The Creation and Use of the SIPP Synthetic Beta[*]

Gary Benedetto, Martha H. Stinson, John M. Abowd [†]

April 2013

1 Introduction

This paper reports on the creation of a partially synthetic Census Bureau data product called the SIPP Synthetic Beta (SSB). Our purpose is to inform users of the SSB about how the file was created and to provide an example of the application of data synthesis methods to those doing research in this area. We also hope to provide some guidance for other organizations which might be interested in creating their own synthetic data products.

We begin by providing a brief overview of how the SSB is created. We then turn to the details of our methodology, beginning with a short review of the literature that supplies the theory for data synthesis as a means of protecting confidential data. We follow with a more detailed description of how we applied this theory. We then explain how we tested the synthetic data for disclosure risk and provide guidance to researchers on how to use the SSB. We provide some short tests of the analytic validity of our latest release, SSB version 5.1, and give some examples of early uses of the SSB. We finish with a discussion of the challenges of creating useful synthetic data and an outline of plans for future development. Appendix A gives a short history of the creation of the SSB and describes the evolution of this product across different versions.

[*] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau, Cornell University, or any of the project sponsors.

[†] Abowd is the Edmund Ezra Day Professor of Economics at Cornell University, Distinguished Senior Research Fellow at the U.S. Census Bureau, Research Associate NBER, Research Affiliate CREST/INSEE, and Research Fellow IZA. Benedetto and Stinson are economists at the U.S. Census Bureau. Contacts: john.abowd@cornell.edu, gary.linus.benedetto@census.gov, martha.stinson@census.gov

2 Overview of the Creation of the SSB

The purpose of the SSB is to provide data from the Survey of Income and Program Participation data linked to administrative earnings and benefits to users outside Census secure facilities. From the beginning of the project, two overarching requirements have guided the decisions about the type of file to create. First, the file should contain micro-data in a format usable by researchers and others familiar with the structure and content of the regular SIPP public use files. Second, the file should stand alone and not be linkable to any of the existing SIPP public use products previously published by the Census Bureau. These two criteria led the Census Bureau to choose partially synthetic data as the primary disclosure avoidance method. The main purpose of this paper is to educate researchers about synthetic data, in particular how these data were created and how they should be used.

As the first step in this process, the Census Bureau created a standardized extract of variables from a set of SIPP panels and merged these extracts with individual administrative earnings and benefits records[1]. These extracts were combined and named the SIPP Gold Standard File (GSF). This file serves as the basis for the creation of the SSB. It establishes the metadata for each variable, determines the sample of people to be included, and serves as the source data for the modeling required to create the synthetic data.

The next step in the process was to handle missing data in the Gold Standard file. We used multiple imputation to create four completed data files that contained imputations for all missing values in the Gold Standard[2]. Finally, again using multiple imputation techniques, we created 16 synthetic data files that replaced all original values with imputations. After the creation of the synthetic data, we then tested for disclosure risk by attempting to link our synthetic data back to the Gold Standard. Even using some inside knowledge not available to a potential intruder, we were not able to reliably match synthetic records to the correct Gold Standard records[3].

The SSB has been extensively tested for analytic validity over the years as new versions have been released. Currently the Census Bureau offers outside researchers the option of having analyses done with the SSB validated using the completed Gold Standard files. The Census Bureau will release results from analyses done on the confidential data so that analysts can know what impact synthesis had on the data relationships they estimated. Feedback from these validation exercises, in turn, helps further the development of the synthesis

[1] Version 5.1 contains data from the 1990, 1991, 1992, 1993, 1996, 2001, and 2004 SIPP panels and the SSA Detailed Earnings Record (DER), Summary Earnings Record (SER), Master Beneficiary Record (MBR), and Supplemental Security Record (SSR).

[2] We did not change item-missing imputations done by standard Census processing of SIPP data. Rather we kept the original imputations and added imputations that handled missing data due to a household missing a wave of the survey and due to individuals missing administrative data.

[3] The probability that the two records deemed "most similar" by our matching strategy were in fact a true match was less than 2%. See Section 4 for full details on our disclosure testing.

process. For more information on using the SSB and doing validation work, please visit www.census.gov/sipp/synth_data.html.

3 Methodology

3.1 Review of Literature on Multiple Imputation

Protecting the identity of individuals whose personal and financial characteristics are released in a micro-data set has long been an important research in statistics. Since its launch in 1984, the public-use SIPP has relied exclusively on top-coding and cell suppression to handle disclosure issues. However the addition of many administrative variables to create the SIPP Gold Standard gave rise to the concern that these methods were no longer sufficient to protect the identify of SIPP respondents. Hence new methods were sought from the research literature and from the examples of other federal data sources. To understand the approach we adopted, we begin by guiding the reader through the development of multiple imputation theory and its subsequent application to data protection methods which came to be called data synthesis.

Rubin first proposed multiple imputation as a way to handle missing data problems. In his seminal book (Rubin 1982), he advocates applying any given imputation method multiple times to create many replacement values for missing data. This approach produces multiple copies of the data set, each copy having its missing values replaced with one of the sets of imputed values. The need for this arises from the fact that extra variability is introduced by the missing data. This variability needs to be taken into account or else the confidence intervals generated for statistics produced using the data will be too small, i.e. parameters will be determined to be significant too often. By generating multiple data sets or implicates, the user can run a standard analysis on each one and then calculate the within-implicate variance (standard variance measure) and the between-implicate variance (variance across the implicates). The total variance formula, discussed in detail in Section 5, then has two components which take account of the standard measure of variance and the variance introduced by the imputation.

The idea that imputation of missing data and creation of synthetic data are related comes from Rubin (1993) and Little (1993). Rubin's original idea was that multiple imputation could be used to fill in survey responses for the entire population of individuals from which the original survey sample had been drawn. In essence, for individuals not sampled by the survey, the survey variables were treated as missing and were multiply imputed. From this population with complete data, new synthetic samples would be created by drawing individuals from the population. The survey responses for these individuals could be released because they were not actual responses but rather multiplely-imputed responses. Little proposed imputation to replace original values as one of many possible mechanisms of disclosure protection.

Rubin's argument for using this method was that researchers using the data would not need special statistical software to analyze such data but rather could use standard methods and then could combine results across implicates using appropriate formulae. All the burden for modeling and creating the synthetic data fell on the data producer, who Rubin felt was most likely to have the necessary resources and expertise. At the same time no actual respondent-reported data would be released so survey response would be helped. Data intruders looking to identify individuals in public-use data products would shy away from synthetic data.

Rubin's original idea for data synthesis was very general and did not suggest a specific imputation method. Early work using data sets with very small numbers of variables, usually of the same kind, were modeled by specifying a distribution of the variables with missing data conditional on all the other observed values and some unknown parameters, which had a specific prior distribution. This model then produced a posterior predictive distribution from which draws were taken to replace the missing values. However explicit multivariate conditional models are difficult to make when the data are complex with many types of variables such as continuous, discrete, and categorical, and when restrictions on one variable are implied by another variable. Raghunathan et al. (2001) proposed a general purpose multivariate imputation procedure called sequential regression multivariate imputation (SRMI) which factors the joint conditional density into a series of conditional density functions where a single variable with missing data was modeled as conditional on other variables with and without missing data and a set of parameters. The imputation proceeds through all the variables with missing data, and as values are imputed, they are included as explanatory variables in the next round of imputation. The imputation process is completed for a certain number of rounds in order to allow all the variables to influence each other regardless of the order in which the data completion is done.

From these original ideas, the idea of partially-synthetic data has been developed. Unlike fully synthetic data, original sample members remain in the file. However their responses are replaced by values which are multiply-imputed. As described by Reiter and Raghunathan (2007), partially synthetic data sets look like data sets that have missing values replaced by multiple imputation methods but in fact the multiple imputation methods produce replacements for self-reported data. One early application of partially-synthetic data to protect confidentiality was the Survey of Consumer Finances, described in Kennickell (1997). Abowd and Woodcock (2001) synthesized an early prototype of linked employee-employer data. Today the Census Bureau releases two other partially synthetic data products in addition to the SSB. The first is the Longitudinal Business Database (LBD) which is described in Kinny et al. (2011) and the second is On The Map which is described in Machanavajjhala et al. (2008).

3.2 Data Synthesis and Completion Methods

3.2.1 Summary of synthetic data production

We now provide specific details about the process used to create our synthetic data product. The first step of the process was to multiply impute all missing data. Data were missing for SIPP respondents either because they missed a wave of the survey or lacked the necessary information to link to administrative records. We kept hot-deck imputations done during regular SIPP processing to complete item-missing values for household that responded in a given wave but failed to answer some questions. We call this first step "completing the data," because the result is data that contain all the original values plus imputed values when the original ones were missing.

In contrast to regular missing data, which we multiply impute, structurally missing data occur when an item is missing due to the logical structure of a set of variables in the survey or administrative record. For survey data, structurally missing values occur when the skip logic of the survey dictates that a question should not be asked because of the response given to a prior question. Administrative record data have a similar, albeit implicit, structure. Lack of participation in the formal labor market or SSA programs will produce structural zeros for earnings and benefits respectively. Structurally missing data were never completed (*i.e.*, imputed) because they do not represent missing information. We use the term "missing" to mean missing-to-be-completed and will explicitly describe any other data that are missing as structurally missing.

In completing data, we followed Raghunathan et al. (2001) and implemented an SRMI framework that estimated the joint conditional density as a series of conditional density functions, thus allowing us to model each variable separately, conditional on the previously modeled data. We chose from among three models for each variable with missing data: linear regression, logistic regression, and Bayesian Bootstrap. For administrative data, we imputed whether the person had earnings and benefits at a particular point in time and then if yes, we imputed the dollar amounts. Likewise for the SIPP variables, we preserved the logical relationships amongst variables by imposing restrictions on down-stream variables (called "child" variables) using values of up-stream variables (called "parent variables"). We describe the types of models and the specification of logical relationships among variables in more detail in Section 3.2.2.

The actual SRMI process of completing data is iterative. The goal of the first iteration is to fill in all missing values with starting values. We cycle through all the variables in the file, estimating models using originally non-missing cases for the dependent variable and a set of explanatory variables which contain no missing data. After a variable has been modeled and its missing values replaced, it becomes eligible to be an explanatory variable for the next modeled variable. As we progress through the variable list and the data set is updated, there will be fewer and fewer variables with missing values, and hence increasingly more cases available for model estimation. The end product of this process is a data set that contains completed administrative and SIPP variables.

After producing an initial set of values for the missing data, we move to the second iteration. As in the first iteration, only originally non-missing dependent variables are used in model estimation. However in this iteration, we can choose explanatory variables from all the variables in our list, not just those which had previously been modeled. The first variable to be modeled uses explanatory variables from the completed data that was the output of the first iteration. The second variable to be modeled uses the most up-to-date values for variable 1, *i.e.,* the values imputed in iteration 2, and the completed data from iteration 1 for every other variable. The sequential estimation progresses until the last variable, which uses imputed values from iteration 2 for all explanatory variables. In this manner, the modeling is always done with the most up-to-date imputed values available, allowing the modeling to improve itself over iterations. At the conclusion of this second step, another completed data set is generated which has updated values for all variables. There is no exact number of iterations dictated by SRMI theory. In creating SSB version 5.1, we did three iterations of data completion.

We impute multiple times, meaning that we run multiple, parallel iterative data completion processes. The data product that results is actually a set of files called the completed data implicates. Each implicate has an identical structure (same number of observations, variables, *etc.*) and contains identical data in cases where the information was originally non-missing. The separate SRMI processes are necessary because of the inter-related nature of the variables. Once a variable has been completed, its updated (i.e. imputed) value is used as a right-hand-side variable in the imputation process for other variables. Thus, in order to maintain internal consistency within an implicate file, each implicate must be generated separately. For version 5.1, we created four missing data implicates.

Once the data are completed and contain no missing data except for structurally missing items, the final step of synthesizing the data requires only one final iteration. We synthesize in the same manner as the first missing data iteration, namely, we build up the synthetic data as a series of conditional marginals, using only previously synthesized variables as explanatory variables. The main difference between synthesis and data completion is that *every* individual has *all* of his or her values imputed, variable by variable, conditional on the completed data. This means that for each variable, a model is estimated using all cases from the completed data that are in universe, regardless of whether they were originally non-missing or imputed in the data completion step. After estimating the model, we impute a value for each variable based upon the most up-to-date synthetic data. Hence while the synthetic variables are not used in the model estimation, they are used to impute other synthetic values in order to keep the synthetic data internally consistent.

Since there are multiple completed data implicates, there are multiple input files to the synthesizing process. For each completed data implicate, we run multiple, parallel synthesis processes which produce multiple separate sets of synthesized values. For version 5.1, we created four synthetic data implicates per completed data implicate for a total of 16 synthetic implicates.

3.2.2 Modeling details

In implementing an SRMI iteration, we made four decisions for each variable that was completed/synthesized. First, we chose what type of model to use (OLS, logistic, Bayesian bootstrap); second, we designated parent-child relationships among variables; third, we defined restrictions to be placed on the values of variables when necessary; fourth, we chose a set of grouping and conditioning variables to use in modeling. In this section we explain the three types of models and describe the process for the last three steps. The SSB version 5.1 code book lists specific modeling details for each variable.

Models of variables The first information the analyst must provide about a variable to be completed and synthesized is the model type. We used three major modeling techniques: normal linear regression (OLS), logistic, and Bayesian bootstrap. The purpose of the modeling step is to estimate a posterior predictive distribution (PPD) for each variable and then to take draws from this PPD to replace either missing values or original data, depending on whether we are completing missing data or synthesizing. The PPD is simply the probability distribution of the data we are trying to produce, conditional on the data we observe. More formally, the PPD for variable y_k is defined as:

$$\begin{aligned} PPD &= P(y_k \mid Y^m, X) = \int p(y_k \mid Y^m_{\sim k}, X, \theta) p(\theta \mid Y^m, X) d\theta \\ X &= \text{non-missing, non-modeled variables} \\ Y^m &= \text{completed data} \end{aligned}$$

We use linear regression models to estimate the PPD for continuous variables. In this case, the parameters, θ, are assumed have normal/inverted gamma distributions and the regression produces estimates of the mean and variance of these distributions, giving us $p(\theta \mid Y^m, X)$. We then use standard techniques to take a draw from the θ distribution to produce a set of parameters ($\beta's$ and σ^2) for predicting values. Using these parameters and the observed values of the other data elements provides us with $p(y_k \mid Y^m_{\sim k}, X, \theta)$, which we also assume is normal with mean βx and variance σ^2. A draw from this distribution is simply a predicted value from the linear regression, given the set of $\beta's$ and σ^2 that we drew earlier.

The basics of this method will seem familiar to most researchers. We estimate a relationship between the observed values of a dependent variable and a set of independent variables also found in the data. This relationship is characterized by a set of regression coefficients and the standard error of the equation and involves assumptions about the model form and the distribution of the model parameters. We use these estimated parameters to predict a value for individuals missing data or for all individuals in the case of synthetic data. The key insight is that the regression parameters are themselves random variables and as such must be sampled. This sampling of parameters replicates the underlying uncertainty due to estimating our model on a sample of data instead of a universe. By taking multiple draws from the regression parameter distribution, we provide data that allows users to take account of this uncertainty.

It is sometimes the case that the univariate distribution of the variable we are trying to synthesize, y_k, differs greatly from conditional normality. This situation will cause the distribution of the synthetic values to differ from that of the confidential values, an undesirable result. To handle these variables, we transform the confidential data so that they have an approximately normal distribution, estimate the posterior predictive model on the transformed data, and perform the inverse transformation on the imputed values. This process is described in detail in Benedetto and Woodcock (2006).

For binary discrete variables, the PPD is based on the asymptotic posterior distribution of the parameters of a logistic regression model. Otherwise the methods are the same as in the linear regression models. Finally for Bayesian bootstrap models, we define the PPD in a non-parametric way. We begin by selecting a set of n individuals who are eligible to be donors for either the missing or synthetic data. In a regular bootstrap, the probability of selecting any given individual to be a donor is $\frac{1}{n}$ and there is no uncertainty in what probability is assigned to a given observation. In contrast, in a Bayesian bootstrap, the probability of individual i being chosen as a donor is p_i, which is modeled from the sample data and is centered around $\frac{1}{n}$. The set of probabilities, p_1 to p_n is the non-parametric representation of the PPD. By not assigning equal probabilities to all donors, the Bayesian bootstrap accounts for the fact that the sample distribution may not be the same as the population distribution. Performing the Bayesian bootstrap multiple times allows users to estimate the uncertainty introduced by imputation and synthesis. See Rubin (1981) for more details on this method.

Parent-child relationships and constrained variables Next the analyst must provide information that appropriately accounts for explicit relationships among the original variables that need to be preserved in the synthetic data. We have developed two tools for handling these relationships.

Our first tool is to specify parent-child relationships. We define parent variables as those that restrict which observations of another variable are present and which observations are structurally missing. These parent-child relations formalize the skip patterns in the SIPP survey instrument and the logical dependencies in the administrative records. A parent variable determines the universe of observations that are in scope to estimate the model for the associated child variable and to receive an imputed value following the estimation. If the parent variable indicates that the child variable is structurally missing (out of the universe) for an individual, then this observation will not be included in the estimation nor will it receive an imputed value.

Our second tool for handling relationships among variables is to place restrictions or constraints on some variables. Constraints specify a minimum and maximum value that restricts the range of draws from the posterior predictive distribution for a given variable. Constraints are specific to an individual and we impose them by forcing the draw to be from the part of the PPD that satisfies the constraints.

Grouping and conditioning variables Finally, for every variable that we modeled, we chose both grouping variables and conditioning variables. Grouping variables define the sets of observations within which regressions will be run and conditioning, or explanatory, variables define what goes on the right-hand side of the model. All three of our model types use grouping variables but only the parametric models use conditioning variables.

We chose grouping variables so that each group met a minimum size requirement and at the same time contained people who were as similar as possible. Adding additional grouping variables is very costly in terms of computational time so we sought to make a parsimonious but effective list to use for group stratification. Each unique group, defined by the values of all the variables in the grouping list, has its own posterior predictive distribution. This is the equivalent of fully interacting every grouping variable with every conditioning variable. Conditioning variables are used so that within homogeneous groups, important relationships between the dependent variables and other variables on the file can be preserved.

Problems develop when the grouping variables produce sub-groups that are too small to estimate a statistically reliable PPD. We use the rule that the number of observations in any sub-group must be at least 15 times the number of conditioning variables or 1,000, whichever is greater. To implement this rule, we first created multiple lists of grouping variables and conditioning variables for each variable. Each set of grouping variables is defined by progressively fewer variables, with the intent of making larger groups of observations. As variables are dropped, they are added to the list of conditioning variables, which hence becomes progressively longer.

We begin with the complete set of grouping variables, form all possible sub-groups, and then check their sample sizes. Sub-groups that are too small are collapsed and then split into a new set of sub-groups, using the next shortest list of grouping variables. This process continues until all the sub-groups meet the minimum observation requirements or until the list of grouping variables provided by the analyst is exhausted, at which point all groups that are still too small are combined.

As with grouping variables, the initial selection of conditioning variables is dependent on the selection made by the analyst. However each time a set of candidate conditioning variables is included in the model for a particular dependent variable in a particular sub-group, a Bayesian variable selection process is used to reduce the variable list by eliminating variables that are deemed to have weak relationships with the dependent variable, as measured by the Bayes Information Criterion (BIC). This statistic estimates an odds ratio of the model with and without each explanatory variable and, based on critical values we choose, decides whether a particular explanatory variable belongs in the model.

4 Analysis of Disclosure Risk

4.1 General Methods

The link between administrative earnings, benefits, and SIPP data adds a significant amount of information to an already very detailed survey and warrants careful investigation of possible disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The creation of synthetic data is meant to mitigate those risks by preventing a link between these new public use files and the original SIPP public use files, which are already in the public domain[4].

We assess the risk of disclosure using the principle that a potential intruder would first try to re-identify the source record for a given synthetic data observation in the existing SIPP public use files. In order to test the effectiveness of the data synthesis in controlling disclosure risk, we used minimum distance matching to attempt to link one SSB implicate to the Gold Standard File[5]. Since the Gold Standard is built from the original SIPP public use files and our methods of creating this file are public, the Gold Standard variables are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the Gold Standard and the synthetic data represent potential disclosure risks. We describe our minimum distance methods in more detail in Section 4.2.

We assume that an intruder attempting to link SSB records to SIPP respondents would block (i.e. stratify) on our two pieces of unsynthesized SIPP information, gender and the spouse-link, and then attempt to link records within these blocks. Hence in our re-identification exercise, we also block on gender. To handle the marital-link, we create a wide-version of both the Gold Standard File and the synthetic data where a single record contains all the data for both members of a linked marriage. If there is no linked marriage, the record only contains data for the single individual. We then match at the couple-level in order to allow the combined synthetic data for both husbands and wives to be used in finding a matching pair in the original data[6].

Individuals from the GSF were only kept in the SSB if they were at least 15 years old at the beginning of their SIPP panel. We implemented this sample restriction in the synthetic data by first synthesizing all the records from the underlying GSF, and then using both the synthetic birthdate and the synthetic panel to determine who met the age criteria. The synthetic files after the sample restriction differed in size and were smaller than the GSF. Thus an intruder cannot tell from looking at the public use SIPP which respondents were dropped and which were kept. Together, the age cutoff and synthetic birthdates add an extra layer of uncertainty to any matching exercise performed by an intruder.

[4] We also note that SSB version 5.1 will not be linkable to SSB version 4.0 or 5.0.

[5] At this point we have only matched the first SSB implicate to the GSF. We would expect that the matching results for implicates 2-16 to be very similar to those for implicate 1.

[6] Co-habitating same-sex partners were not allowed to declare themselves married in the SIPP panels contained in SSB v5.1. Hence a married couple always has both a male and female.

In our matching exercise, however, we wished to be very conservative, and so we used the full set of observations in the first synthetic implicate (prior to the age cutoff) in all of our re-identification exercises. Hence our synthetic implicate file has the same sample size as the Gold Standard, and we know that a "true match" between the two files exists.

Importantly, simply linking a record in the SSB to a matching record in the public-use SIPP would be insufficient for an intruder to identify a SIPP respondent. Re-identification would also require the intruder to make a second link to some additional source that contained personal identifiable information such as names, addresses, telephone numbers, *etc.* Hence, the results from our matching process are a very conservative estimation of re-identification risk.

4.2 Distance Matching

Distance-based record linking is a common approach to estimating the risk of re-identification in micro data. For example, Domingo-Ferrer, Abowd, and Torra (2006) use distance-based methods to re-identify records on two synthetic micro-data samples. They find that distance-based metrics perform similarly to (if not better than) more commonly used probabilistic methods. Domingo-Ferrer, Torra, Mateo-Sanz, and Sebe (2006) conduct similar comparisons of distance-based and probabilistic record linking methods. This body of work suggests that distance-based methods provide reliable measures of re-identification risk.

The basic re-identification method we employed was to calculate the distance between a given record in the Gold Standard and every record in the synthetic implicate. The j closest records were then declared potential candidates for a match to the source record. In our analysis we considered $j = 3$. We began by sub-dividing the data in two stages. First, we split both the Gold Standard and the first synthetic implicate into groups based on the unsynthesized variables. In this case, marital status(married/single) and gender were the only two unsynthesized variables. We next split each blocking group into smaller segments of approximately 10,000 observations in order to decrease the processing time, which is quadratic in the size of the largest files compared. We performed the segment split on both the Gold Standard and synthetic file so that the correct match in the Gold Standard was always in the same block and segment of the synthetic data used for comparison. In other words, we forced the segmentation of the files to guarantee that the correct match could always be found in the block/segments being compared. The segmentation of the blocks used our prior knowledge of which records were actual matches and hence our matching results are conservative–overestimates as compared to a distance record link that could not segment the comparison files because the intruder did not have access to person identifiers that linked between the synthetic implicate and the Gold Standard. After splitting the data into blocking groups and segments, we then calculated the distance between a given Gold Standard record and every record in the synthetic file in its corresponding blocking group and segment using the set of matching variables listed in Table 1. For couples, we used the small set of variables that were common to both partners and then used both the husband and wife values for all other variables. For singles, we used the person's own values for every matching variable. The list includes the SIPP point-in-time variables and summary measures from the SIPP and SSA/IRS time series variables. The three closest records were then declared possible matches.

We used four distance metrics. Each metric is a special case of either Mahalanobis or Euclidian distance. The concept of Euclidean distance is fairly intuitive. Two variables measuring the same thing in different sources are compared and we determine how "close" they are. This measure is combined across many variables to create an overall distance measure. Mahalanobis distance is simply a different weighting scheme for combining the distance between many variables, using as weights the inverse of the variance/covariance matrix of the matching variables from both sources.

In order to formally define these distance metrics, we first define some notation. Let A and B represent the two data sets being matched. For our purposes, conceptualize the block and segment of the Gold Standard as the A file and the block and segment of the synthetic implicate as the B file. Denote α as the vector of matching variables from an observation in the A file and β as the analogue for the B file. Given this notation we define the distance between a given vector α in the A file and a given vector β in the B file as follows:

$$d(\alpha, \beta) = (\alpha - \beta)\prime[Var(A) + Var(B) - 2Cov(A, B)]^{-1}(\alpha - \beta)$$

We consider four specific cases of the general distance. In the first case we assume that the intruder can properly calculate the $Cov(A, B)$. We denote this distance $MAHA1$, and note that it is a true Mahalanobis distance; hence we expect that this distance measure will give us the highest match rates since it uses all of the available information, including the correct covariance structure of the errors in synthesizing all matching variables. In the second case, we assume that the $Cov(A, B) = 0$. This is equivalent to assuming that we do not know how to link the observations across the A and B files and cannot compute $Cov(A, B)$. A real intruder would not have access to $Cov(A, B)$. We denote the second distance $MAHA2$, and note that it is a "feasible" Mahalanobis distance. In the third case, we assume $[Var(A) + Var(B) - 2Cov(A, B)] = I$, where I is the identity matrix. We denote the third measure as $EUCL1$, which is a Euclidian distance with unstandardized inputs. For the fourth measure, we transform all of the matching variables in the A and B files to $N(0, 1)$ variables. Call the transformed files $\tilde{A}$ and $\tilde{B}$. We then calculate the distance using $[Var(\tilde{A}) + Var(\tilde{B}) - 2Cov(\tilde{A}, \tilde{B})] = I$. We denote this fourth metric $EUCL2$, and note that it is a standardized Euclidian distance.

For specific results on minimum distance matching for version 5.1, please see *DRBMemoSSBv5_ 1.pdf* posted at www.census.gov/sipp/synth_data.html.

5 Using the SSB

Many potential users may be concerned about how to begin using synthetic data and multiple implicate files. In this section we give some advice for using these data sets to perform analyses and provide the exact formulae for combining results from multiple implicates.

We suggest that users begin with one synthetic implicate and write code to prepare variables and verify the specification of statistical models for this single data set. Since all the synthetic implicates are identical in terms of file structure, number of records, variables names, *etc.*, any code that works on one implicate also works on the remaining implicates. Users can debug their models and, once they are satisfied with the programming specification, run the model on all 16 implicates. In this sense, synthetic data are no different from any other micro-data set. Analyses are run in exactly the same manner but are repeated multiple times. We recommend saving analysis results such as regression coefficients or summary statistics in a data set that can be manipulated on its own. This will be useful for combining results. We also recommend that users base all their statistical inferences by properly combining results from all the implicates. That is, we do not recommend that users conduct statistical specification searches on a single implicate and then estimate “final” standard errors with the proper formulae. The statistical inference theory that underlies partially synthetic data with multiple imputation relies on the multiple analyses, conducted on independently drawn implicates, to reflect the model uncertainty inherent in the original confidential data.

Any statistic of interest to a researcher can be calculated from the synthetic data by calculating it once per synthetic implicate and then averaging across the 16 implicates. If the researcher wants to know the mean of variable x, he or she should calculate the mean of x in each of the 16 implicates and then calculate the simple average of these 16 separate means to get one grand mean. If the researcher wants to know the variance of x, he or she should follow the same procedure: calculate the variance in each implicate and then calculate the simple average of these 16 statistics to get one grand variance. Point estimates for any statistic of interest from regression results to moments or percentiles of a distribution can be obtained in this manner. In the standard combining formulae, every implicate is equally weighted, so simple averaging is all that is required. Formally, for a statistic $q^{(\ell,k)}$ calculated from the k^{th} synthetic implicate created from the ℓ^{th} completed data implicate, a point estimate is created by averaging across the r synthetic implicates associated with each completed data implicate and then across all m completed data implicates using the following formula:

$$\text{average across synthetic and missing data implicates: } \bar{q}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{mr}$$

The calculation of the estimated total variance of a statistic of interest, from which one might compute a confidence interval or test statistic, is more complicated but still can be performed with standard software. In addition to the statistic of interest, the user should save the estimated sampling variance of this statistic for each of the mr synthetic implicates. For example, if calculating the mean of x, the user should calculate the sampling variance of the mean of x for each implicate.[7] The within-implicate sampling variances are then averaged to estimate the average within-implicate variance, the first component of the total variance. Thus if $u^{(\ell,k)}$ is the variance of $q^{(\ell,k)}$, then the within-implicate variance is formally defined as:

$$\begin{aligned} & \text{average variance across synthetic and missing data implicates} \\ \bar{u}_M \quad = \quad & \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{u^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{u}^{(\ell)}}{m} \end{aligned}$$

The next piece of the total variance formula is the between-synthetic-data-implicate variance which quantifies the variation introduced by differences between synthetic implicates that were generated from the same missing data implicate, *i.e.,* deviations of the synthetic implicate from the average across all synthetic implicates generated from the same completed data implicate, $q^{(\ell,k)} - \bar{q}^{(\ell)}$, defined as:

$$\begin{aligned} & \text{average across the synthetic implicates} \\ \bar{q}^{(\ell)} \quad = \quad & \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{r} \\ & \text{variance across synthetic implicates} \\ b^{(\ell)} \quad = \quad & \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{r-1} \\ & \text{average variance across synthetic implicates} \\ b_M \quad = \quad & \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{m\,(r-1)} = \sum_{\ell=1}^{m} \frac{b^{(\ell)}}{m} \end{aligned}$$

To calculate b_M, the user first calculates the variance of the statistic across the four r implicates associated with a particular m implicate. There will be m of these variances: one per completed data implicate. These m variances are then averaged to give the overall between-synthetic-data-implicate variance.

The final piece of the total variance is the between-missing-data-implicate variance:

$$\text{variance across missing data implicates:} \quad B_M = \sum_{\ell=1}^{m} \frac{\left(\bar{q}^{(\ell)} - \bar{q}_M\right)\left(\bar{q}^{(\ell)} - \bar{q}_M\right)'}{m-1}.$$

[7] The reader is cautioned to be certain to perform all calculations on variances and not standard deviations. To compute a standard deviation or standard error, the square root operation should be peformed on the total variance that has been computed by combining all of the component variances appropriately.

The user calculates the mean of the statistic of interest for all the synthetic implicates associated with a particular completed data implicate. Again, there will be m of these means. The between m implicate variance, B_M, is the variance of these m means.

The variance pieces are then combined to create the total variance for the statistic using the following formula:

$$\text{total variance:}\ \ T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{b_M}{r} + \bar{u}_M.$$

When m and r are moderate and the estimator $\bar{q}_M$ is univariate, inference is based on $(\bar{q}_M - Q) \sim t_{\nu_M}(0, T_M)$ where the degrees of freedom ν_M are defined as

$$\text{degrees of freedom:}\ \ \nu_M = \frac{1}{\left(\frac{\left(\left(1+\frac{1}{m}\right)B_M\right)^2}{(m-1)T_M^2} + \frac{(b_M/r)^2}{m(r-1)T_M^2}\right)}$$

Proofs and details can be found in Reiter (2004). In the case that the total variance becomes negative, we recommend not subtracting the between-synthetic-data-implicate variance when calculating the total variance. The confidence interval can be calculated using the asymptotic assumption of normality instead of the finite sample $t-$distribution.

When presenting research results, users should not report the results from a single synthetic implicate. This is not an accurate representation of either the point estimates or their associated variances. This is especially important when comparing synthetic and completed data in order to determine analytic validity. No synthetic implicate can be judged for accuracy as a stand-alone file. It must be considered in conjunction with the other synthetic data sets. Likewise, all implicates of the completed data must be used together in order to create a comparison basis. The formulae for combining completed data implicates are similar to those for combining synthetic implicates and are as follows:

$$\text{average across implicates:}\ \ \bar{q}_m = \sum_{\ell=1}^{m} \frac{q^{(\ell)}}{m}.$$

$$\text{variance across implicates:}\ \ b_m = \sum_{\ell=1}^{m} \frac{\left(q^{(\ell)} - \bar{q}_m\right)\left(q^{(\ell)} - \bar{q}_m\right)'}{m-1}$$

$$\text{variance on each implicate file:}\ \ u^{(\ell)} = u\left(D^{(\ell)}\right)$$

$$\text{average variance across implicates:}\ \ \bar{u}_m = \sum_{\ell=1}^{m} \frac{u^{(\ell)}}{m}.$$

$$\text{total variance:}\ \ T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

$$\text{degrees of freedom:}\ \ \nu_m = (m-1)\left(1 + \frac{\bar{u}_m}{\left(1 + \frac{1}{m}\right) b_m}\right)^2$$

6 Analytic Validity

Many potential SSB users are concerned about the analytic validity of this data product and ask whether they will get the same answers using the synthetic data as they would using the internal confidential data. How the synthetic data compare to the confidential data typically depends on the research question and the sample of individuals chosen. Due to the experimental nature of the SSB and to faciliate further development of the synthesis process, Census will conduct a validation exercise for any researcher who submits error-free programs via the Cornell Virtual RDC Synthetic Data Server (SDS). After review of the confidential results by authorized Census employees, disclosable results will be released to the researcher for use in papers and publications. In this way, researchers can have confidence that they will be able to identify any differences in results due to synthetic data. At the same time, Census researchers can track the performance of the SSB and make improvements to the modeling process that enhance analytic validity.

7 Challenges and Future Research

The Census Bureau envisions the SSB as a constantly evolving data product. Because it provides researchers with access to (synthetic) administrative data without requiring special permission or use of a secure Census computing environment, demand continues to grow. Many researchers request additional SIPP variables. Unfortunately the synthesis process is long and complicated enough that producing new versions has only been possible every 2-3 years. This has made meeting researcher demand for new variables and new SIPP panels difficult.

In 2014, the SIPP will be conducted using a completely re-designed survey instrument. Interviews will happen only once a year and the format of the data will be quite a bit different. While much of the content will remain the same, assimilating the 2014 panel into the GSF will be challenging. The SSB development team is currently considering whether a separate GSF file will be required for panels beginning in or after 2014.

Two major areas of current SSB research are in progress in the Census Bureau. The first involves developing the link between parents and children and testing the research value of this link and considering the disclosure issues surrounding this link. Currently version 5.1 does not link any family members except spouses. The addition of a link between parents and children and consequently siblings would perhaps have to be synthesized in order to protect confidentiality. There are currently no existing methods for doing this, and hence this work will represent new research in the field of synthetic data creation.

The second area of research involves the creation of a job-level file for SSB respondents that would link individuals to their employers over time and would provide information such as an industry and firm size history, as well as earnings by employer. SSB staff have created the basic structure of this person-employer match file using the administrative earnings records and are now working on integrating SIPP job reports using name and address linking techniques. The administrative data will add more historical firm-level information to the relatively short employment history collected by the survey, whereas the SIPP will add more detail about labor supply to the jobs captured by both the survey and the administrative data. The release of an employee-employer match file will also present challenges, of the same nature as family links but even more complicated because of the number of employers per individual.

For now, the full family and employer links are being created as part of the next version of the GSF but will most likely not be released as part of the next version of the SSB. Rather some summary measures such as total number of employers, industry of main employers, earnings of parents attached to the records of their children will most likely be employed while Census continues to research methods for protecting confidential linked data.

In spite of the challenges of creating synthetic data, users are increasingly finding the SSB to be a useful product that allows access to data that have previously been unavailable to non-government researchers. The continued development and availability of this data product depends in large part on the successful interaction between the government and the research community. As researchers provide feedback on how effectively they can use the SSB for research and as statistical research on data synthesis methods progresses, the SSB will continue to expand and improve, covering more topics with more reliable data.

8 Bibliography

Abowd, J. M., Woodcook, S.D. (2001), "Disclosure Limitation in Longitudinal Linked Data," *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, L. Zayatz, J. Theeuwes (eds), Amsterdam: North Holland, 215-277.

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

Kennickell, A. (1997), "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," Survey of Consumer Finances Working Paper.

Kinney, S.K., Reiter, J.P., Reznek, A. P., Miranda, J., Jarmin, R., Abowd, J.M. (2011), "Towards Unrestricted Public use Business Microdata: The Synthetic Longitudinal Business Database," *International Statistical Review*, 79 (3), 362-384.

Machanavajjhala, A., Kifer, D., Abowd, J.M., Gehrke, J., Vilhuber, L. (2008), "Privacy: Theory Meets Practice on the Map," *International Conference on Data Engineering (ICDE)*, 277-286.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models," *Survey Methodology*, 27, 85-96.

Raghunathan, T.E., Reiter, J.P. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102 (480), 1462-1471.

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

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

Rubin, D. B. (1993), "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics*, 9, 462-468.

9 Appendix A: History of the SSB

In February 2001, a temporary U.S. Treasury Regulation went into effect that allowed the U.S. Census Bureau to obtain administrative W-2 earnings data for certain survey respondents from the Social Security Administration (SSA) and the Internal Revenue Service (IRS) for the purpose of improving core Census Bureau data products[8]. One of the first primary goals was to create a new public use file that linked existing public-use survey data from the Survey of Income and Program Participation with the W-2 data and administrative benefits data maintained by SSA. The creation of this new product was a joint effort of Census, IRS, and SSA, with all three agencies contributing data and statistical expertise and Census and SSA providing funding.

In consultation with outside researchers and the Congressional Budget Office (CBO), the Census Bureau created a standardized extract of variables from five SIPP panels (1990, 1991, 1992, 1993, and 1996) and merged these extracts with individual administrative earnings and benefits records. These extracts were then combined to create the first version of the Gold Standard in 2002. The Census Bureau produced the first synthetic version of these data in late fall 2003, and called it the SIPP/SSA/IRS Public Use File version 1.0. However this file was always viewed as preliminary and was never released to the public. Three other preliminary public use files were created: version 2.0 (fall 2004), version 3.0 (December 2005), and version 3.1 (June 2006). The Census Bureau completed work on version 4.0 in December 2006 and this version was released to the public in the spring of 2007.

SSB v4.0 contained the following unsynthesized variables: gender, marital status at time of wave 2, link to spouse (if married), type of OASDI benefit at time of initial claim and type of OASDI benefit in the year 2000. It did not contain any indicator for SIPP panel and so all SIPP respondents were required to have the same data present, regardless of their source panel. This design decision meant that large amounts of missing data had to be completed for respondents in years when they were not surveyed. For example, total income for SIPP respondents in the 1990 panel had to be imputed from mid-1992 onward because the 1990 panel ended part-way through 1992. SSB v4.0 also contained

[8] In February 2003, the temporary Treasury Regulation became final (see *Federal Register*, Vol. 68, No. 13 Tuesday, January 21, 2003, Rules and Regulations, pp. 2691-5).

a weight, meant to make the full set of respondents age 15 and older from all five panels representative of the civilian, non-institutionalized national population in the year 2000.

After extensive analytic validity testing by Census, SSA, and outside researchers, some design changes were made and version 5.0 was created. This version added the 2001 and 2004 SIPP panels and an indicator for the source panel was included in the data. This decision was made in order to prevent the necessity of imputing so much missing data. The OASDI benefit variables were expanded and SSI benefit variables were added. Version 5.0 had the same set of unsynthesized variables as version 4.0 but it did not contain the cross-panel year 2000 weight. This weight had not been successful in re-producing population statistics and efforts to correct it were postponed until another version. In order to speed the release of new data, time-varying SIPP variables were left off when SSB version 5.0 was released to the public in December 2010.

Version 5.1 uses the same SIPP panels as version 5.0 but adds a substantial number of SIPP variables, in particular, ones which vary over time. This version is being released in May 2013. Version 5.1 also has fewer unsynthesized variables, with only gender and spouse link being unsynthesized. We have also made significant improvements to our modeling of the earnings tax data by first cleaning the underlying data in order to prevent administrative data error from skewing our synthesis process. Version 5.1 is also the first SSB to contain geography (state at time of interview).