

Disclosure Review Board Memo: Second Request for Release of SIPP Synthetic Beta Version 6.0

Gary Benedetto and Martha Stinson
Survey Improvement Research Branch,
Social, Economic, and Housing Statistics Division (SEHSD)

January 15, 2015

1 Introduction

We are requesting the approval of the Census Disclosure Review Board (DRB) for the release of the SIPP Synthetic Beta (SSB) v6.0, produced by the Survey Improvement Research Branch (SIRB) of the Census Bureau's Social, Economic, and Housing Statistics Division. This data product is an update to the previously released SSB v5.1. In this memo we provide a brief review of the creation of the SSB and then describe our disclosure-risk analysis. From the results of this analysis, we conclude that the release of SSB 6.0 would not risk disclosing the identity of any SIPP respondent.

This memo is a follow-up to our request for approval to release version 6.0 in May 2014. After DRB approval of this earlier request, further review of the product prior to release revealed some processing errors. Hence SSB 6.0 was not released in the summer or early fall of 2014 as planned. Instead we fixed the errors and re-synthesized the data. This memo describes our disclosure-risk analysis of the new synthetic data.

2 SSB Creation

To create the SSB v6.0, we first combined nine SIPP panels (1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, 2008) to create the Gold Standard File (GSF). We chose a subset of SIPP variables and then standardized these data elements across panels and merged them with SSA-provided administrative data from the Summary Earnings Records (SER), the Detailed Earnings Records (DER), the

Master Beneficiary Record (MBR), the Supplemental Security Record (SSR), the 831 Disability File (F831), and the Payment History Update System (PHUS). We then employed regression-based multiple imputation to fill in the missing data of the GSF to create four *completed* data sets, called implicates. These implicates are identical to the GSF except that missing data are replaced with independent draws from a probability distribution. We refer to these four datasets as the completed data. We then use the same modeling techniques to create 16 *synthetic* data sets or implicates (4 synthetic implicates per completed implicate). These are produced by setting all but two variables to missing for every record and then applying the above methods for replacing missing data with independent draws from the estimated probability distributions. The two variables left unsynthesized are gender and the first marital link observed in the SIPP. As was the case in version 5.1, these are the only variables in the SSB that contain actual data from any source.

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. We also note that SSB version 6.0 will not be linkable to SSB version 4.0, 5.0, or 5.1.

The Internal Revenue Service (IRS) and Social Security Administration (SSA) will also review this memo and their approval is required in order for SSB 6.0 to be released.

3 Disclosure Testing Methods

3.1 Overview

Our disclosure avoidance analysis uses 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 File¹. 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. A realistic intruder would have the SIPP public use data, and not the administrative records, so by using

¹At 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.

the full Gold Standard file we are producing a very conservative estimate of re-identification risk.

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 data².

In version 6.0 of the SSB, we relax the age restriction imposed in earlier versions and keep all individuals in the synthetic data regardless of age. This means that our synthetic data implicates now have the same number of observations as each other and as the GSF (783,781 observations). We used the full set of observations in the first synthetic implicate in all of our re-identification exercises. Since the synthetic implicate file has the same sample size as the Gold Standard, we know that a “true match” between the two files exists.

Importantly, simply linking a record in the SSB v6.0 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.

3.2 Differences between version 6.0 and 5.1

The main purpose of releasing a new version of the SSB is to include two additional SIPP panels (1984, 2008) and to lengthen the administrative data time series to include data through 2011 for the earnings variables and 2012 for the benefits variables (previously in version 5.1, end years were 2006 and 2007 respectively). We have included two new SIPP point-in-time variables (a base weight and an ownership of a life insurance policy indicator), eight new SIPP time-series variables that include monthly data for the time period covered by the respondent’s SIPP panel (AFDC/TANF receipt and amounts; veterans’ benefits receipt and amounts; SNAP/Food Stamp receipt and amounts; workers’ compensation receipt and amounts), and have made modifications to the content of a few other SIPP variables (fertility history and self-reported disability status). All other SIPP variables are in the same format and have the same

²Co-habiting same-sex partners were not allowed to declare themselves married in the SIPP panels contained in SSB v6.0. Hence a married couple always has both a male and female.

content as in version 5.1, with the only exception being the correction of small errors discovered since the release of version 5.1. We have added a significant number of disability application and receipt variables from the MBR and F831 data, expanding from 6 variables to 48. Likewise we have expanded the number of SSI application and receipt variables from the SSR data that we include from 3 to 9. Appendix A provides the complete list of new SSDI and SSI variables that we include. All new SIPP and SSA variables are **synthesized** and hence we do not expect this increased number of variables to add to the disclosure risk.

As in version 5.1, the unsynthesized SIPP variables create only four cells: married/male, married/female, single/male, single/female which all contain more than 100,000 individuals. Hence linking synthetic implicates to each other is no longer a concern because there are not any small cells that would make this possible. Also consistent with version 5.1, we include a synthesized state variable. The 1984-2001 SIPP panels combined some states into groups and reported only the group instead of the actual state. We respected these state groups in our synthesis. The synthesis process was engineered to assign state codes that were consistent with the synthetic panel variable. We continue to use state as one of our matching variables in the re-identification analysis but the different groupings of states across panels created some challenges. For example, in the 1990-1993 panels, state code 62 included Iowa, North Dakota, and South Dakota and state code 63 included Alaska, Idaho, Montana, and Wyoming. In the 1996 and 2001 panels, state code 62 included Wyoming, South Dakota, and North Dakota and Alaska, Idaho, and Montana were each given individual state codes. All panels from 1990 to 2001 grouped Maine and Vermont in code 61. The 2004 panel did not contain any state groupings. If we had simply matched on the state codes as given, an SSB record with panel=2001 and state=62 when compared to a GSF record with panel=1993 and state=63 would not have agreed on state even though the respondent might have lived in Wyoming in both records. This would have artificially lowered the matching rate because the different state codes would not have agreed. Hence we created for our disclosure-risk analysis a new set of codes that represented the coarsest grouping possible and matched on these. Our coarsest grouping contained two state groups: state code=62 for Iowa, North Dakota, South Dakota, Wyoming, Alaska, Idaho, and Montana; code=61 for Maine and Vermont. All other states had individual state codes. We discuss the implications of this approach at the end of Section 3.3, after explaining our minimum distance matching methodology in more detail.

3.3 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) used 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:

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

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.

We make special note of the implications for using distance measures to match on the state variable. We created a set of dummies that represented each state or state-group and then the Euclidian distance measure, for each given state indicator, was either zero or one, with zero representing agreement in state and one representing disagreement. Because of our coarsening of the state categories for the purposes of matching, we actually create more agreement when matching state than would otherwise be present. For example consider a GSF record with panel=2004 and state=Alaska compared to a synthetic record with panel=2001 and state=Idaho. Our coarsened state variable will have the same code for both these states, since we grouped them together to respect the 1990-1993 set of state codes, and hence it will appear as if state matches when in reality it does not. This will lower the distance between the records. If these two records are a “true match” then this will raise the probability of this match being one of the top three. However synthetic records with panel=2001 and state=Idaho which are “false matches” will also have lower distance scores and hence the effect on the overall likelihood of finding the “true match” is ambiguous.

4 Results

The Census Bureau Disclosure Review Board has used two standards for disclosure avoidance in past reviews of SSB disclosure requests. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. Second, given a strategy for choosing a best match, the ratio of true matches to the total number of best matches (true and false) should be close to one-half or smaller. To this end, we report the percentage of records that were declared a best match by virtue of being the closest record and were in fact true matches.

Even with our conservative approach, for most blocking group-distance measure combinations, the minimum distance match represented the true match much less than 1% of the time. Table 2 shows the results for the four minimum distance strategies across each blocking group. We found that the strategy that used the greatest amount of information about the covariance of the variables (*MAHA1*) performed the best, and even that strategy matched to the true record about 0.1% of the time for singles, and about 0.3% of the time for couples.

Moreover, the last two columns of Table 2 show that the optimal strategy of taking the minimum distance match as the candidate match was barely better than sub-optimal strategies of taking the 2nd smallest distance pair or 3rd smallest distance pair. Column 4 is the ratio of column 1 and column 2. If this ratio were exactly one, then the smallest distance strategy would produce exactly the same number of correct matches as the 2nd smallest distance strategy. Likewise column 5 is the ratio of column 1 and the sum of columns 2 and 3. For all the distance measures, the ratios in Column 4 are similar, ranging from slightly higher than 1 to approximately 1.5. These ratios show that, at best, the optimal strategy only marginally out-performs the second-best strategy. In column 5 we see that the number of combined correct matches from the second and third best matches is always higher than the number of matches from the first best group (ratio is less than one). Given these results, it is clear an intruder would have difficulty finding a reliable strategy for producing correct matches.

5 Conclusion

Given the results shown in Table 2, we conclude that the SSB v6.0 is not a threat to the confidentiality of the identities of SIPP respondents. The results suggest that an intruder, even when armed with far more information than we can reasonably assume any intruder to have, can have very little confidence that a re-identification strategy on the synthetic data will produce a match to the public-use SIPP files. The results from our conservative re-identification exercises produce true match rates far below the 50% upper bound, and optimal re-identification strategies barely outperform sub-optimal strategies, both of

which indicate that the synthetic data introduces a great deal of uncertainty for any intruder.

6 References

Domingo-Ferrer, Josep and Abowd, John M. and Torra, Vicenc, "Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment," in *Privacy in Statistical Databases*, ed. Domingo-Ferrer, J. and Franconi, L., Springer-Verlag, 2006.

Domingo-Ferrer, Josep and Torra, Vicenc and Mateo-Sanz, J.M. and Sebe, F., "Empirical disclosure risk assessment of the IPSO synthetic data generators," in *Monographs in Official-Statistics-Work Session on Statistical Data Confidentiality*, Eurostat, 2006.

7 Appendix A List of new SSDI and SSI variables included in version 6.0

mbr_ssdi_applied_{n}: indicator for application{n=1,2,3,4} submitted

mbr_ssdi_entitled_{n}: indicator for application{n=1,2,3,4} entitles recipient to benefits (i.e. approved)

mbr_ssdi_ddo_{n}: date of disability onset from application{n=1,2,3,4}

mbr_ssdi_doed_{n}: date of disability entitlement from application{n=1,2,3,4}

mbr_ssdi_dsd_{n}: date of disability adjudication for application{n=1,2,3,4}

mbr_ssdi_ddbc_{n}: date of disability benefits cessation for application{n=1,2,3,4}

mbr_ssdi_dig_group_{n}: diagnosis group for application{n=1,2,3,4}

mbr_ssdi_benefit_totamt_{n}: monthly benefit amount for application{n=1,2,3,4}

mbr_ssdi_ceased_{n}: indicator for benefits ceased for application{n=1,2,3,4}

PHUS_ssdi_benefit_stdate_{n}: start date of disability benefits reported in PHUS{n=1,2,3,4}

PHUS_ssdi_benefit_totamt_{n}: amount of disability benefits reported in PHUS{n=1,2,3,4}

PHUS_ssdi_pos_totamt_{n}: indicator for positive payment in PHUS{n=1,2,3,4}

ssr_ssi_appl_dt : Application date

ssr_ssi_applied: Applied for SSI benefits

ssr_ssi_benefit: Received SSI benefits
ssr_ssi_benefit_fed_totamt: Total federal benefit amount
ssr_ssi_ceased: Benefits ceased
ssr_ssi_dig_group: Diagnosis code
ssr_ssi_first_pmt_dt: First payment date
ssr_ssi_last_pmt_dt: Last payment date
ssr_ssi_type: Type of benefit

Table 1: Variables Used to Match

Couple-level variables

For couples, these variables have the same values for husband and wife.

Singles do not have the marriage variables in this list.

panel
 start date of linked marriage
 end date of linked marriage
 indicator for linked marriage ending
 reason for linked marriage ending
 state indicators

Person-level variables

Married couples: match on the husband and wife's values for each of these variables.

Singles: match on the person's own values only.

black other race education: 5 categories Hispanic foreign born own a home deathdate exists: person dies by end of 2008 indicator for SIPP reported disability indicator for SIPP reported disability that prevents work if foreign born, decade arrive in the USA valid SIPP industry exists 4 category SIPP-reported industry valid SIPP occupation exists 3 category SIPP-reported occupation home equity non-housing wealth in scope to be asked pension question because employed indicator for dc pension indicator for db pension deathdate birthdate number of biological children ever number of marriages number of divorces year finish high school year finish post-high school education year of bachelor degree field of bachelor degree	number of years with positive SER/DER earnings: 1951-2006 number of years with positive deferred DER earnings: 1990-2006 average total SER/DER earnings: 1951-2006 average deferred DER earnings: 1990-2006 first four moments of distribution of years <i>weighted by earnings (e.g. first moment is</i> <i>"average year" where each year is weighted</i> <i>by the % of lifetime earnings in that year)</i> number of months of positive SIPP-reported earnings average monthly SIPP-reported earnings number of months with positive SIPP-reported work hours average weekly SIPP-reported work hours number of months with health insurance coverage number of months with employer-provided health insurance coverage total weeks with job during SIPP panel total weeks with pay during SIPP panel number of months with positive income average monthly SIPP-reported income number of months of welfare (Food Stamps/AFDC/TANF) indicator for receive retirement OASDI benefits indicator for receive widow OASDI benefits indicator for receive spouse OASDI benefits indicator for receive disability OASDI benefits indicator for receive SSI benefits
--	--

Table 2: Matching Results by Distance Matching Method and Type of SIPP Respondent

Method and SIPP Marital Status Group	Number of Blocks	Average Block Size	Percent Correctly Matched:			Ratio of best to:	
			(1) Best (smallest distance)	(2) Second Best	(3) Third Best	(4) Second Best	(5) Second Best + Third Best
MAHA1							
couples	15	10638	0.26%	0.19%	0.16%	1.38	0.75
single women	24	10162	0.08%	0.08%	0.06%	1.01	0.54
single men	22	10034	0.06%	0.06%	0.05%	1.09	0.57
MAHA2							
couples	15	10638	0.04%	0.03%	0.04%	1.09	0.50
single women	24	10162	0.04%	0.03%	0.03%	1.49	0.68
single men	22	10034	0.04%	0.03%	0.04%	1.35	0.61
EUCLIDIAN							
couples	15	10638	0.07%	0.05%	0.04%	1.44	0.77
single women	24	10162	0.10%	0.08%	0.06%	1.21	0.68
single men	22	10034	0.07%	0.06%	0.04%	1.26	0.70
EUCLIDIAN STD							
couples	15	10638	0.05%	0.05%	0.04%	0.85	0.48
single women	24	10162	0.04%	0.04%	0.04%	1.01	0.50
single men	22	10034	0.03%	0.03%	0.03%	1.05	0.53

Data: SIPP Synthetic Beta Implicate 1 and SIPP Gold Standard File, panels 1984, 1990, 1991, 1992, 1993, 1996, 2001, 2004, 2008