https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html

The Creation and Use of the SIPP Synthetic Beta v7.0*

Gary Benedetto, Jordan C. Stanley, and Evan Totty†

November 2018

1 Introduction

This paper reports on the creation of a fully synthetic Census Bureau data product called the SIPP Synthetic Beta (SSB). It serves as an update to a previous paper, Benedetto, Stinson, and Abowd (2013), which described version 5.0 of the 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 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.

2 Overview of the Creation of the SSB

The purpose of the SSB is to provide to researchers outside Census-secure facilities data from the Survey of Income and Program Participation linked to administrative records pertaining to earnings and benefits. From the beginning of the project, two over-arching requirements have guided the decisions

*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 or any of the project sponsors.

†Benedetto, Stanley, and Totty are economists at the U.S. Census Bureau. Contacts: gary.linus.benedetto@census.gov, jordan.c.stanley@census.gov, evan.scott.totty@census.gov

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 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.

For version 7.0, we synthesize four implicates from the GSF. Unlike in past versions of the SSB, every variable is now synthesized as is the missing data pattern. The synthesis process will assign missing values with a distinction between structurally missing and non-structurally missing (i.e., missing-to-be-replaced) data.[2] It is then up to the user to determine how they wish to address these missing-to-be-replaced values.[3] 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.[4]

Over time, the SSB has been extensively tested for analytic validity 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 Gold Standard File. If the requested output passes disclosure review, the Census Bureau will release results from analyses done on these 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 process. For more information on using the SSB and doing validation work, please visit https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html.

[1] Version 7.0 contains data from the 1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, and 2008 SIPP panels and the SSA Detailed Earnings Record (DER), Summary Earnings Record (SER), Master Beneficiary Record (MBR), Supplemental Security Record (SSR), and Payment History Update System (PHUS).

[2] For version 7.0, we maintain in the GSF non-missing values for SIPP-based variables only if the value was either logically imputed or not imputed. Such determinations are made based on SIPP allocation flags. If an allocation flag was not available for a particular variable, we use the given SIPP value as we cannot distinguish between non-imputed and imputed values.

[3] See Section 5.2 for greater details.

[4] See Section 4 for full details on our disclosure testing.

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 topic in statistics. Since its launch in 1984, the public-use SIPP has relied exclusively on top-coding and cell suppression to handle disclosure issues. The addition of many administrative variables to create the SIPP Gold Standard File raised concern that these methods were no longer sufficient to protect the identify of SIPP respondents. As a result, new methods were sought from the research literature as well as the examples of other federal data sources. To help the reader understand the approach we adopted, we begin by describing 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 with each copy having its missing values replaced with one of the sets of imputed values. The extra variability introduced by the missing data 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 then has these two components which take account of the standard measure of variance and the variance introduced by the imputation.[5]

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 then multiply imputed. From this population with complete data, new synthetic samples would be created by drawing individuals from the population. The multiply imputed survey responses for these individuals could be released because they were not actual 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. Rather, they could use standard methods and then 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

[5] We provide a detailed discussion of the total variance formula in Section 5 of this paper

resources and expertise. At the same time, survey response would theoretically improve since no actual respondent-reported data would be released. Further, data intruders looking to identify individuals in public-use data products could be expected to 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 modeled data sets with very small numbers of typically same-kind variables. The distribution of the variables with missing data was specificed to be 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 (e.g. continuous, discrete, or categorical), as well as 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). SRMI 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 partially synthetic data products in addition to the now fully synthetic 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 create a "snapshot" of the internal SIPP GSF. This file only includes variables that go onto the corre-

sponding version of the SSB. As the SIPP GSF changes or grows internally, the snapshot for a given version of the SSB remains unchanged, so that analysis on the SSB and its snapshot remain comparable.

There are two types of missing data in the snapshot: missing-to-be-replaced values and structurally missing values. In contrast to missing-to-be-replaced data, 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. In this paper, we use the term "missing" to mean missing-to-be-replaced and will explicitly describe any other data that are missing as structurally missing. For the SIPP variables, we preserved any logical relationships among 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.

From the GSF snapshot, we synthesize four distinct files by building up the synthetic data as a series of conditional marginals, using only previously synthesized variables as explanatory variables. 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. Note that in version 7.0, we model missing values, making a distinction in the data between structurally missing values and non-structurally missing values. This change from past versions grants the data user greater freedom of choice in how to handle missing values. Potential options for addressing missing values are described in Section 5.2.

3.2.2 Modeling details

We model the joint distribution of all the variables in the snapshot as a sequence of conditional marginals.

$$p(Y_1, ..., Y_K) = p(Y_1)p(Y_2|Y_1)...p(Y_K|Y_1, Y_2, ..., Y_{K-1})$$

In implementing this sequential regression approach, we made four decisions for each variable that was 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. Indicators for missingness were treated as variables to be modeled and synthesized in this sequence, and their synthesized values then determined the universe for the

regular variables downstream. The SSB version 7.0 codebook lists some specific modeling details for each variable.

Models of variables The first information the analyst must provide about a variable to be synthesized is the model type. We used three major modeling techniques: normal linear regression (OLS), logistic regression, and Bayesian bootstrap. The purpose of the modeling step is to estimate a posterior predictive distribution (PPD) for each variable and then take draws from this PPD to replace the values. 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:

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

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 from 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 is undesirable as it will cause the distribution of the synthetic values to differ from that of the confidential values. 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.

Synthesis of Familial Linkages Variables providing record identifiers indicating family relationships between respondents cannot be synthesized with the methods described above, but they can potentially reveal private information. In previous versions of the SSB, we left the first observed spousal linkage between spouses in the SIPP unsynthesized, and we did not provide any parent-child linkages. We did this by making a wide file of male variables and female variables, and for linked couples, all of these columns were in scope.[6] This meant synthesizing twice as many variables in order to preserve correlations between spousal variables, and yet we still often observed quite a bit of decay in these correlations in the synthetic data. Expanding this method to parent-child links would mean expanding the width of the file by at least a factor equivalent to the maximum number of children of a parent in addition to many more practical problems. To get around these issues, our approach for version 7.0 was to synthesize all the columns of a person-level file and then randomly link rows as spouses or parent-child in a way intended to preserve correlation between variables of family members.

To synthesize familial linkages between synthetic records, we developed a new approach that is similar in concept to predictive mean matching. For ease of explanation, we will describe the process for spouses, but the same process is used for matching children to mothers. The set of potential wives and husbands is determined from the sequential regression synthesis of the indicator for whether a record is a member of a linked couple in the survey.

Consider a set of variables for potential wives, $x_1 - x_{k_w}$, and a set of variables for potential husbands, $y_1 - y_{k_h}$ (in theory, these can be different variables, but for spouses, these were the same). Using the same non-parametric transform

[6] Co-habitating same-sex partners were not allowed to declare themselves married in the SIPP panels contained in SSB v7.0. Hence a married couple and any associated spousal linkage always have both a male and female component.

used in the linear regression, we transform all of these variables to approximately standard normal distributions. Call the vector of transformed variables, $[\tilde{x}_1 - \tilde{x}_{k_w}\ \tilde{y}_1 - \tilde{y}_{k_h}]$, which, for spouses, are approximately distributed as a multivariate normal with mean vector, $\mu = [\mu_w\ \mu_h]$, and covariance matrix, $\Sigma = \left[\begin{smallmatrix} \Sigma_{ww} & \Sigma_{wh} \\ \Sigma_{hw} & \Sigma_{hh} \end{smallmatrix}\right]$. From the observed spouses in the internal data, we can estimate μ and Σ as $\hat{\mu}$ and $\hat{\Sigma}$. We do this on a Bayesian Bootstrap of the internal data so as to account for sample uncertainty and follow posterior predictive sampling.

We then take the synthetic values of $\tilde{x}_1 - \tilde{x}_{k_w}$ for our set of potential wives in the synthetic data, and draw for them candidate husband values from the conditional multivariate normal distribution of $\tilde{x}_1 - \tilde{x}_{k_h}$ given $\tilde{x}_1 - \tilde{x}_{k_w}$. Finally, we go through the wife set in random order, and assign each wife the husband from the potential husband set whose synthetic values of $\tilde{x}_1 - \tilde{x}_{k_h}$ are closest (determined by standardized Euclidean distance) to the candidate husband values. Once a husband is assigned to a wife, that pair is removed from the potential sets of husbands and wives. This continues until every potential wife has a husband, or until the set of potential husbands is empty. For the case of linking children to mothers, once a child has been assigned to a mother, that child is removed from the set of potential children, but the mother is not removed until she has been assigned a number of children on the survey roster which was synthesized in the sequential regression step (or until the potential set of children is empty).

4 Analysis of Disclosure Risk

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.[7]

While the type of data synthesis used for this product is not a quantifiable privacy technique, we have attempted to measure the privacy protection and data utility of the synthetic data. We have a measure of data utility (see section 4.1 below) and a measure of privacy protection (see section 4.2). Given our posterior predictive sampling method of data synthesis, we cannot easily vary the level of noise introduced through synthesis, so we cannot generate the theoretical graph of the trade-off of privacy protection (noise injection) and data utility. We will, however, use our measures of privacy protection and data utility to try to give a sense of where the synthetic data lies in this graph.

[7] We also note that SSB version 7.0 will not be linkable to past SSB versions.

4.1 Data Utility Measure

For our measure of data utility, we stack the internal data and the synthetic data and make an indicator variable which equals 0 if the record is internal, and 1 if the record is synthetic. We then run a logistic regression of this indicator on all the variables in the file (sometimes making summaries of groups of variables). Then we calculate the mean squared error (MSE) of the predicted probability from 0.5 and call this pMSE. The optimal data utility would lead to a predicted probability of 0.5 for all records (internal or synthetic) and a MSE of 0. If we were able to perfectly predict the source with certainty, then the squared error would always be 0.25 ($(1-0.5)^2$ for synthetic records and $(0-0.5)^2$ for internal records). Therefore, the worst case MSE is 0.25. To make our final data utility measure go from 0 (worst data utility) to 1 (optimal data utility), we set it equal to 1-($\frac{pMSE}{0.25}$). This measure is averaged across the 4 synthetic implicates to give us a single value.

4.2 Privacy Protection Measure

The most basic thing we can do to check privacy protection is to find the synthetic record with the minimum standardized Euclidean distance from each internal record and check to see if the underlying record was the same. Given that we have synthesized every variable, it is perhaps not surprising that this kind of re-identification almost always fails. In fact, in three of the four synthetic implicates there was only one match, and in the other there were zero matches. This is what we would expect by sheer random chance that $\frac{1}{N}$ records would happen to line up perfectly.

We also are concerned with what the closest synthetic record might tell an intruder even if that record is not the same underlying record. Using the standardized Euclidean distance to find the closest synthetic record to each internal record, we want to see how much that synthetic record tells us about the values of the variables on the internal record. For each variable we calculate the root mean squared error (RMSE) of the nearest synthetic record's values to the internal values. We standardize this measure by dividing the RMSE by the sample standard deviation (STDDEV) of that variable on the internal file. We then average this across all the variables, to get a grand average of RMSE/STDDEV for the given synthetic file. Finally, we average this across the 4 synthetic implicates to get our measure of privacy protection. The worst possible protection would result in this measure being 0 (RMSE=0 implies the synthetic values of the nearest synthetic record are exactly equal to the internal record for all variables and all records). If the protection was so strong that we know nothing more about an internal record from its nearest synthetic record than we would know from the overall distribution of all the variables, then the RMSE would equal the STDDEV for every variable, and our final measure would be 1. So while, in theory, this measure could be greater than 1, in practice we can expect this measure to range from 0 being the worst protection to 1 being the best protection.

Figure 1 shows our measure of privacy protection and data utility plotted on the theoretical graph of the optimal trade-off of privacy and utility. To give the reader a better intuition for what this level of privacy means in practice, Tables 1 and 2 show the RMSE of the synthetic value of the nearest synthetic record from each internal record by quantile bins for a couple of sensitive variables. Table 1 shows this for total net worth, which was directly conditioned on in the nearest neighbor match. Table 2 shows this for total FICA covered earnings in 2010 from the Detailed Earnings Record, which was not directly conditioned on in the nearest neighbor match (for the longitudinal arrays we used summaries of the arrays for dimension reduction). As a result, this should give a sense of what an intruder with a certain set of information on a respondent can learn about a variable for that respondent outside of that set. We see in Table 1 that even for a variable with a ratio of RMSE to STDDEV equal to only 0.285 (calculated using the natural log of total net worth as the matching variable), the RMSE for a sensitive variable like total net worth on its original scale is very large, increasingly so as we get into higher parts of the distribution. Therefore, even for outliers, the nearest synthetic record tells the intruder very little. Table 2 shows us that if an intruder is trying to find out about a variable that he does not already have when attempting a nearest neighbor attack, the uncertainty is so high that the RMSE is actually much larger than the overall STDDEV of that variable. Further, if one is most concerned about outliers, the RMSE gets larger in the higher parts of the distribution as before.

5 Using the SSB

5.1 Working with the Data

Many potential users may be concerned about how to begin using synthetic data. 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 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 multiple synthetic data files and 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 all implicates. If the researcher wants to know the mean of variable x, they should calculate the mean of x in each implicate and then calculate the simple average of these separate means across implicates to get one grand mean. If the researcher wants to know the variance of x, they should follow the same procedure: calculate the variance in each implicate and then calculate the simple average of these statistics across implicates 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. 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.

The combining formulae depend on how the SSB user chooses to handle the missing values in the synthetic files. If the user chooses to ignore the missing values and perform listwise deletion (also known as complete case analysis) or pairwise deletion (available case analysis), then the combination formulae are as follows, based on the $Y_{inc} = Y_{obs}$ case in Reiter (2004):

$$\bar{q}_r = \sum_{l=1}^{r} \frac{q^{(l)}}{r} \tag{1}$$

$$b_r = \sum_{l=1}^{r} \frac{(q^{(l)} - \bar{q}_r)^2}{(r-1)} \tag{2}$$

$$\bar{u}_r = \sum_{l=1}^{r} \frac{u^{(l)}}{r} \tag{3}$$

and

$$T_r = (\frac{b_r}{r}) + \bar{u}_r \tag{4}$$

$$v_r = (r-1)\left(1 + \frac{\bar{u}_r}{(b_r/r)}\right)^2 \tag{5}$$

Where q is our estimand, r is the number of synthetic replicates, $q^{(l)}$ is the estimate of q on replicate l, and $u^{(l)}$ is the variance of $q^{(l)}$. We use these to calculate our final estimate of q to be $\bar{q}_r$, which is student-T distributed with variance, T_r, and degrees of freedom, v_r. We see that the variance is a combination of the variance across replicates of the point estimate in each replicate, b_r, and the average across replicates of the variance in each replicate. Thus, in addition to the statistic of interest ($q^{(l)}$), the user should save the estimated sampling variance of this statistic for each of the synthetic implicates ($u^{(l)}$). For example, if calculating the mean of x, the user should also calculate the sampling variance of the mean of x for each implicate.[8]

While the inclusion of missing values in SSB version 7.0 allows users to perform listwise or pairwise deletion, we recommend that users handle missing data with the multiple imputation approach described earlier. Listwise or pairwise deletion assumes missing values are missing completely at random (MCAR), which is a very strong assumption that is likely to fail in most applications, thus leading to biased estimates. Multiple imputation relaxes this assumption and also accounts for missing data uncertainty. In order to assist users with how to address missing data and sample uncertainty, we have developed programs for SSB users that perform different solutions including multiple imputation. These programs and their methods are described in the next two sections.

Multiple imputation changes the combination formulae. If we were to multiply impute missing data for synthetic replicate, l, then for that replicate, we would have:

$$\bar{q}^{(l)} = \sum_{i=1}^{m} \frac{q_i^{(l)}}{m}$$

$$b^{(l)} = \sum_{i=1}^{m} \frac{(q_i^{(l)} - \bar{q}^{(l)})^2}{(m-1)}$$

$$\bar{u}^{(l)} = \sum_{i=1}^{m} \frac{u_i^{(l)}}{m}$$

and

$$T_m^{(l)} = (1 + 1/m)b^{(l)} + \bar{u}^{(l)}$$

So for equations (1) and (3), the $u^{(l)} = T_m^{(l)}$ and the $q^{(l)} = \bar{q}^{(l)}$, so that gives us:

$$\bar{q}^r = \sum_{l=1}^{r} \sum_{i=1}^{m} \frac{q_i^{(l)}}{mr}$$

[8]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 performed on the total variance that has been computed by combining all of the component variances appropriately.

$$\bar{u}^r = \frac{1}{r}\left(\sum_{l=1}^{r}\left((1+1/m)\left(\sum_{i=1}^{m}\frac{(q_i^{(l)}-\bar{q}^{(l)})^2}{(m-1)}\right)+\sum_{i=1}^{m}\frac{u_i^{(l)}}{m}\right)\right)$$

If we break up $\bar{q}_M$ into the two pieces and rename things according to the notation in Reiter (2004), then our final combination formulae for combining completed implicates across synthetic replicates is as follows:

$$\bar{q}^M = \bar{q}^r = \sum_{l=1}^{r}\sum_{i=1}^{m}\frac{q_i^{(l)}}{mr} \tag{6}$$

$$B_M = b_r = \sum_{l=1}^{r}\frac{(q^{(l)}-\bar{q}_M)^2}{(r-1)} \tag{7}$$

$$\bar{b}^M = \frac{1}{r}\sum_{l=1}^{r}\left(\sum_{i=1}^{m}\frac{q_i^{(l)}-\bar{q}^{(l)}}{(m-1)}\right) \tag{8}$$

$$\bar{u}^M = \sum_{l=1}^{r}\sum_{i=1}^{m}\frac{u_i^{(l)}}{mr} \tag{9}$$

and equations (4) and (5) become:

$$T_r = (\frac{B_M}{r}) + (1+1/m)\bar{b}_M + \bar{u}_M \tag{10}$$

$$v_r = (r-1)\left(1+\frac{(1+1/m)\bar{b}_M+\bar{u}_M}{(B_M/r)}\right)^2 \tag{11}$$

Where $\bar{q}_M$ is the final estimate for q, B_M is the variance of our estimate of q across replicates, $\bar{b}_M$ is the average of the variance of the estimate across replicates, and $\bar{u}_M$ is the average variance across replicates and implicates. The final estimate, $\bar{q}_M$, is student-T distributed with variance, T_r, and degrees of freedom, v_r. Proofs and details can be found in Reiter (2004).

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 multiply imputed data must be used together in order to create

a comparison basis. The formulae for combining results based on the internal GSF are similar to those for combining synthetic implicates. If the user chooses not to multiply impute missing data, then analysis for their validation will be based on a single data file and no combining will be needed. If the user chooses to multiply impute missing data, then combination formulae are as follows:

$$\text{average across implicates:} \quad \bar{q}_m = \sum_{i=1}^{m} \frac{q_i}{m}$$

$$\text{variance across implicates:} \quad b_m = \sum_{i=1}^{m} \frac{(q_i - \bar{q}_m)^2}{(m-1)}$$

$$\text{variance on each implicate file:} \quad u_i = u\left(D_i\right)$$

$$\text{average variance across implicates:} \quad \bar{u}_m = \sum_{i=1}^{m} \frac{u_i}{m}$$

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

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

5.2 Addressing Missing Values

This section provides additional guidance on how users can handle missing values in the SSB. There are four broad categories for handling missing values: dropping missing values, weight-based adjustments, imputing missing values, and model-based inference. There are trade-offs associated with each method. Imputation allows users to maintain larger sample sizes and relax MCAR assumptions, but computing reliable imputations can be difficult and introduces additional uncertainty into the data. Dropping missing values, also known as listwise or pairwise deletion, avoids issues of imputation uncertainty, but does so at the cost of smaller and potentially biased samples. Weight adjustments can be applied to the complete- and available-case samples to bring the weights back into alignment with stratified population totals which may help address bias, but small sample issues may remain. While a robust literature exists on weight adjustments, imputation, and model-based inference, many research studies simply drop missing values with little or no discussion about how this may affect the results.

This section focuses on two missing data methods: model-based inference via sequential regression multivariate imputation (SRMI) and a weight-based adjustment known as raking. The next two sections describe these two approaches both in theory and in application to the SSB. Example programs in

SAS and Stata for using each approach with the SSB data are available to SSB users on the synthetic data server.[9]

5.2.1 Sequential regression multivariate imputation

Common imputation methods include hot deck imputation, which substitutes observed values for missing values; mean imputation, which substitutes group-specific means for missing values; and regression imputation, which predicts missing values based on the relationship between variables in the observed data. These methods can be useful but may miss important associations between missing or observed values and may underestimate variability. Furthermore, by replacing the missing values with a single value, they fail to take into account the uncertainty associated with missing values.

Multiple imputation addresses these shortcomings by imputing missing values several times. Multiple imputation has been shown to outperform the other imputation-based methods in many situations, particularly when the missing-data model is sound and for inference in smaller samples (Rubin, 1996; Fay, 1996; Rao, 1996). Multiple imputation can be performed in a variety of ways. The discussion and analysis below uses Sequential Regression Multivariate Imputation (SRMI) (Raghunathan et al., 2001).

Examples programs on the synthetic data server show how to impute missing values in the synthetic data using IVEware (Raghunathan et al., 2016). The SAS procedure PROC MI is an alternative, but previous work has suggested that IVEware performs better and is more user friendly than PROC MI (Vizcarra and Sukasih, 2013). Because IVEware is written in macro language, it can be combined with other procedures and data steps, both before and after imputation, which makes it easy to combine with other user-written programs. Additionally, IVEware has a number of features that can be used to specify detailed model settings for the imputation. Examples of these include the ability to model different types of variables, drop variables from the model, transfer variables to the output dataset without being used in the imputation model, restrict imputation to sub-populations, place bounds on imputation values, and include interactions. The IVEware programs are available on the synthetic data server for users to access in conjunction with their own programs. IVEware provides SAS, Stata, and R versions of their programs.

5.2.2 Raking

Listwise and pariwise deletion are alternatives to imputation. Listwise deletion refers to analyzing survey respondents who lack missing values for any variables. Pairwise deletion refers to analyzing survey respondents who lack missing values for only the variables of interest. In many cases, little attention is paid to selection bias that may result from complete- or available-case analysis.

[9]The programs not only provide the raw code needed to perform the multiple imputation or raking, but also incorporate the results into example analyses that provide additional useful features, such as code for combining results across synthetic implicates as described earlier.

One technique to account for potential bias arising from this approach is raking adjustments to sample weights. Sample weights can become inaccurate when respondents in the survey are dropped from analysis due to missing values; if missing values occurs not at random but at differential rates across subsets of the population, then certain sampling groups within the survey design may become over- or under-represented in the sample when respondents with missing values are dropped. This is where raking adjustments become useful: in these cases, the relationship between the sample and population can often be improved by adjusting the sample weights such that marginal totals match specified control totals for particular cross-classifications of variables. The process of raking is performed by iteratively adjusting marginal weighted population totals to match that of a specified control total. The iteration continues until each marginal weighted total is within a pre-specified threshold of the control total. Additional discussion on raking adjustments can be found in Ireland and Kullback (1968), Bishop et al. (1975), Kalton (1983), and Battaglia et al. (2013).

Example programs for raking survey weights and using the raked weights for analysis are available in SAS and Stata on the synthetic data server. The raking program was written to be similar to the raking programs used for creating the original SIPP weights. Thus, choices such as which variables, variable values, and cross-classifications to use for constructing the marginal totals to which the base weights are raked were largely based on similarity to SIPP programs, availability of SIPP variables in the SSB, and publicly-available U.S. population data for the necessary cross-classifications.

5.3 Addressing Sample Uncertainty

Surveys select samples of respondents from within specified groups, according to sample selection criteria. This introduces sample uncertainty into the data, which represents the uncertainty that the particular sample of the population available to the researcher is representative of the entire population that the sample is intended to represent. This uncertainty may be related to sampling error, which results from the fact that the sample is not the entire population, or non-sampling error, which results from flawed or biased sampling and collection methods (Assael and Keon, 1982; Lessler and Kalsbeek, 1992). To account for sample uncertainty, the sample selection, data collection, and analysis process would need to be carried out many times. Variance estimates for parameters of interest could then be constructed using the variance of parameter estimates across samples. This is infeasible given the limited resources associated with most surveys, so sample standard errors are often estimated using traditional variance formulae that assume simple random sampling.

This simple random sampling assumption is problematic because expressions of variance estimation based on complex survey designs can be quite complicated; survey designs often produce data that do not satisfy the conditions required for the application of variance formulae that assume simple random sampling (McCarthy, 1966). A practical alternative is to repeatedly draw sub-

samples from the full sample and then use the variance of estimates across sub-samples to construct variance estimates. This generates more informed standard error estimates that mimic the theoretical basis of standard errors while retaining information about the sample design (McCarthy, 1966). These sub-samples are commonly referred to as replicates. However, similar to the tendency of research studies to drop missing values, many studies also do not use this replicate process for generating variance estimates. The remainder of this section describes the process of raking based on Fay's Balanced Repeated Replicate (BRR) method (Judkins, 1990) and how this approach can be used with SSB data. Examples programs for constructing and using replicate weights with the SSB are provided on the synthetic data server.

The BRR method is a half-sample replicate method for sample designs that feature exactly two groups or clusters within each sample stratum. These clusters are also sometimes referred to as half-samples. They can be used to split the survey sample into two halves for each stratum. Performing BRR requires two pieces of information from the sampling procedure: the stratum from which each respondent was selected and the cluster from within that stratum to which the respondent belongs.

The BRR method forms a series of half-sample replicates by selecting one of the two half-samples for each stratum. The half-samples are selected according to a published BRR replicate structure such that the samples are balanced, fully orthogonal, and produce an unbiased variance estimator (McCarthy, 1966). Replicate weights are then generated by multiplying the weights of units in the selected half-sample by two and dropping units from the other half-sample. Fay's method of BRR imposes one small variation: rather than doubling the weights of the selected half-sample and zero-weighting the other half-sample, a small perturbation factor, α, is inserted such that the weights of the selected half-sample are multiplied by (2-α) and the units of the other half-sample are multiplied by α. This avoids small sample sizes that may arise from dropping half of the sample. A perturbation factor of 0.5 is used in the example programs on the synthetic data server.

For accurate replicate weights that fully account for sample design effects and weighting adjustments, the replicate weights should be computed starting with the base weight, which is the inverse of the probability of selection (Lemeshow, 1979; Chowdhury, 2013). Then, all post-stratification and raking adjustments can be applied to each replicate separately.

The stratum and half-sample variables needed in order to construct replicate weights are now included in the SSB. This allows users to construct replicate weights based on the base SIPP weight, which is also included in the SSB. The base SIPP weight is the inverse probability of selection for the sample unit adjusted only for non-response and sampled units which turned out to represent more than one household. Example programs for creating replicate weights with the SSB and applying the replicate weights are provided on the synthetic data server.

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

Due to changing demands within the Census Bureau, there are currently no production plans for a another release of the SSB, although releases could resume in the future. 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 different. While much of the content is similar, assimilating the 2014 panel into the GSF would be challenging. The SSB development team currently expects that a separate GSF file would be required for SIPP panels beginning in or after 2014.

Research has been done involving 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 com-

plicated because of the number of employers per individual. If a new version of the SSB were to include such data, 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. Continued statistical research on data synthesis methods coupled with feedback from researchers using the SSB will help the SSB maintain data privacy protection while expanding and improving in both topical coverage and data reliability.

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.

Benedetto, Gary, Martha H. Stinson, and John M. Abowd (2013), "The Creation and Use of the SIPP Synthetic Beta." Available at: `http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf`

Assael, Henry, and John Keon. 1982. "Nonsampling vs Sampling Errors in Survey Research." *Journal of Marketing* 46: 114-123.

Battaglia, Michael P., David D. Hoaglin, and Martin R. Frankel. 2013. "Practical Considerations in Raking Survey Data." *Survey Practice* 2 (5).

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Chowdhury, Sadeq R. 2013. "Effects of Poststratification and Raking Adjustments on Precision of MEPS Estimates." Agency for Healthcare Research and Quality.

Fay, R.E. 1996. "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91: 490-498.

Ireland, C.T., and S Kullback. 1968. "Contingency Tables with Given Marginals." *Biometrika* 55: 179-188.

Judkins, David R. 1990. "Fay's Method for Variance Estimation." *Journal of Official Statistics* 6(3): 223-239.

Kalton, G. 1983. "Compensating for Missing Survey Data." Survey Research Center, Institute for Social Research, University of Michigan.

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.

Lemeshow, S. 1979. "The use of unique statistical weights for estimating variances with the balanced half-sample technique." *Journal of Statistical Planning and Inference* 3: 315-323.

Lessler, Judith T., and William D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: Wiley.

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

McCarthy, Philip J. 1966. "Replication: An Approach to the Analysis of Data from Complex Surveys." *Vital and Health Statistics Series* 2-No. 14.

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.

Raghunathan, T.E., P Solenberger, P. Berglund, and J. van Hoewyk. 2016. "IVEware: Imputation and Variance Estimation Software (Version 0.3)."

Rao, C.R. 1996. "On Variance Estimation with Imputed Survey Data." *Journal of the American Statistical Association* 91: 499-506.

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.

Rubin, D. B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91: 473-489.

Figure 1: Privacy protection versus data utility trade-off

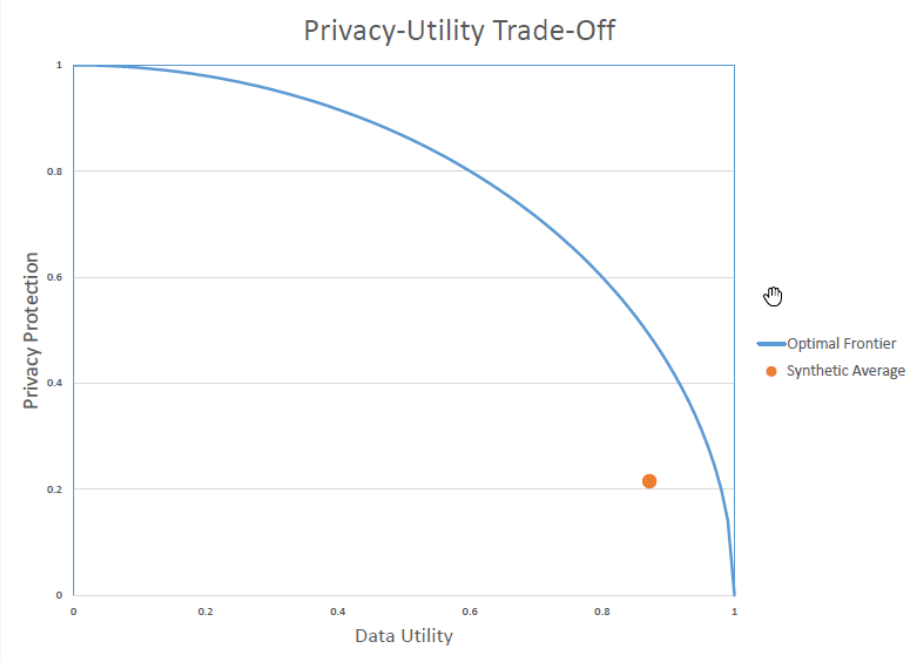

Table 1: RMSE of the synthetic value of the nearest synthetic record: Total net worth

Select Quantiles						
p10	p25	p50	p75	p90	p95	p99
$ -	$ -	$ 17,000.00	$ 111,000.00	$ 300,000.00	$ 499,000.00	$ 1,134,000.00

Wealth Category	n	RMSE
<0	18000	$ 35,600.00
0	240000	$ 15,320.00
<p50	131000	$ 75,560.00
<p75	203000	$ 113,300.00
<p90	118000	$ 173,400.00
<p95	38000	$ 353,600.00
<p99	28500	$ 463,100.00
>=p99	6500	$ 5,586,000.00

Overall		
RMSE	Std. Dev.	Ratio
$ 521,000.00	$ 548,400.00	0.95

Table 2: RMSE of the synthetic value of the nearest synthetic record: 2010 total FICA covered earnings

Select Quantiles						
p10	p25	p50	p75	p90	p95	p99
$ 2,859.00	$ 10,150.00	$ 26,420.00	$ 50,420.00	$ 83,950.00	$ 115,000.00	$ 246,800.00

Earnings Category	n	RMSE
0	282,000	$ 58,750.00
<p10	30,500	$ 59,720.00
<p25	45,500	$ 82,170.00
<p50	76,000	$ 80,760.00
<p75	76,000	$ 123,300.00
<p90	45,500	$ 169,200.00
<p95	15,000	$ 296,100.00
<p99	12,000	$ 594,200.00
>=p99	3,000	$ 1,501,000.00

Overall		
RMSE	Std. Dev.	Ratio
$ 173,700.00	$ 78,430.00	2.21

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 from the Social Security Administration (SSA) and the Internal Revenue Service (IRS) for certain survey respondents for the purpose of improving core Census Bureau data products.[10] 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. All three agencies contributed data and statistical expertise and Census and SSA provided 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 File 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. 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, 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 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 source panel was included to reduce missing data imputation. 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 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. In order to speed the

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

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 was released in May 2013. It used the same SIPP panels as version 5.0 but added a substantial number of SIPP variables, in particular, ones which vary over time. Version 5.1 had fewer unsynthesized variables as only gender and spouse link were not synthesized. Significant improvements were made to the modeling of the earnings tax data by first cleaning the underlying data in order to prevent administrative data error from skewing the synthesis. Version 5.1 was also the first to contain geography (state at time of interview).

SSB version 6.0 was released in February 2015. Version 6.0 added the 1984 and 2008 SIPP panels while extending the years of administrative record availability. New SIPP variables were added including time series variables pertaining to government program receipt and amounts. Additional information on SSDI and SSI application submissions were pulled from the administrative records. Finally, this version sought to improve fertility history variables by reconciling child birthdays from administrative data with the self-reported birthdays provided by mothers in the SIPP. Version 6.02 also altered the naming convention for monthly SIPP variables by changing the time identifier from calendar year and month to month number since start of panel. For example, the first interview of the 1996 panel was given in April 1996 with four months retrospective extending back to December 1995. So, April 1996 would correspond to variable_19964 in Version 6.0 and variable_5 in Version 6.02.

Version 7.0 is the most recent release of the SSB. The SIPP panels are the same while the administrative record coverage extends through 2014. The creation process has been altered in several ways. First, all variables are now synthesized as is the missing data pattern. Second, the process of producing the SSB from the GSF has changed. We now begin by creating four synthetic files from a "snapshot" of the internal Gold Standard File. This snapshot contains all the variables which will be featured in the SSB. With missing values included in the synthetic files, users are now granted more freedom to choose how to address missing values.

The naming convention for monthly SIPP variables has also changed again. Such variables are now defined by the year number within SIPP panel and calendar month. Using the previous example for the 1996 SIPP panel, calendar year 1 is 1995 since the earliest observations correspond to December of 1995; therefore, a variable representing April 1996 would correspond to variable_2_4. Several variables from the previous version have been updated while some new variables have been added. For example, the available categories for the industry and occupation variables have been expanded while the variables themselves are now time series (by wave). The variable indicating state of residence during the SIPP is now dis-aggregated rather than grouping some states together for some years. New variables include respondent birthdate information from the SIPP, date of filing for SSDI, birth years of a respondent's first child and last child, and the stratum and half-sample variables. The SSB v7.0 codebook contains the full list and descriptions of variables.