33	0.0	18.4	24.5
Indicator for $M=2$	−1.153	−0.011	0.90	2.0	−2.8	−0.5
Indicator for $M=3$	−1.237	−0.380	0.69	0.0	−7.6	−4.1
Indicator for $M=4$	−0.607	−0.486	0.20	19.0	−6.9	−11.9
Indicator for $M=5$	−1.00	−0.427	0.57	0.0	−15.3	−13.7
Indicator for $M=6$	−1.58	−0.385	0.76	0.0	−12.4	−6.0
Indicator for $M=7$	−1.20	−0.380	0.68	0.0	−17.9	−11.8
Indicator for $R=2$	−0.570	−0.705	0.24	24.6	−8.1	−20.9
Indicator for $R=3$	−0.181	−0.267	0.48	85.4	−1.0	−3.2
Indicator for $R=4$	−0.861	−0.840	0.02	98.6	−6.7	−14.3
Indicator for $E>43$	0.019	−0.023	2.18	90.4	0.2	−0.6
Indicator for $E<39$	−0.598	0.051	1.08	0.0	−10.0	1.8
Indicator for HY>0	0.214	0.398	0.86	0.0	4.0	15.9

Simulations that are not shown here using an analytical model that matches the imputation model result in similar synthetic data and observed data inferences. The analyst can use model diagnostics, as well as information provided by the imputer about the imputation model, to determine that the analytical model does not fit the synthetic data and therefore differs from the imputation model. Assuming that the imputations are reasonable, such a lack of fit should give the analyst a hint to alter the analytical model.

As suggested by the theory of Raghunathan *et al.* (2003) and the literature on multiple imputation (Rubin, 1987; Meng, 1994), synthetic data and observed data inferences should not be too dissimilar when imputations are drawn from close approximations to the distributions of the survey data. The large differences that are seen here suggest that the imputation model for P can be improved to approximate better the relationships between P and other variables. Such improvements are not pursued here so that the consequences of uncongeniality with imperfect imputations are clearly illustrated.

6.3. Illustrative evaluation of protection of confidentiality

As discussed in Section 3, fully synthetic data sets in general should be well protected from re-identifications, but predictive disclosures may occur. This section suggests and illustrates some methods of evaluating predictive disclosure risks for fully synthetic data.

When a particular target unit's covariate pattern is known, the analyst can estimate that unit's sensitive variable by using the imputed values for synthetic units with covariate patterns that match the target's pattern. This suggests a prescription for evaluating predictive disclosure risks for fully synthetic data:

(a) repeatedly simulate values for the observed units and
(b) examine the distributions of the simulated values of sensitive variables across synthetic data sets to ensure sufficient protection.

Since predictive disclosure risk depends on the amount of information about target units that is known by the analyst (Willenborg and de Waal (2001), pages 42–46), it may be necessary to simulate values assuming different amounts of known information.

To illustrate, we examine predictive disclosure risks for income under two assumptions about the available information on target units:

(a) the only known covariates include age, race, sex and household size, and
(b) all survey variables are known except income and property taxes.

The observed data are comprised of a random sample of $n = 10\,000$ households from the population of Section 4. For each household h, 100 synthetic values of income, $I_{h,i}$, for $i = 1, \ldots, 100$, are generated on the basis of the imputation models from Appendix A, and the following quantities are computed:

$$\mathrm{DIF}_h = \left| I_{h,\mathrm{obs}} - \frac{\sum_i I_{h,i}}{100} \right|; \tag{6}$$

$$\mathrm{SD}_h = \sqrt{\left\{ \frac{1}{99} \sum_i \left(I_{h,i} - \frac{\sum_i I_{h,i}}{100} \right)^2 \right\}}; \tag{7}$$

$$\mathrm{RSE}_h = \sqrt{(\mathrm{DIF}_h^2 + \mathrm{SD}_h^2)}. \tag{8}$$

RSE_h is a typical prediction error when using the imputed income from one synthetic unit with covariate pattern h to predict the income for household h. When there are k_h units with covariate pattern h, the analyst can reduce RSE_h by averaging the imputed values for those k_h units. The reduced RSE_h is determined by dividing SD_h^2 by k_h.

The imputer can examine these quantities for households with sensitive income values. When the imputer judges the RSE_h to be sufficiently large for these households, their incomes can be considered safe from predictive disclosure. For simplicity, we assume that $k_h = 1$ for all units and examine the households with the largest and smallest incomes in the sample.

When only age, race, sex and household size are known, incomes are imputed by using the model that is described in Table 2 but using only the known variables. The resulting medians of DIF_h, SD_h and RSE_h across households equal 20 150, 42 900 and 50 700 respectively. The median RSE_h is close to the standard deviation of I in the sample (48 100), which indicates that predicting income with a single imputation is typically as accurate as using the average

income. As argued by Duncan and Mukherjee (2000), such equivalences represent good protection of confidentiality. For the household with the largest income (618 600) in the sample, $\text{DIF}_h = 575\,000$ and $\text{SD}_h = 76\,500$; for the household with the smallest income (−13 500) in the sample, $\text{DIF}_h = 46\,400$ and $\text{SD}_h = 41\,700$. These amounts of variation in imputations should amply protect these units' incomes.

When all variables except income and property taxes are known, incomes are imputed by using the model in Appendix A.6 and the actual values of the covariates. The resulting medians of DIF_h, SD_h and RSE_h across households equal 16 400, 35 400 and 42 700 respectively. For these data and imputation models, knowing the other variables typically does not help analysts to improve predictions substantially relative to knowing only age, race, sex and household size. For the household with the largest income, $\text{DIF}_j = 436\,300$ and $\text{SD}_j = 80\,900$; for the household with the smallest income, $\text{DIF}_j = 67\,800$ and $\text{SD}_j = 36\,700$. Once again, these households' incomes are well guarded.

7. Concluding remarks

This empirical study clearly illustrates the importance of specifying accurate imputation models when generating fully synthetic data. Imputers should take full advantage of subject-matter and statistical expertise, and build the most inclusive models that are permitted by the observed data. The results also suggest a method of checking imputation models. Imputers can fit several regressions or other models involving different and relevant specifications of the outcome and predictors. When the observed data and synthetic data inferences differ markedly across analytical models, it may be necessary to modify the imputation models.

The results in Sections 6.1 and 6.2 are in accordance with the theory of Raghunathan *et al.* (2003): parameters from models that are congenial to the imputation models can be estimated reasonably well from the synthetic data. To help analysts to judge whether their models are congenial, imputers should release information about the imputation models. For example, imputers can include the models as attachments to public releases of data. Or, they can include generic statements that describe the imputation models, such as 'Main effects for age, sex and race are included in the imputation models for education'. Analysts who desire finer detail than is afforded by the imputations may have to apply for special access to the observed data.

Releasing or describing the imputation models is necessary, but it is not sufficient: imputers also should release synthetic data generated from the models. Some analysts cannot generate synthetic data given the models; they need imputers to do it for them. Even when analysts can do so, it is a cumbersome burden to place on them. Additionally, when analysts want to compare competing analyses, it is advantageous if these analyses are performed on the same data sets, thereby eliminating simulation variance from comparisons. Finally, analysts may desire some function of the synthetic data that is difficult to estimate from the model parameters, but easy to determine from the synthetic data.

It may be possible to weaken the dependence of synthetic data inferences on the specification of the imputation models, while maintaining some of the disclosure limitation benefits of fully synthetic data. For example, instead of releasing non-sampled units with imputed data, imputers can release the units that were originally sampled but replace only some of the collected values with imputed values (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2001; Liu and Little, 2002; Reiter, 2003). Because partially complete, and possibly fully complete, actual records are released in these multiply imputed data sets, the resulting inferences are less sensitive to the specification of the imputation models. However, the protection of confidentiality is decreased relative to fully synthetic data, since some actual values are purposefully released

and available for matching to administrative records. Another possibility for decreasing this dependence is to use nonparametric imputation methods, e.g. imputations based on Polya trees (Lavine, 1992). Research into the effectiveness and feasibility of such methods is needed.

This study does not empirically address the relative merits of fully synthetic data as compared with other disclosure limitation techniques. Because disclosure limitation is in general context specific, and because various techniques have competing advantages, general claims are difficult to make. Empirical studies implementing multiple approaches on the same data set would provide insight into the comparative advantages of approaches. Such comparisons should consider the protection of confidentiality, properties of point estimates and confidence intervals for a wide range of estimands, and the ease of use and implementation for analysts and imputers.

Concerns over confidentiality seem to be only growing. In the future, it is conceivable that agencies may not be allowed to release any genuine data. If so, the synthetic data approach may be the only way to provide society with public use data. Further illustrations and empirical investigations will help analysts and imputers to understand the benefits and limitations of the fully synthetic approach.

Appendix A: Imputation models

This appendix describes the imputation models that are used in the empirical study. In the model formulae, an indicator variable set equal to 1 when some condition holds is described by the notation (*condition*). Variables in bold are treated as continuous. Variables in plain type are treated as a series of indicator variables. The notation $U * V$ denotes the inclusion of the main effects and interactions between two variables U and V. The generic phrase 'people' is used to mean US residents.

A.1. Education

Education is modelled by using multinomial logit regressions with E as the outcome. For household heads, models are built for two groups: people older than age 24 years and people between ages 17 and 25 years. The predictors in each of the two models include $X + R$, and additionally

$$17 < G < 25 \qquad \boldsymbol{G} + \boldsymbol{G}^2,$$
$$G > 24 \qquad (54 < G < 65) + (G > 64).$$

The models for other household members are similar, except that they include a variable for the education of the household head. This variable is labelled as HE. The predictors in each age group's model include $X + R$, and additionally

$$17 < G < 25 \qquad \boldsymbol{G} + \boldsymbol{G}^2 + (\mathrm{HE} < 39) + (\mathrm{HE} > 39),$$
$$G > 24 \qquad (54 < G < 65) + (G > 64) + (\mathrm{HE} > 44) + (42 < \mathrm{HE} < 45) + (\mathrm{HE} < 39).$$

One model for E is used for all people between age 15 and age 18 years, using $\boldsymbol{G}$ as the sole predictor.

Separate models are fitted for these age groups to reflect differences in the distributions of educational attainments. People under age 18 years are almost exclusively in high school, so their values of E are nearly a linear function of age and less than 39 years (high school graduate), regardless of race or sex. Between ages 18 and 25 years, many people are pursuing their undergraduate educations, so few of them have $E > 43$ (college graduate) and many have $E = 40$ (some college). After age 25 years, most people have completed their schooling. Data analyses on the full population indicate that educational attainment is nearly independent of age for people between ages 25 and 55 years, and that there are slight differences for people over age 55 years. The age-related indicator variables capture these population trends.

The models include race and sex terms so that the synthetic data reflect differences in the distributions for these groups. Interaction terms for sex and race are not included because they usually are not statistically significant in the observed data models, suggesting that these interactions are not well estimated and not particularly important. The models for other household members include indicator variables for education of the household head to capture some of the within-household relationships in educational attainments.

A.2. Marital status

The models for marital status are multinomial logit regressions with M as the outcome. For household heads, single-person and multiple-person households are modelled separately, because it is not possible to have $M = 1$ and $\mathrm{HS} = 1$ simultaneously. The predictors in the two models for household heads include $X * R + \boldsymbol{G} + \boldsymbol{G}^2$, and additionally

$$\begin{aligned} \mathrm{HS} = 1 \qquad & (E < 35) + (34 < E < 39) + (39 < E < 43) + (E = 43) + (E > 43), \\ \mathrm{HS} > 1 \qquad & (E < 35) + (34 < E < 39) + (39 < E < 43) + (E = 43) + (E > 43) \\ & + (\mathrm{HS} = 2) + (\mathrm{HY} > 0). \end{aligned}$$

The models for other members of households include indicator variables for the marital status of the household head, labelled HM. Including HM maintains some of the within-household relationships for marital status. The predictors in the model for other household members include $X * R + \boldsymbol{G} + \boldsymbol{G}^2$, and additionally

$$(E < 35) + (34 < E < 39) + (39 < E < 43) + (E = 43) + (E > 43) + (\mathrm{HY} > 0) + \mathrm{HM}.$$

All models include race, sex and their interactions so the strong relationships between marital status and these demographic characteristics are also found in the synthetic data. The quadratic term in age parsimoniously addresses the fact that the youngest and oldest people are likely to be unmarried (e.g. single or widowed). Education is split into indicator variables that reflect relationships that are found by using data analyses on the full population. The variable $(\mathrm{HY} > 0)$, an indicator for youths in the household, is a very strong predictor of marital status.

A.3. Alimony payments

A logistic regression is fitted to predict whether or not a household receives positive alimony payments. The logistic regression uses the predictors

$$X + (M = 5) + (E < 39).$$

Positive alimony payments are imputed by using a Bayesian bootstrap from the observed positive alimony payments.

On average only 40 households out of 10 000 have $A > 0$, so parsimonious models are needed in the simulation. This ensures that coefficients can be almost always estimated by using different random samples of observed data. Data analyses on the full population show that the three predictors X, $M = 5$ and $E < 39$ are the most useful for predicting whether $A > 0$. Alimony payments mostly go to women rather than to men, and to divorced heads of household $(M = 5)$ than to other marital statuses. Women with at least a high school degree are more likely to receive alimony than women without one. There are not enough households with $A > 0$ to obtain reliable estimates of coefficients for finer subdivisions of marital status or education.

The Bayesian bootstrap, described in Appendix B, samples from the observed positive alimony payment values. This implies conditional independence between A and other variables in the synthetic data, given $A > 0$. Data analyses on the full population suggest that this is a reasonable assumption. The Bayesian bootstrap ensures that the marginal distribution of positive alimony payments is accurately reproduced.

A.4. Child support payments

A logistic regression is fitted to predict whether or not a person receives positive child support payments. The logistic regression uses the predictors $X * R + \boldsymbol{G} + \boldsymbol{G}^2$, and additionally

$$(M = 5) + (M = 6) + (M = 7) + (E < 39) + (E > 42) + (\mathrm{HH} = 1) + (\mathrm{HY} > 0).$$

Positive child support payments are modelled with a linear regression with Gaussian errors of $\sqrt{C}$ on the predictors $X + \boldsymbol{G} + \boldsymbol{G}^2$, and additionally

$$(R = 1) + (M = 5) + (M = 6) + (M = 7) + (E < 39) + (E > 42) + (\mathrm{HH} = 1) + \mathbf{HY}.$$

The models for child support payments are determined primarily from the results of data analyses on the full population. These analyses suggest that child support payments are more common for divorced,

separated or single people than for people with other marital statuses. The quadratic function in age reflects the fact that the youngest and oldest people are less likely to receive child support than middle-aged people. The indicator $HH = 1$ captures the fact that household heads are more likely to receive child support payments than are other members. Finally, the strongest determinant of child support status is whether or not there is a child, $HY > 0$.

For units with $C > 0$, using $\sqrt{C}$ as the outcome makes the errors in the linear regression approximately Gaussian. Predicted values of $\sqrt{C}$ are squared to obtain the synthetic values of C.

A.5. Social security payments

Social security payments are modelled in multiple steps. First, a logistic regression is used to predict whether or not the household head has $S > 0$. It uses the predictors

$$\begin{aligned} &X + R + M + (E < 39) + (E = 43) + (E = 44) + (E = 45) + (E = 46) \\ &\quad + (29 < G < 40) + (39 < G < 55) + (54 < G < 60) + (G = 60) + (G = 61) \\ &\quad + (G = 62) + (G = 63) + (G = 64) + (G = 65) + (G = 66) + (G = 67) \\ &\quad + (G = 68) + (G = 69) + (G = 70) + (G > 70). \end{aligned}$$

Second, a logistic regression is used to predict whether those household heads with $S > 0$ have $S \geqslant 20000$. This logistic regression uses the predictors

$$X + (G < 62) + (M = 4) + (E < 39) + (E > 44).$$

For household heads who are imputed to have $S \geqslant 20000$, a Bayesian bootstrap of observed values of $S \geqslant 20000$ is used to impute S. For households that are imputed to have $0 < S < 20000$, the imputation model is a linear regression of $\sqrt{S}$ on the predictors

$$\begin{aligned} &X + R + (M = 4) + (M = 5) + (M = 6) + (M = 7) + (E < 39) + (42 < E < 45) \\ &\quad + (E > 44) + (29 < G < 40) + (39 < G < 55) + (54 < G < 60) + (G = 60) + (G = 61) \\ &\quad + (G = 62) + (G = 63) + (G = 64) + (G = 65) + (G = 66) + (G = 67) + (G = 68) \\ &\quad + (G = 69) + (G = 70) + (G > 70). \end{aligned}$$

For other household members, the same sequence of models is fitted using only people with $HH = 0$. These models also include an indicator variable for whether or not the household head has $S > 0$. This captures some of the within-household correlation in social security payments.

In the USA, eligibility for social security payments depends almost entirely on age. Currently, by law most people under age 62 years are not eligible for social security benefits, and most people over age 67 years do receive social security benefits. Exceptions include people who receive benefits when a spouse dies and people who have certain disabilities. The models include indicator variables for ages beyond 55 years because US employees generally retire from employment during these years, and analysts analysing social security benefits may desire finer detail at retirement ages.

Splitting positive social security payments into two models, one for $S < 20000$ and one for $S > 20000$, improves the overall model fit relative to one continuous model. This split is the result of trial and error, searching for models that produce predicted values of S that are similar in distribution to the observed social security values. Typically, only a few people have $S > 20000$, so a parsimonious model is necessary for the logistic regression to predict whether $S > 20000$. The Bayesian bootstrap used to impute large values of S preserves the distribution of large social security payments.

A.6. Income

Income is modelled as a sequence of multinomial logit and linear regressions. First, incomes are split into five categories:

(a) $I < 0$;
(b) $I = 0$;
(c) $0 < I < 6000$;
(d) $6000 \leqslant I < 205000$;
(e) $I \geqslant 205000$.

Using these categories as the outcome, a multinomial logit regression is fitted to the predictors

$$X + (R = 1) + \boldsymbol{G} + \boldsymbol{G}^2 + M + (\mathrm{HS} = 1) + (E < 39) + (39 < E < 43) + (E = 43) + (E = 44) + (E = 45) + (E = 46) + \log(A + S + C + 1).$$

For incomes in the range $6000 < I < 205\,000$, a linear regression with Gaussian errors is fitted, using $\log(I)$ as the outcome and the predictors

$$X * R + X * M + S * (\mathrm{HS} = 1) + (\mathrm{HS} = 2) + (\mathrm{HS} = 3) + (G < 18) + (17 < G < 25) + (24 < G < 30) + (29 < G < 35) + (34 < G < 40) + (39 < G < 45) + (44 < G < 50) + (49 < G < 55) + (54 < G < 60) + (59 < G < 65) + (64 < G < 70) + (69 < G < 75) + (E < 35) + (34 < E < 39) + (39 < E < 43) + (E = 43) + (E = 44) + (E = 45) + (E = 46) + \log(A + S + C + 1).$$

Bayesian bootstraps are used to generate incomes for households with income categories that are imputed to be $I < 0$, $0 < I < 6000$ or $I \geqslant 205\,000$.

A multinomial model is used so that negative and zero incomes can be easily imputed. Positive incomes are split at 6000 and 205 000 to help to fit the tails of income accurately. These split values are determined by trial and error, searching for models that produce predicted values of I that are similar in distribution to observed incomes. Households that are imputed in any particular category are constrained to have income in that category. The sets of predictors are chosen primarily by trial and error: they provide predicted incomes in line with the actual incomes while capturing many of the important relationships in the full population. The Bayesian bootstraps are used so that the distributions of the tails of synthetic I are like those from the observed data. Using bootstrapped values implies independence of I from other variables for households within these three categories. Data analyses on the full population suggest that this assumption is reasonable.

A.7. Property tax

A logistic regression is fitted to predict whether a household has $P > 0$. The logistic regression uses the predictors $X + R + \boldsymbol{G} + \boldsymbol{G}^2 + \boldsymbol{E} + (\mathrm{HY} > 0)$, and additionally

$$(I > 100\,000) + (90\,000 < I < 100\,000) + (80\,000 < I < 90\,000) + (70\,000 < I < 80\,000) + (60\,000 < I < 70\,001) + (50\,000 < I < 60\,001) + (40\,000 < I < 50\,001) + (30\,000 < I < 40\,001) + (20\,000 < I < 30\,001) + (10\,000 < I < 20\,001).$$

Positive property taxes are modelled from a linear regression with Gaussian errors of $\log(P)$ on the predictors $X + R + \boldsymbol{G} + \boldsymbol{G}^2 + M + \mathbf{HY} + (E < 39) + (E > 43)$ and additionally

$$(I > 100\,000) + (90\,000 < I < 100\,000) + (80\,000 < I < 90\,000) + (70\,000 < I < 80\,000) + (60\,000 < I < 70\,001) + (50\,000 < I < 60\,001) + (40\,000 < I < 50\,001) + (30\,000 < I < 40\,001) + (20\,000 < I < 30\,001) + (10\,000 < I < 20\,001).$$

Income is broken into categories to capture non-linearities in its relationship with P. The squared term for age captures the fact that the youngest and oldest people are less likely to own houses, as well as less likely to own expensive houses. Households with youths ($\mathrm{HY} > 0$) are more likely to own houses and therefore to have $P > 0$. The numerical **HY** is used to predict positive values of P because the physical dimensions of houses, which are positively correlated with P, typically increase with the number of youths in the household.

Appendix B: Imputation methods

For some parameters, draws from their actual posterior distributions require Monte Carlo simulation methods. These are difficult to monitor and computationally expensive to run over repeated simulations. To avoid excessive computing expense, parameters of some models are drawn from approximate rather than actual posterior predictive distributions. Whenever it is computationally convenient, e.g. for linear

regressions, values of the regression parameters are drawn from exact posterior distributions. The methods of drawing parameters are described below. To save additional notation, W_{obs} and W_{syn} are used to represent generically matrices of values of relevant predictors for the observed and synthetic units respectively.

For linear regressions with Gaussian errors, standard Bayesian posterior predictive distributions are used to draw new values $\mathbf{y}^*$ given predictors W_{syn} as follows.

(a) Draw a value σ^* of the regression variance σ^2 from a scaled inverse χ^2-distribution on $n-p$ degrees of freedom with scale parameter $(n-p)s_{\text{e}}^2$, where s_{e}^2 is the unbiased, ordinary least squares estimate of σ^2.
(b) Draw a value $\boldsymbol{\beta}^*$ of the vector of regression coefficients $\boldsymbol{\beta}$ from a multivariate normal distribution with mean $\hat{\boldsymbol{\beta}}$ and variance $(W_{\text{obs}}^{\text{T}} W_{\text{obs}})^{-1}\sigma^*$, where $\hat{\boldsymbol{\beta}}$ is the unbiased, ordinary least squares estimate of $\boldsymbol{\beta}$.
(c) Draw the relevant units' synthetic values $\mathbf{y}^*$ from $N(W_{\text{syn}}\boldsymbol{\beta}^*, \sigma^*\mathbf{I})$, where $\mathbf{I}$ is an identity matrix.

For multinomial logit regressions, an approximate Bayesian posterior predictive distribution is used to generate new multinomial values $\mathbf{y}^*$ given predictors W_{syn} as follows.

(a) Draw values $\boldsymbol{\alpha}^*$ of the multinomial logit regression coefficients $\boldsymbol{\alpha}$ from a multivariate normal distribution with mean $\hat{\boldsymbol{\alpha}}$ and variance $\hat{V}$, where $\hat{\boldsymbol{\alpha}}$ and $\hat{V}$ are respectively the maximum likelihood estimates of $\boldsymbol{\alpha}$ and its covariance matrix.
(b) Determine the probabilities π_{jk}^* for each multinomial category k and each unit j, where

$$\pi_{jk}^* = \frac{\exp(\mathbf{w}_{\text{syn},\,j}^{\text{T}}\boldsymbol{\alpha}_k^*)}{\sum_k \exp(\mathbf{w}_{\text{syn},\,j}^{\text{T}}\boldsymbol{\alpha}_k^*)}.$$

(c) Draw each synthetic unit's y_j^* using its corresponding $\pi_j^* = \{\pi_{jk}^*\}$.

Draws from binomial distributions in logistic regressions are simulated similarly, using only two categories in the steps above. The normal distribution approximation to the posterior distribution of $\boldsymbol{\alpha}$ simplifies parameter simulation considerably. For large observed data sample sizes n, this approximation should be reasonable. That said, draws from the true posterior predictive distributions of $\boldsymbol{\alpha}$ are preferable and are recommended in implementations of the synthetic data approach, particularly when n is not large.

Some synthetic values are simulated by using a Bayesian bootstrap. The Bayesian bootstrap (Rubin (1987), pages 123–124) is a nonparametric method of drawing from posterior predictive distributions. It is similar to the usual bootstrap, in that values of some $\mathbf{y}$ are drawn from a donor pool comprised of observed values. Let $\mathbf{y}_{\text{elig}}$ be the $n_0 \times 1$ vector of values of $\mathbf{y}_{\text{obs}}$ that make up the pool of potential donors for the bootstrap. The Bayesian bootstrap proceeds as follows.

(a) Draw $n_0 - 1$ uniform random numbers. Sort these numbers in ascending order. Label these ordered numbers as $a_0 = 0, a_1, a_2, \ldots, a_{n_0-1}, a_{n_0} = 1$.
(b) Draw n_{syn} uniform random numbers, $u_1, u_2, \ldots, u_j, \ldots, u_{n_{\text{syn}}}$. For each u, impute $\mathbf{y}_{\text{elig},j}$ when $a_{j-1} < u < a_j$.

Unlike the usual bootstrap, the Bayesian bootstrap leads to proper imputations (Rubin, 1987), which justifies its use for generating multiply imputed, synthetic data.

References

Abowd, J. M. and Woodcock, S. D. (2001) Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies* (eds P. Doyle, J. Lane, L. Zayatz and J. Theeuwes), pp. 215–277. Amsterdam: North-Holland.

Dandekar, R. A., Cohen, M. and Kirkendall, N. (2002a) Sensitive micro data protection using Latin hypercube sampling technique. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 117–125. Berlin: Springer.

Dandekar, R. A., Domingo-Ferrer, J. and Sebe, F. (2002b) LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 153–162. Berlin: Springer.

Duncan, G. T., Keller-McNulty, S. A. and Stokes, S. L. (2001) Disclosure risk vs. data utility: the R-U confidentiality map. *Technical Report*. Durham: US National Institute of Statistical Sciences.

Duncan, G. T. and Mukherjee, S. (2000) Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. *J. Am. Statist. Ass.*, **95**, 720–729.
Fienberg, S. E., Makov, U. E. and Steele, R. J. (1998) Disclosure limitation using perturbation and related methods for categorical data. *J. Off. Statist.*, **14**, 485–502.
Franconi, L. and Stander, J. (2002) A model-based method for disclosure limitation of business microdata. *Statistician*, **51**, 51–61.
Franconi, L. and Stander, J. (2003) Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statist. Comput.*, **13**, 295–306.
Fuller, W. A. (1993) Masking procedures for microdata disclosure limitation. *J. Off. Statist.*, **9**, 383–406.
General Accounting Office (2001) *Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information*. Washington DC: United States General Accounting Office.
Kennickell, A. B. (1997) Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques, 1997* (eds W. Alvey and B. Jamerson), pp. 248–267. Washington DC: National Academy Press.
Lavine, M. (1992) Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.*, **20**, 1222–1235.
Little, R. J. A. (1993) Statistical analysis of masked data. *J. Off. Statist.*, **9**, 407–426.
Liu, F. and Little, R. J. A. (2002) Selective multiple imputation of keys for statistical disclosure control in microdata. In *Proc. Joint Statistical Meet.*, pp. 2133–2138. Blacksburg: American Statistical Association.
Meng, X.-L. (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.*, **9**, 538–573.
Polettini, S. (2003) Maximum entropy simulation for microdata protection. *Statist. Comput.*, **13**, 307–320.
Polettini, S., Franconi, L. and Stander, J. (2002) Model-based disclosure protection. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 83–96. Berlin: Springer.
Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003) Multiple imputation for statistical disclosure limitation. *J. Off. Statist.*, **19**, 1–16.
Reiter, J. P. (2002) Satisfying disclosure restrictions with synthetic data sets. *J. Off. Statist.*, **18**, 531–544.
Reiter, J. P. (2003) Inference for partially synthetic public use microdata sets. *Surv. Methodol.*, to be published.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
Rubin, D. B. (1993) Discussion: Statistical disclosure limitation. *J. Off. Statist.*, **9**, 462–468.
Willenborg, L. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*. New York: Springer.
Yancey, W. E., Winkler, W. E. and Creecy, R. H. (2002) Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases* (ed. J. Domingo-Ferrer), pp. 135–152. Berlin: Springer.