Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project

John M. Abowd, Martha Stinson and Gary Benedetto

November 5, 2006

This report was produced by the Longitudinal Employer-Household Dynamics Program at the U.S. Census Bureau, Jeremy S. Wu, Assistant Division Chief, Data Integration Division. The report is required by the Jointly Financed Cooperative Agreement between the Census Bureau and the Social Security Administration for fiscal year 2006 (SSA agreement number BC-05-05, as amended; Census Bureau agency reference number 0084-2005-043-002-001, project account 7675084). John Abowd participated in the project in his capacity as Distinguished Senior Research Fellow at the Census Bureau (on IPA from Cornell University). Martha Stinson and Gary Benedetto are economists on the LEHD staff. In addition to the authors named above, Lisa Dragoset (Census-LEHD), Sam Hawala (Census-SRD), Karen Masken (IRS), Bryan Ricchetti (Census-LEHD), Lars Vilhuber (Census-LEHD), and Simon Woodcock (Simon Fraser University) all contributed to the research. Josep Domingo-Ferrer (University of Rovira and Virgili), Jerome Reiter (Duke University), Vicenc Torra (Artificial Intelligence Lab, University of Barcelona), and Simon Woodcock, operating with the support of the Census Bureau through a subcontract to the main Research and Development contract between the Census Bureau and Abt Associates, Inc. (Census Bureau contract number 50YABC-2-66036, task order number TO002) to Cornell University (OSP reference number 47632), provided substantial consulting on the creation of the SIPP/SSA/IRS-PUF. The National Science Foundation through Grants ITR-0427889 and SES-0339919 to Cornell University with subcontracts to the Census Bureau (Census Bureau agreement number 0063-2005-003-000-000, project account 9098000) also provided substantial support for this project.

Contents

1 Executive Summary

1.1 Purpose and brief history

The creation of public use data that combine variables from the Census Bureau's Survey of Income and Program Participation (SIPP), the Internal Revenue Service's (IRS) individual lifetime earnings data, and the Social Security Administration's (SSA) individual benefit data began as part of ongoing collaborative research at the Census Bureau and SSA. The current project had its genesis with the formation of a joint committee containing representatives from the Census Bureau, SSA, IRS, and the Congressional Budget Office (CBO) that designed a prospective public use file. Aimed at a user community that was primarily interested in national retirement and disability programs, the selection of variables for the proposed SIPP/SSA/IRS-PUF focused on the critical demographic data to be supplied from the SIPP, earnings histories from the IRS data maintained at SSA, and benefit data from SSA's master beneficiary records.

After attempting to determine the feasibility of adding a limited number of variables from the SIPP directly to the linked earnings and benefit data, it was decided that the set of variables that could be added without compromising the confidentiality protection of the existing SIPP public use files was so limited that alternative methods had to be used to create a useful new public use file. The committee agreed to allow the Census Bureau to experiment with the confidentiality protection system known generically as "synthetic data." The actual technique adopted is called partially synthetic data with multiple imputation of missing items. As the term is used in this report, "partially synthetic data" means the release of person-level records containing some variables from the actual responses and other variables where the actual responses have been replaced by values sampled from the posterior predictive distribution for that record, conditional on all of the confidential data.

From 2003 until the present, four preliminary versions of the SIPP/SSA/IRS-PUF have been produced. This final report accompanies the delivery of version 4.0 to SSA as part of the fiscal year 2006 Jointly Financed Cooperative Agreement between the Census Bureau and SSA.

1.2 Structure of the inputs to the SIPP/SSA/IRS public use file

The SIPP/SSA/IRS-PUF contains data from the records of individuals who responded to the SIPP panels conducted in 1990-1993 and 1996 A standardized extract of approximately 125 variables from all waves of each of these panels was created. We included the following demographic variables: gender, marital status, race (black), five categories of education, Hispanic ethnicity, birth date, death date, disability status, number of children, marital history, foreign born, decade arrived in United States if foreign born, and a spouse identifier that links to the marriage partner if the respondent is married and the spouse was also surveyed. We took the values for these variables at a point in time. For the time-invariant variables–gender, race, and Hispanic ethnicity, values were taken from the point in the SIPP when they were first reported, generally wave 1. Values for the other demographic variables were generally chosen from month 8 of the respective SIPP panel (*i.e.,* the last reference month of the second interview). We chose this point because marital, immigration, and disability histories were collected in the wave 2 topical modules and we wanted to take all the variables from the closest possible interview dates. For education, we searched over all reported education values in each wave of the SIPP and chose the highest level of education ever reported. Thus gender, marital status, race, education, Hispanic ethnicity, and spouse identifier are never missing in the standardized extract because these variables are all reported at least once, and we chose to take the self-reported values whenever they were available. Disability status, number of children, marital history, foreign born, and decade arrived in United States if foreign born are sometimes missing because individuals did not answer the relevant topical modules or because we chose not to search over every available month of SIPP data. All item missing data, with the exception of structurally missing items, were flagged for imputation.

This standardized extract was linked using the respondent's validated Social Security Number (SSN) to the following data provided by SSA:

- From SSA's Summary Earnings Record (SER), a longitudinal history of all FICA-covered wage and salary income earned since 1937, we linked the annual summary and the quarters-worked summary. These are the only earnings data available from the SSA and IRS files prior to 1978. This array is capped at the FICA taxable maximum;[1]

[1]These data, as well as the Detailed Earnings Record data cited in the next bullet, are also confidential under the protocol defined in Title 26 of the U.S. Code. Prior permission from the IRS disclosure officer is required before they can be used in a project in combination with Title 13

- From SSA's Detailed Earnings Record, a longitudinal history of wage and salary items from the employer-filed W-2 form by employer, we linked annual total wage and salary income and deferred earnings from all FICA-covered jobs. We also linked an analogous set of variables for non-FICA-covered jobs;
- From SSA's Master Beneficiary Record (MBR), a longitudinal history of type and amount of all benefits paid to an individual, we linked the entire history and created variables for type of benefit initially received, type of benefit received in April 2000, and the monthly benefit amount associated with those two benefit receipt dates.
- From the Census Bureau version of the master Social Security Number data base, known as Numident when sorted in SSN order, we linked the administrative birth and death dates.

Next, we added variables that were not destined for the public use file but would provide additional information useful in the process of completing the missing data, synthesizing the variables to be protected, creating a weight for the merged SIPP panels, and assessing the quality and disclosure risk of the final product. The documented, standardized extract from the SIPP 1990-1993 and 1996 panels, the linked SSA and IRS data, the supplemental variables added to facilitate processing and review, and the customized weight collectively define what we call the "Gold Standard" file. The codebook and technical description of the Gold Standard Version 4.0 accompanies this report. This codebook also documents the variables found in the completed Gold Standard files and the synthetic data files.

1.3 Completion of the missing data and synthesis of the confidentiality-protected data

Although the existing SIPP public use files have had all item non-response allocated using the methods developed for this purpose as part of the regular SIPP data processing, the Gold Standard version of the consolidated 1990-1993 and 1996 panels has item missing data for two basic reasons. First, SIPP respondents in the Gold Standard file for whom the Census Bureau does not have validated SSNs were missing all data items whose linkage depends upon the SSN; that is, all earnings, benefit, and administrative birth and death data. Second, because one of the critical components of the confidentiality protection is to prevent identifying the source record of the synthetic data in the existing SIPP public use files, all information regarding the dating of variables whose source was a SIPP response, and not administrative data, has to be made consistent across individuals regardless of the panel and wave from which the response was taken. This requirement resulted in the creation of ten-element arrays that contained all dated SIPP items, like family total income, with values inserted for each year from 1990 to 1999. No SIPP respondent household ever provided all ten of these items. Those array elements that were available for a particular respondent, which depend upon which panel the respondent answered, were populated by the actual value (from the public use version of the variable). All other elements in the array were item missing data. All missing data items that resulted from either missing validated SSN or missing items in an array were multiply imputed using the techniques described in the report. The imputation models were based on Bayesian bootstrap and Sequential Regression Multivariate Imputation methods for estimating and sampling from multivariate posterior predictive distributions.

There is a third source of missing data in the Gold Standard file. Some data items are structurally missing because it is not logically possible for the item to have a value; for example, no data are available concerning the second marriage of individuals who never married or married only once. Structurally missing data remain in the Gold Standard file and in the synthetic data implicates that constitute the SIPP/SSA/IRS-PUF.

The public use file contains several variables that were never missing and are not synthesized. These variables are: gender, marital status, spouse's gender, initial type of Social Security benefits, type of Social Security benefits in 2000, and the same benefit type variables for the spouse. All other variables in the public use file were synthesized.

In order to preserve exact logical relations among the variables, the first step of the missing data imputation process, and the first step of the data synthesizing process, is to implement a binary tree of parent-child relations among all the variables. This tree guides the execution of first the missing data imputation and then the synthetic data phase. We created the binary tree to organize the data processing by summarizing all of the assumptions and logical restrictions that must be preserved in the final data product.

The top level of this binary tree contains all variables that exhibit no logical dependencies on any other variables in the file, for example birth date. The tree has nine levels. At each level below the top, variables depend upon their

confidential data. Permission to conduct the present research is monitored by the Census Bureau under Adminstrative Records Tracking System project 458, which contains a copy of the IRS approval.

parents, and are only processed when appropriate. In the intermediate levels of the tree, a variable can be both a parent and a child, for example, whether or not there is a second marriage is a child of the same variable for the first marriage and a parent of the variable for the third marriage. The terminal level and all leaves of the binary tree contain only child variables.

For each iteration of the missing data imputation phase and again during the synthesis phase, we estimate a joint posterior predictive distribution for all of the required variables according to the following protocol. At each node of the parent/child tree, a statistical model is estimated for each of the variables at the same level. The statistical model is a Bayesian bootstrap, logistic regression, or linear regression (possibly with transformed inputs). All statistical models are estimated separately for detailed groups of individuals based on the values of categorical variables that include both demographic and economic controls. Logistic and linear regressions also include additional linear controls that are selected from a long list of potential right-hand-side control variables on the basis of the Bayes Information Criterion. Once the analyst specifies the grouping variables and their associated control variables, the estimation of a proper posterior predictive distribution from which to impute or synthesize, as appropriate, is fully automated. On the basis of the estimated models, and taking proper account of parameter uncertainty, each variable is imputed (missing data phase) or synthesized (synthetic data phase) conditional on all values of all other variables for that individual. The missing data phase included nine iterations of estimation. The synthetic data phase occurred on the tenth iteration. Four missing data implicates were created. These constitute the completed data files that are the inputs to the synthesis phase. Four synthetic implicates were created for each missing data implicate. Thus, there are a total of sixteen synthetic implicates in the SIPP/SSA/IRS-PUF Version 4.0.

A complete diary of the assumptions used to synthesize every variable in the PUF: parent/child relations, synthesizer method, statistical model, grouping variables, control variables, allowable values, logical limitations, synthesizer restrictions, and usage notes is included as an Excel workbook accompanying this report.

The software to implement the missing data imputation and confidentiality synthesis is written in SAS as a massively parallel application. Running on two 64-processor large memory computers at the Census Bureau the estimation phase for completing all 616 variables can be accomplished in about two months. Given completed data, a full run of the synthesizer (16 implicates) takes about three weeks.

1.4 Development of the weights

The final Gold Standard file contains data drawn from the survey responses and administrative records of individuals who responded to the Survey of Income and Program Participation in the 1990-1993 and 1996 panels. The design of the 1990-1993 panels envisioned combining data from waves of different SIPP panels that corresponded to the same calendar dates. Consequently, there are explicit instructions for recalibrating the SIPP weights when using individuals or households from the same year who were surveyed in different panels. The recalibrated weights account for the design and ex-post differences across the panels. The data collected as part of the 1996 panel do not overlap the time periods covered by the 1990-1993 panels. Hence, no official formulae exist for recalibrating the weights when combining data from the 1996 panel with data from the earlier panels.

The linkage of longitudinal lifetime earnings data from SSA's Summary and Detailed Earnings Records to individuals from these five SIPP panels implies that records that correspond to the same calendar year will come from all of the panels. Analyses that use these longitudinal earnings data cannot use any combination of the official SIPP weights to produce an estimate that has a fully specified reference population. This conundrum has faced analysts who used linked SIPP/SSA/IRS data, such as internal researchers at SSA and the Census Bureau, for years. In order to allow users of the SIPP/SSA/IRS-PUF Version 4.0 to conduct analyses with a known reference population, we created an ex post weight for the PUF. This weight can be used to make estimates representative of individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000.

Our method for creating an ex-post weight for the merged SIPP panels involved seven steps. First, we reproduced the major component of the 1996 sampling frame (the unit frame) in the Census 2000 micro-data, updating the SIPP reference population to April 1, 2000. Next, we divided the Decennial individual records (long and short form) into strata according to the 1996 SIPP sampling plan. Then we standardized all five SIPP panels with respect to geographic definitions and strata used in the 1996 sampling plan. Each individual observation in the merged panels was then placed in a stratum according to the 1996 SIPP sampling plan. The fifth step was to link each SIPP person to a person in the Census 2000 reference population. This match was accomplished using probabilistic record linking. Most observations could be linked on the basis of the PIK[2] that had been assigned to the SIPP or Decennial individual. For those SIPP individuals who did not link by PIK, a cruder probabilistic record linkage based on characteristics used to define the sampling strata was used. Having accomplished this linkage for all in-scope individuals in the 1990-1993 and 1996 SIPPs, we created a preliminary weight as the ratio of in-scope individuals in Census 2000 to in-scope individuals in the merged SIPPs within each final (stage-2) SIPP sampling cluster. The final weight was created by raking the preliminary weights to agree with official U.S. population control totals for the sex/age/race/ethnicity demographic breakdown of the reference population, as supplied by the Census Bureau's Population Estimates Division. This final raking was controlled to exactly the same population categories as the official 1996 SIPP weights.

The final weight was tested for analytical validity by creating weighted tables summarizing earnings and benefit measures from the administration of the Old Age, Survivor, Disability Insurance (OASDI) program. The estimates from the PUF were compared to SSA's published statistical summaries for the year 2000. When the final weight is applied to the completed Gold Standard data, the results reproduce most aspects of published SSA data derived from the universe of OASDI recipients.

Because copying the final weight to each implicate of the synthetic data would have provided an additional unsynthesized variable with 55,552 distinct values, the disclosure risk associated with the weight variable had to be addressed. We created a synthetic weight using a posterior predictive distribution based on the Multinomial/Dirichlet natural conjugate likelihood and prior. The likelihood component was created by modeling the 55,552 distinct cells created by all feasible combinations of the six variables used to create the final sampling clusters. The cell counts were the sums of the weights in each cell. The Dirichlet prior was uniform over all 55,552 cells with a prior sample size selected to insure adequate confidentiality protection. We sampled a complete table from the Dirichlet posterior for each synthetic implicate. An observation was assigned a weight equal to the posterior probability in its final sampling cluster times the civilian non-institutionalized U.S. population as of April 1, 2000 age 18 or older.

The synthetic weight was tested for analytical validity by comparing the pooled results from the 16 synthetic implicates to the analysis from the Gold Standard file of the same earnings and benefit measures that were studied to assess the quality of the final weight itself. When weighted analyses from the synthetic implicates are combined according to the correct formulae, the synthetic weight is just as reliable as the final weight we created for the Gold Standard file. The maximum discrepancy between the weighted Gold Standard analysis and the weight synthetic data analysis is -4.44% and most discrepancies are less than 2% in absolute value.

1.5 Analytical validity testing

Although synthetic data are designed to solve a confidentiality protection problem, the success of this solution is measured by both the degree of protection provided and the user's ability to estimate scientifically interesting quantities reliably. The latter property of the synthetic data is known as analytical (or statistical) validity. Analytical validity exists according to Rubin Rubin (1987) when, at a minimum, estimands can be estimated without bias and their confidence intervals (or the nominal level of significance for hypothesis tests) can be stated accurately. The estimands can be summaries of the univariate distributions of the variables, bivariate measures of association, or multivariate relationships among all variables.

When creating synthetic data, the analyst's goal must be to refrain from imposing prior beliefs about the relationships among the variables. Instead, the synthesizer must be constructed in a manner that allows existing relationships to be expressed with approximately the same degree of precision as they have in the underlying original data. When modeling a particular variable using the Sequential Regression Multivariate Imputation method that is our primary technique, all other variables, powers of these variables and interactions among these variables can potentially be used as explanatory variables even when such a relationship might not seem sensible to a researcher. Of course, due to feasibility constraints, the analyst must choose some subset of variables to go on the right-hand side of the predictive regressions but the goal remains to impose as few prior beliefs as possible. If the analyst is successful in specifying these components of the synthesizer, the result should be analytically valid synthetic data.

Section 6 gives a complete summary of the inference framework and computational formulae for assessing analytical validity. From a theoretical framework, the synthetic data will be analytically valid for the precise set of relations embodied in the posterior predictive distributions used for the synthesis. This theoretical result is reassuring to the

[2] A PIK is the Census Bureau's internal unique person identifier that replaces Privacy Act protected identifiers, like SSN, on files that have been approved for linking at the individual level.

extent that substantial computational power and flexible methods for estimating complex multivariate distributions can produce reliable posterior predictive distributions. Given the limits of current technology, however, the analytical validity of a particular synthetic data product must be directly assessed. Our method of assessment proceeds as follows. Parallel analyses of a large number of estimands are conducted on the completed Gold Standard data and on the synthetic data. The estimand is averaged over all implicates: four in the case of the completed confidential data and 16 in the case of the synthetic data. Next, the within and between implicate components of the estimand's variance are combined, according to rules that depend upon the precise multiple imputation method used, to generate an estimate of the total variance. The square roots of the diagonal elements of the total variance matrix and the appropriate degrees of freedom are used to form 95% confidence intervals for the completed and synthetic data estimates of the estimand. Ideally, the confidence intervals computed from the synthetic data should cover the confidence intervals computed from the completed data. At a minimum, there should be substantial overlap in the confidence intervals. The point estimates should also be "close," but this result is a by product of confidence interval coverage. In general, the confidence interval in the synthetic data will be wider than the interval computed from the completed confidential data for a specified nominal level, and this loss of precision is part of the cost of confidentiality protection. However, the width of the synthetic confidence interval can be reduced by increasing the number of synthetic implicates. A summary of our analytical validity results follows.

1.5.1 All univariate distributions

We compared the results for univariate distributions of all continuous variables using the first, fifth, tenth, twenty-fifth, fiftieth, seventy-fifth, ninetieth and ninety-fifth percentiles, and the means. Our synthesizer was designed to reproduce univariate distributions through the use of kernel density estimator transformations and inverse transformations. Our comparison of univariate results confirms that the synthesizer worked as designed. Only the wealth-related variables, which have notoriously skewed and multi-modal distributions, proved difficult to synthesize as measured by the univariate distributions. Even the kernel density estimators were not completely successful. One could as reliably compute univariate statistical tables representative of the civilian, non-institutional population age 18 and over on April 1, 2000 from the synthetic data as from the completed data.

We also compare the frequency distributions for categorical variables. These, too, are analytically valid as regards their univariate distributions.

1.5.2 Summary statistics for all workers and for OASDI beneficiaries

Although our synthesizer automatically develops models for subgroups when there are adequate sample sizes, the order in which the subgroups will be formed and tested for sample size adequacy is specified in advance. Consequently, one cannot say *a priori* that results will be analytically valid for all subsamples. We compared the results for all workers and for all OASDI beneficiaries using subsamples constructed on demographic variables and benefit type. This testing focused on important earnings and benefit measures. Work histories, average annual earnings, average indexed monthly earnings (AIME) or average monthly wages (AMW), primary insurance amount (PIA), lifetime earnings, and personal savings account accumulated balances are very similar between the synthetic and completed data files for all major demographic subgroups and all types of benefits. In general, the univariate confidence intervals in the synthetic data cover those in the completed data and are not excessively wide. These results hold whether the reference group is all persons age 18 or older or only OASDI benefit recipients. Overall, the version 4.0 synthetic data have almost complete analytic validity for these tests. This is a notable improvement over all previous versions and, in particular, over version 3.1. See section 6.4.3 for the detailed summary of the results for all workers and section 6.4.2 for the detailed summary of the OASDI recipient results.

1.5.3 Summary statistics by education and foreign born

We also studied summary statistics for several important variables by three-way interactions of race, gender, and education category. This analysis focused on earnings and benefits in 2000. Again, most point estimates were very close and synthetic confidence intervals covered the completed data intervals. In earlier versions, there were problems with certain educational categories. Some problems remain with the groups having no high school diploma or graduate degrees. These problems are usually that the confidence interval in the synthetic data is excessively wide–indicating that the synthesizer had trouble simulating these relationships and reflected a great deal of model uncertainty in the posterior predictive distribution. This is not surprising since these education categories, when cross-classified by race and gender, contain relatively few individuals in the Gold Standard file. See section 6.4.4 for a detailed summary of these results.

The same analysis was repeated for foreign born individuals, which is a four-way interaction in the underlying data. The results are generally encouraging but in this case the small Gold Standard sample sizes frequently produce very wide (or undefined) synthetic confidence intervals. See section 6.4.5 for a detailed discussion of the foreign born results.

1.5.4 Selected regression model results

We studied the coefficients in selected regression models, fit for the entire sample and for demographic subgroups. Our analysis of the logarithm of total Detailed Earnings Record wage and salary income (deferred and non-deferred at FICA and non-FICA-covered jobs) is representative of earnings analyses. All analyses are markedly improved over version 3.1 of the synthetic data. Most coefficients have some analytic validity–point estimates are similar and synthetic data confidence intervals significantly overlap completed data intervals. There is a detailed discussion of both the successes (most education categories, ethnicity) and the relative failures (actual labor force experience). The earnings analysis is repeated for other earnings measures with similar results.

The analysis of regressions modeling the logarithm of $AIME/AMW$ shows analytical validity for all major demographic groups and virtually all studied variables. The synthetic data can be used to model this variable almost as reliably as the completed data. This is remarkable considering that $AIME/AMW$ was not directly synthesized. Rather, it is derived from the synthetic earnings data.

We also studied regression models for the monthly benefit amount. We believe that the monthly benefit amounts are some of the most accurately synthesized variables and these regression results support this conclusion.

The analytical validity results for measures of earnings and wealth from the SIPP data were less successful. These variables have proven very resistant to synthesis by SRMI and, to date, no adequate alternative technology has been developed.

See section 6.4.8 for a detailed discussion of the regression results.

1.6 Disclosure avoidance assessment

The link of administrative earnings, benefits and SIPP data adds a significant amount of information to an already very detailed survey and could pose potential disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The creation of partially synthetic data is meant to prevent a link between these new public use files and the original SIPP public use files, which are already in the public domain. In addition, the synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information found in the summary and detailed earnings histories used to create the longitudinal earnings variables in the Gold Standard and public use files. Our disclosure avoidance research 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, which were used to create the SIPP component of our Gold Standard file.

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted two distinct matching exercises between the synthetic data and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, 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. However, for an actual re-identification of any of records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required–the intruder must make another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc*. Hence, the results from our experiments are very conservative estimates of re-identification risk. Nevertheless, we find that the re-identifiable records represent only a very small proportion of the candidate records, less than three percent using the most aggressive technology, and that these correct re-identifications are swamped by a sea of false re-identifications, which a real intruder would not be able to distinguish from the true re-identifications.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. In our case, the true match rate never exceeds three percent of the relevant candidate records. Second, the ratio of true matches to the total number of matches (true and false) should be close to one-half. We have performed two types of matching exercises, probabilistic and distance-based. The first criterion ensures that very few candidate re-identifications occur. The second criterion ensures that those candidate re-identifications are surrounded by substantial uncertainty as regards their correctness.

We conducted two types of record linking experiments to assess disclosure risk. The first experiment used the Census Bureau's internal probabilistic record linking software to attempt to re-identify the source record of a synthetic file observation in the Gold Standard file. The second experiment used four recently proposed distance-based record linking metrics to attempt the same reidentification. Both experiments were aggressive and conservative.

Aggressive record linking experiments use information that should not be available to a potential intruder but which is available to the analysts conducting the experiment. In our probabilistic record linking, we made aggressive use of the fact that we know the correct linkages between the Gold Standard and synthetic records to estimate the parameters of the agreement score that is used to find candidate matches. In our distance record linking, we made aggressive use of this same knowledge to estimate the full Mahalanobis distance between two records. Such a distance measure uses the covariance structure of the errors in synthesizing the data.

Conservative record linking strategies ensure that the estimated linking rates are upper bounds to what an intruder would calculate. In both experiments, we blocked on the unsynthesized SIPP variables. An intruder would do likewise. To reduce computational burdens, we also segmented the comparison files in a manner that ensured that the true match was always in the segment of the Gold Standard file that was compared to a segment of the synthetic file. Without prior knowledge of the true matches, which would make the record linking exercise superfluous, no intruder could reduce the computational burden with a similar strategy. Because both experiments could always find a correct link in their candidate records, while at the same time the number of at-risk records was artificially limited to reduce computation time, all estimated true match rates are over-estimates.

In the probabilistic record linking experiments, we found true match rates that never exceeded 1.2% overall. The ratio of true to false matches is always around 1/100 and never even approaches unity. In our distance record linking experiments, we found true match rates that never exceed 3%. The ratio of the true to false match rate is not as useful in distance record linking because the false match rate is always one minus the true match rate–every synthetic record has a best match in the Gold Standard file using distance linking techniques. We substituted an analysis of the true match rates based on using the best, second best, and third best distance record linking match candidate. Our analysis shows that there is considerable uncertainty regarding which of these three candidates is the correct match. The ratio of true matches associated with the second or third best candidate to true matches associated with the best candidate hovers around unity.

Both experiments clearly demonstrate that the partially synthetic SIPP/SSA/IRS PUF meets the standards of the Census Disclosure Review Board, which is expected to formally declare the version 4.0 file releaseable before the end of November. Because the public use file is also based on data from SSA and IRS, the consent of their disclosure review officers is also required before the file can be officially released.

1.7 Using the SIPP/SSA/IRS-PUF

This report includes a brief primer on using synthetic data. We explain how to calculate statistical measures on the different synthetic implicate files. Then we explain how to use the control variables placed on those files to properly compute confidence intervals and hypothesis tests. Our primer is not intended to be exhaustive. Rather, it provides a beginning user the wherewithal to process the PUF using standard statistical programming languages like SAS.

1.8 Next steps

Given the length and scope of this project, it is perhaps beneficial at this point to consider what has been accomplished. This collaboration between four government agencies has produced several new data products and advanced the body of knowledge on missing data imputation, assessing the validity of automated statistical modeling, disclosure avoidance techniques, and disclosure risk analysis. In the past six years, we have produced a highly useful compilation of SIPP data that combines five separate panels with edited administrative data from IRS and SSA, a weight to allow meaningful analysis of these combined panels, a set of files that multiply impute all missing data, and a set of synthetic data files that meet disclosure standards of the Census Bureau, the Internal Revenue Service, and the Social Security Administration. For the first time in 30 years, it appears that it will be possible to release lifetime earnings histories taken from administrative records, an accomplishment that will be of enormous benefit to the research community and the general population. This project has been a model for what inter-agency cooperation can accomplish by pooling the expertise of researchers from the Census Bureau, IRS, SSA, and CBO.

When we began this project, there was a great deal of uncertainty over whether synthesizing techniques could produce micro-data that would preserve relationships among variables and mitigate disclosure risk. In fact, almost none of the enhanced theory or experience with these methods required to complete the project existed. Based on the results at this point, we feel that both these questions can be answered in the affirmative. It is now imperative that outside users be given a chance to test these synthetic data and that the agencies involved develop a system for validating outside results using the Gold Standard in order to promote general confidence in the methods and to permit quality improvements. This process will help us to discover remaining flaws in the synthetic data and improve the synthesizing process, both of which will enable the collaborators to provide useful future updates to this data product, as funding resources permit.

2 Project Background

2.1 Purpose and brief history

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. To accomplish the goal of improving the Survey of Income and Program Participation (SIPP), the Census Bureau created an approved project entitled the "Demographic Survey Improvement Project" as a part of the Longitudinal Employer-Household Dynamics (LEHD) Program. Work began on the improvement of the SIPP and on the creation of a new public use file, which is the subject of this report. 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).

One of the primary goals of the survey improvement project was to create a new public use file that linked existing SIPP data with the administrative earnings data as well as administrative benefits data maintained by SSA. To this end, a joint committee was created with members from the Census Bureau, the SSA, the IRS, and the Congressional Budget Office (CBO). Individuals with related interests from the staff of the Joint Committee on Taxation (JCT) were also invited to participate. Committee members from the Census Bureau included John Abowd, Nancy Bates, Gary Benedetto, Pat Doyle, Judy Eargle, Sam Hawala, and Martha Stinson, who has served as the coordinator of the project since 2003. SSA has been represented by Susan Grad, Brian Greenberg, Howard Iames, and Dawn Haines. IRS members included Nick Greenia and Karen Masken. John Sabelhaus participated for the CBO. This committee has guided all major decisions concerning the creation of the public use file.

Beginning with fiscal year 2004, an Inter-Agency Agreement and subsequently a Jointly Financed Cooperative Agreement established an official jointly financed and sponsored project between the Census Bureau and SSA whose main purpose was the research leading to the improvement of the SIPP and the creation of the new public use file. Those agreements provide the basis of the financial and intellectual support for this work. This report summarizes the work done during fiscal year 2006 to finish the creation of the public use file. Inasmuch as the goal is to release a file for use by others outside the development group, this report also includes some history of the project where necessary to understand the final product.

2.2 Overview project description

From the beginning of the project, two over-arching requirements have guided the decisions made by the committee about the type of public use 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 criteria led to several other early decisions.

The first major design decision was that the file would contain records for individuals surveyed in one of five SIPP panels, 1990, 1991, 1992, 1993, and 1996, but the panel of origin for each individual would not be revealed. The decision to suppress the panel of origin for the individual was part of the overall confidentiality protection plan for the new PUF. The second major design decision was that the number of variables on the new public use file that came from the SIPP would be limited and would be chosen to facilitate national studies by retirement and disability researchers. The third major design decision was that the primary disclosure avoidance method would be to produce partially synthetic micro-data that could not be re-identified in the existing SIPP public use files. Thus, instead of containing the actual values of SIPP-reported variables, administrative earnings and benefits, the file would contain values that were draws from the joint posterior predictive distribution of the underlying variables conditional on the existing confidential data. The process of synthesizing data is described in detail in section 4.

The committee began its work by selecting the variables to include on the file. The selection process involved detailed discussions between all four agencies and consultations with outside researchers. As part of the process, the Census Bureau created a standardized extract of variables from each SIPP panel and merged these extracts with individual administrative earnings and benefits records. These extracts were combined and named the "Gold Standard" file. (See section 3 for a detailed description of this file.) The Gold Standard file has been revised many times during the past five years as new variables have been added, old ones dropped, and formatting for some variables changed. This file serves as the basis for the creation of the SIPP/SSA/IRS-PUF. 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 Gold Standard file contains data that are Title 13, Title 26, and Title 42 confidential because it commingles Census Bureau , IRS, and SSA data.

The next step in the process was to create a set of synthetic files that replicated the structure of the Gold Standard data. The Census Bureau produced the first such files in late fall 2003, and called it preliminary SIPP/SSA/IRS-PUF version 1.0. Since that time there have been three other preliminary public use files: version 2.0 (fall 2004), version 3.0 (December 2005), and version 3.1 (June 2006). The current preliminary public use file, which is expected to become final, is version 4.0, and is being delivered in conjunction with this report.

After each preliminary public use file was produced, committee members from each agency were responsible for reviewing the file to assess analytical validity and disclosure risk. Analytical validity tests have consisted of comparing univariate distributions, cross-tabulations, moments, and regression coefficients calculated from the synthetic and the completed Gold Standard data. (See section 6 for a detailed discussion of the analytical validity assessment.) Disclosure risk analysis has included probabilistic and distance-based record linking between the synthetic and the Gold Standard files. (See section 7 for a detailed discussion).

3 Creation of the Gold Standard File

The work of creating a new public use SIPP product with linked administrative data began with the creation of a base data set called the "Gold Standard" file. To create this file, we extracted variables from the five SIPP panels conducted in the 1990s (beginning in 1990, 1991, 1992, 1993, 1996, respectively) and merged SSA-provided administrative data from the Summary Earnings Records (SER), Detailed Earnings Records (DER), and the Master Beneficiary Record (MBR). This data compilation serves as the basis for the public use file. We refer to these data as the Gold Standard because they represent the available confidential micro-data that would be used for analysis by an authorized researcher working in a restricted-access facility. Any public use version of these data must, of necessity, closely reproduce the characteristics of the Gold Standard while at the same time taking steps to ensure the confidentiality of the actual data on the sampled individuals.

In this section, we describe each data source, list the variables chosen for inclusion in the Gold Standard file, and explain the major decisions made regarding different types of variables. A complete data dictionary for Gold Standard Version 4.0 accompanies this report. The data dictionary provides exact details about the creation of every variable in the Gold Standard, including the specific source SIPP, SER, DER and MBR variables used.

3.1 SIPP data

We chose the following demographic variables to be included on the Gold Standard file: gender, marital status, race (black/African-American), five categories of education, Hispanic ethnicity, birth date, death date, disability status, number of children, marital history, foreign born indicator, decade arrived in the United States if foreign born, and a spouse identifier that links to the marriage partner if the respondent is married and the spouse was also surveyed. We took the values for these variables at a point in time and thus none of these variables are time-varying in the Gold Standard file. For the time-invariant variables–gender, race, and Hispanic ethnicity, values were taken from the point in the survey when they were first reported, generally wave 1. Values for the other demographic variables were generally chosen from month 8 of the respective SIPP panel (*i.e.,* the last reference month of the second interview). We chose this point because marital, immigration, and disability histories were collected in the wave 2 topical modules and we wanted to take all these variables from the same point in time as nearly as possible. However, if an individual was not surveyed in wave 2 of the SIPP panel either because he or she exited the sample due to attrition from the panel after wave 1 or joined the panel in wave 3 or later, we took values for marital status and the spouse identifier from the closest available point in time. In other words, if marital status was missing in month 8, we checked for a marital status value in months 7, 9, 6, 10, 5, 11, 4, 12, 3, 13, 2, 14, 1, 15, 16, 17, and so on until the end of the panel. We chose the first non-missing value that was found. For individuals whose marital status was taken from a month other than 8, we chose the value for number of children from the same month as marital status. If this value was missing, we did not search in any additional months. For education, we searched over all reported education values in each wave of the SIPP and chose the highest level of education ever reported. Thus gender, marital status, race, education, Hispanic ethnicity, and spouse identifier are never missing in the Gold Standard data because these variables are all reported at some point in the SIPP, and we chose to take self-reported values whenever they were available. Disability status, number of children, marital history, foreign born indicator, and decade arrived in the United States if foreign born are sometimes missing because individuals did not answer the relevant topical modules or because we chose not to search over every available month of SIPP data.

The marital history variables are some of the most complicated historical variables on the Gold Standard file. Most of the information in this history came from the marital history topical module collected in wave 2 of each panel. We supplemented this information by creating a short marital status history that covered the period of the panel from wave 2 forward by using the marital status reported in each month. In the topical module, individuals could report 0, 1, 2, 3, or more than 3 marriages. Dates for the beginning and end of first, second, and most recent marriages were then collected, as well as the reason for a marital termination (death or divorce/separation). If an individual had more than 3 marriages, no dates for those marriages between the second and most recent were collected. We used our short history from the panel period post-wave 2 to check for additional marital events: beginning of a new marriage or ending of an existing marriage. We took account of at most one additional marital history event for an individual. We summarized all this information in a set of 16 variables. They include: $mh\_category$, a categorical variable that classifies individuals according to their number of marriages and the type and order of the endings of those marriages; $mh1 - mh7$, a set of flags that provides the same information as $mh\_category$ but broken down by event; $flag\_mar4t$, an indicator for whether the individual was missing a marriage because the SIPP only collected information on three marriages; age at the time of every reported marital history event.

It is important to understand that the marital history variables may differ from the marital status variable described earlier. In particular if a person reports being married in wave 2, month 8 but is not married at the end of his or her history, this is because a divorce or death occurred over the course of the SIPP panel. Similarly, if a person reports not being married in wave 2, month 8 but is married at the end of his or her history, this is because a marriage occurred during the course of the SIPP panel. Although a person may report only 3 marriages during the topical module, it is possible to have a fourth marriage as part of the marital history because the last marriage occurred during the course of the SIPP panel. However no more than 4 marriages can be recorded in the SIPP from all available sources. There are also some things that must be consistent between the marital history and the wave 2 marital status. If the person reports being divorced or widowed in wave 2, month 8 then at least one divorce or widowhood much occur during that individual's marital history. Likewise if the person reported being married, then the history must contain at least one marriage.

Birth date and death date are unique variables in both their source and treatment. We originally extracted birth date from the first self-reported value in the SIPP survey. However, after several discussion between the Census Bureau and SSA about the measurement error likely to be contained in this variable, we switched to using an administrative birth date. Thus, in the final version of the Gold Standard, we create a variable called $birthdate\_pcf$ from the Census Personal Characteristics File (PCF), an administrative database that has as one of its inputs the Social Security Numident file. Any individual that has applied for a Social Security Number (SSN) has a record in the Numident file that contains, among other things, SSN and birth date. The administrative record birth date variable ($birthdate\_pcf$) serves as the basis for the synthesis process and is comparable to the variable $birthdate$ in the synthetic public use files. We chose the administrative source for this important variable in order to insure as much consistency as possible between the administrative earnings and benefits variables and age. Using administrative birth date helps to avoid cases where it appears that people receive retirement benefits prior to age 62, a legal impossibility caused by self-reported birth dates that are several years later than the actual dates.[3] We also included a variable called $birthdate\_sipp$ on the Gold Standard file in order to facilitate re-identification tests that attempt to link the synthetic data back to the Gold Standard. Since the administrative birth date is not an existing SIPP public use variable, anyone attempting to link the new synthetic SIPP/SSA/IRS-PUF to the existing SIPP public use data would have to use the SIPP reported birth date for this purpose.

Death date was extracted from the same PCF file as administrative birth date. In this case, however, the original sources were the Numident file and SSA's Death Master File (DMF), another supplementary file that the Census Bureau receives from SSA that reports deaths every month. The link between SIPP respondents and the PCF was performed using a validated SSN, a process described in more detail in section 3.2.

For economic variables, we included the following annual time series, beginning in 1990 and ending in 1999: weeks worked with pay, weeks worked part-time, total annual hours, family poverty threshold, total family income, total personal income, total personal earnings, welfare program participation and amount of payments, private health/disability program participation and amount of payments, and health insurance coverage and type. Since no individual was followed by a SIPP panel for more than 4 years, these time series arrays contain at least 6 years of missing data for every individual. The exact number and timing of missing years depends upon the original panel. Individuals surveyed in the 1990 panel are missing 1993-1999 data whereas individuals surveyed in the 1996 panel are missing 1990-1995 data. We included the following point-in-time variables: industry and occupation for the main job, chosen from the first available wave, and total net worth, home ownership, home equity, non-housing wealth, and indicators for defined benefit and defined contribution pensions, all taken from topical modules.

Some SIPP variables were purposely omitted from the public use file in order to minimize disclosure risk. Specifically, no data are provided on geography. We include a state of residence variable Gold Standard file but will not release this variable on the public use file. The exact linkage of spouses is the only family relationship data on the file. No other family relationship data are provided on either the Gold Standard or public use files. No panel dating information is provided on the public use file although we retained the panel source variable in the Gold Standard data

[3]It is worth noting that the administrative birth date is not without some error. Unlike the SIPP reported birthdate, which was edited prior to public release to produce a set of plausible ages, the administrative birthdate contains some values that make individuals in our sample 100 years old or more. However the number of these cases is very limited and we feel that this small error is out-weighed by the general accuracy gains and the benefits to disclosure avoidance.

to facilitate evaluation and testing.

An individual was eligible to be included in the Gold Standard file if he or she met one major requirement: the individual must have been at least 15 years old at the time of the second wave of the SIPP panel in which that person was interviewed. We chose this age because at 15 or older, the SIPP considered the individual to be an adult and asked the full battery of questions. In order to make this determination, we used the variables $popstat$ (1990-1993 panels) or $epopstat$ (1996 panel) from the wave 2 core data. For those who were not interviewed in wave 2, their age at the end of wave 2 was calculated and if they would have been at least 15, they were kept in sample. It is important to note that these age calculations were done using the self-reported SIPP birth date. We did this in order to reproduce the survey determination of who was eligible to be treated as an adult as accurately as possible in the new public use file.

3.2 IRS/SSA earnings data

Administrative earnings data were extracted from the Master Earnings File (MEF), a historical compilation of earnings reports filed with the IRS by employers (most commonly using the W-2 form). This administrative database is maintained by SSA for the purpose of calculating benefits when workers retire or become disabled. We receive earnings in two forms: Summary Earnings Records (SER) and Detailed Earnings Records (DER). The SER data contain total personal earnings capped at the FICA taxable maximum for each year from 1951-2003. The DER data contain uncapped earnings broken out by employer from 1978-2003. In the Gold Standard file we include the entire annual SER history plus total earnings from 1937 to 1951. From the DER, we create total annual earnings from FICA covered jobs by summing earnings from each employer that was required to withhold social security tax. We also create total annual deferred earnings from FICA covered jobs by summing deferred wages (*i.e.*, contributions to 401(k) plans) from these same employers. We create an analogous set of earnings and deferred earnings variables that pertain to jobs not covered by FICA. Thus, in each year from 1978-2003, the SER earnings variable indicates the amount of FICA covered earnings in a year, up to the taxable maximum, and the set of four DER earnings variables indicates total earnings, uncapped, split between deferred and paid, and FICA and non-FICA covered jobs. The sum of the two DER earnings variables that represent paid wages gives total wages and salary that an individual would report on IRS Form 1040 and which would be taxable under federal income tax laws.

These IRS/SSA earnings variables are matched to the SIPP extracts using a validated Social Security Number. The 1990s SIPP panels collected the SSN from respondents. Using name, address, birth date, gender, and race information, the Census Bureau and SSA validated these self-reports against the SSA Numident file. If the demographic variables collected by the SIPP matched the demographic variables associated with the reported SSN on the Numident, then the SSN was declared valid.[4] If the demographic variables did not match, an alternative SSN was sought. For individuals who reported that they did not know their SSN, an SSN was sought in the PCF file based on these demographic variables. For individuals who refused to provide an SSN, no match was sought in the PCF and we did not receive earnings records for these individuals. Thus, for individuals without valid SSNs, all the administrative earnings arrays described above were treated as missing data. Approximately 12% of individuals in the gold standard did not have valid SSNs and were, consequently, missing MBR, SER and DER data.

3.3 SSA data

In addition to administrative earnings records, the Census Bureau also received records for SIPP respondents containing information about the type and amount of benefits paid under the Old Age, Survivor, and Disability Insurance Program (OASDI). These SSA data were contained in the Master Beneficiary Record (MBR) file and were linked to the SIPP data using the same method as the earnings data (*i.e.,* validated SSN). The MBR is an extensive and complicated file and, after much deliberation, the decision was made to include only a few variables from it on the Gold Standard file. Specifically, we included the date of initial entitlement to OASDI benefits, the initial reason for receiving these benefits (TOB), and the initial monthly amount paid (MBA). We also included the type and amount of benefit received in April 2000. Using the formulae published by SSA, we calculated average indexed monthly earnings ($AIME$) or average monthly wage (AMW) and the primary insurance amount (PIA) from the administrative earnings history and included these on the Gold Standard as a help for researchers. However, it is important to note that the $AIME/AMW$ and the PIA contain no information not already represented in the earnings history. Thus, they can be recreated in the Gold Standard, the completed Gold Standard, or the SIPP/SSA/IRS-PUF data using

[4] Prior to 2003, the process of validating an SSN was performed by a clerical edit using the same information.

alternative assumptions.

3.4 Weight creation and use

One concern that arose early in the process of creating the public use file was the provision of proper weights for a file that pooled SIPP respondents from five separate samples. The 1990-1993 panels contain some overlapping time periods. The official SIPP public use file documentation explains how to pool the published weights for those panels in order to construct a weight that has a well-defined reference population and reference date. However, there is no design guidance for pooling the individuals from all five SIPP panels in order to produce estimates representative of a well-defined target population at a known reference date. In addition, the different SIPP panels over-sample low income individuals and other groups at differential rates. Hence, these survey data can only be used to construct estimates representative of the U.S. civilian non-institutionalized population as of a particular date if an appropriate weight is provided. Thus, another major data activity for this project has been to create an *ex post* weight for the individuals in the Gold Standard file such that each person's weight indicated how many persons in the reference population that SIPP person represented as of a known date. The designated reference population is all individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000. A full report on the details of this process is provided in section 5.

3.5 Gold Standard data dictionary

The Gold Standard data dictionary for version 4.0 is included as an appendix to this report.

4 Data Completion and Synthesis

4.1 General methodology

In this section, we describe the basic theoretical framework for creating synthetic data. The notation and definitions follow Rubin (1987), which treats multiple imputation of missing data, and Rubin (1993), which is the first paper to define the use of fully synthetic data for confidentiality protection. We adopt enhancements for the application of Sequential Regression Multivariate Imputation (SRMI) to synthetic data from Raghunathan, Reiter and Rubin (2003). We use the formal inference methods for multiple imputation-based partially synthetic data from Little (1993) and Reiter (2003). Finally, we incorporate the formal inference methods for multiple-imputation based partially synthetic data that also have missing data from Reiter (2004). We have attempted to make the notation consistent in this section. Hence, it does not match the original authors' notation.

A finite population contains N entities whose characteristics are known and constitute the f columns of X, $(N \times f)$. A sample of size $n < N$ is drawn from the population. Let the vector $I\ (N \times 1)$ be defined as $I_i = 1$ if entity i is sampled and $I_i = 0$, otherwise. Data are collected for p variables denoted by the matrix $Y\ (N \times p)$. Note that the matrix Y is defined for the entire population, not just for the sampled units. Of course, some elements of Y are missing because the entity that constitutes that row was not sampled. Other elements of Y are missing because of item non-response in the sample. (In administrative data, item non-response is equivalent to missing data items on an in-scope administrative record.) Let the matrix $R\ (N \times p)$ be defined as $R_{ij} = 1$ if the data represented by item Y_{ij} are available in the sample and $R_{ij} = 0$, otherwise. Certain submatrices of Y and R are of interest. Let Y_{inc} $(n \times p)$ be the submatrix of Y that corresponds to the rows for which $I_i = 1$. So Y_{inc} contains the data for all the sampled entities. The complement of Y_{inc} is Y_{exc}, the rows of Y that correspond to the rows for which $I_i = 0$. So Y_{exc} contains the data for all the unsampled entities. Similarly, let $R_{obs}\ (n \times p)$ be the submatrix of R corresponding to the item missingness for the sampled entities; *i.e.*, those rows for which $I_i = 1$. Finally, define the submatrices Y_{obs} and Y_{mis} as follows

$$Y_{obs,ij} = \left\{ \begin{array}{c} Y_{ij},\ \text{if } I_i = 1 \text{ and } R_{ij} = 1 \\ undefined, \text{otherwise} \end{array} \right\}$$

and

$$Y_{mis,ij} = \left\{ \begin{array}{c} Y_{ij},\ \text{if } I_i = 1 \text{ and } R_{ij} = 0 \\ undefined, \text{otherwise} \end{array} \right\}$$

So, the matrix Y_{obs} contains all the sampled values of Y_{ij} that contain data and the matrix Y_{mis} contains all the sampled values of Y_{ij} that are item missing. The observed data are summarized by the set D=$\{X, Y_{obs}, I, R\}$. The following table gives a summary of all these definitions.

		General Definitions
N	=	number of individuals in the population
$X\ (N \times f)$	=	population characteristics of f variables for N individuals,
		f variables are known for all N individuals in the population
p	=	number of variables for which survey/admin. systems will collect data
$Y\ (N \times p)$	=	data on p variables for N individuals; only sampled values available
$I\ (N \times 1)$	=	identifies which individuals from the population were sampled,
		i.e. tells which rows of Y are non-missing
$R\ (N \times p)$	=	identifies non-missing elements,
		i.e. tells which variables are non-missing for which individuals
Y_{obs}	=	observed data, submatrix of $Y\ (N \times p)$ that contains only elements
		where individual was sampled and provided data on specific variable
Y_{mis}	=	missing data, submatrix of $Y\ (N \times p)$ that contains elements where
		individual was sampled but did not provide data on specific variable
D	=	$\{X, Y_{obs}, I, R\}$ or all known data about individuals in survey sample

In the context of our public use file, the above notation applies as follows: $Y\ (N \times p)$ is a matrix with one row for every member of the U.S. population age 15 and older at any time between January 1, 1990 and January 1, 1996 and one column for each of p variables that describe these individuals. In our case, there are 173 SIPP variables, 443 SSA/IRS earnings variables, and 5 SSA benefit variables so $p = 621$ and $N = 287$ million. $I\ (N \times 1)$ contains one row for every member of the U.S. population age 15 and older and $I_i = 1$ when an individual was surveyed by the Census Bureau using the 1990, 1991, 1992, 1993, or 1996 SIPP survey instrument. The matrix I defines $Y_{inc}\ (n \times p)$ which is a submatrix of $Y\ (N \times p)$. The I matrix tells which n rows from the population Y matrix were sampled into one of the five SIPP panels and eligible according to age to be in the gold standard: $n = 261,000$. $R(n \times p)$ is a matrix that records which of the n SIPP respondents are missing responses to which of the p variables. $R_{ij} = 1$ if person i has non-missing data for variable j. The R matrix defines $Y_{obs}\ (n \times p)$ which contains the data we actually observe. The following table provides a summary of these definitions.

Specific Definitions for SIPP/SSA/IRS-PUF

N	$=$	287 million, i.e. population of U.S.
$X\ (N \times f)$	$=$	race, gender, and birth date, known for all individuals on Census short form
p	$=$	173 SIPP variables, 443 SER/DER variables, 5 SSA benefit variables
$Y\ (N \times p)$	$=$	data on all the above variables for entire U.S. population
$I\ (N \times 1)$	$=$	identifies which individuals from the population were sampled by the SIPP and included in gold standard
$R\ (N \times p)$	$=$	identifies which SIPP and administrative variables are non-missing for which individuals
Y_{obs}	$=$	observed data, submatrix of $Y\ (N \times p)$ that contains only elements where individual was sampled by SIPP and data is non-missing
Y_{mis}	$=$	missing data, submatrix of $Y\ (N \times p)$ that contains elements where individual was sampled by SIPP but did not provide data on specific variable
D	$=$	$\{X, Y_{obs}, I, R\}$ or all known data about individuals in the SIPP samples

In the classic Rubin (1987) missing data application, Y_{mis} is imputed m times by sampling from $p\left(Y_{mis}\left|D\right.\right)$, the posterior predictive distribution of Y_{mis} given D. The completed data consist of m sets $D^{(\ell)} = \left\{D, Y_{mis}^{(\ell)}\right\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw from $p\left(Y_{mis}\left|D\right.\right)$ and is called the ℓ^{th} implicate. The basic insight for using synthetic data as part of a confidentiality protection system is that sampled individuals can be treated as having missing data for some or all variables even if they provided valid data. When these data are "completed" in the same manner as described above, namely by drawing from the posterior predictive distribution of Y_{mis}, $p\left(Y_{mis}|D\right)$, a file is produced that remains statistically valid but no longer contains the sampled individuals values for the variables that were synthesized.

In our application of synthetic data methods to the linked SIPP and administrative data, we first use Rubin's general multiple imputation method to complete our missing data. Next, we used this same method to create synthetic data. It is important to note that data resulting from this process are most accurately described as "partially" synthetic data. The terms "partially synthetic" and "fully synthetic" are now used in the statistics literature to distinguish between two related synthetic data generating models. Partially synthetic data are created using an actual sample of the population (*i.e.* the actual SIPP surveys) as source records so that a record in the partially synthetic data is based upon an actual record from the underlying survey. Fully synthetic data are created by sampling from a synthetic population in which the unsampled entities from the original survey have synthetic values for all variables from the survey. Fully synthetic samples are created by using all the known population characteristic variables to generating synthetic values for all survey variables conditional on the known population characteristics. Thus fully synthetic implicates do not have an actual source record in the original survey and can be described as fictitious entities. This project did not attempt to create fully synthetic data.

The major focus of the synthesizing process is to obtain a good estimate of the posterior predictive distribution

(PPD) for all the variables to be completed and synthesized. We now discuss the computational formulae for estimationg and sampling from the PPD. More general methods exist, such as Markov Chain Monte Carlo, but the methods summarized herein are the ones used by this particular project.

To begin, an explicit representation of D is required. As defined above $D = \{X, Y_{obs}, I, R\}$. While, in principle, the analyst at the Census Bureau has access to X, the population characteristics, in the applications described in this section, only the rows of X corresponding to $I_i = 1$ are used.[5] Hence, there is no practical difference between X and Y_{obs} for our synthetic data modeling. Complete data are guaranteed for X but nevertheless many variables in X require confidentiality protection before they can be placed in a public use data file. In this section, we adopt the notational convention that a variable appears in X if it is always available when $I_i = 1$ and it never requires confidentiality protection. Otherwise, the variable is included in Y_{obs}. This set of X variables can be empty without affecting the discussion below.

We describe two methods: Bayesian bootstrap (BB) and SRMI.[6] In both of these methods, we apply the principle of estimating the conditional distribution of group of variables (columns of Y) conditional on all other columns. For each distinct group of variables in Y, the columns of D are partitioned into four mutually exclusive sets: grouping variables, conditioning variables, dependent variables, and ignored variables. Grouping variables are used to stratify D such that a separate PPD is estimated in each stratum. Conditioning variables are a list of potential right-hand-side variables to be entered linearly in model-based estimation of the PPD. Dependent variables are those for which the PPD is being estimated. Finally, ignored variables are all other columns of (X, Y_{obs}). For purposes of doing the computations below, the data matrix (X, Y_{obs}) should be interpreted as including any variables that have been calculated as exact functions of the available data. Hence, the dimensionality of the matrices used below potentially exceeds $f + p$.

4.2 Bayesian Bootstrap

The Bayesian bootstrap was originally defined by Rubin (1981). As explained therein, the BB is used to simulate the posterior distribution of the parameter whereas the regular bootstrap simulates the sampling distribution of the parameter. Whereas a conventional bootstrap assumes that the sample CDF is equal to the population CDF, the BB properly accounts for the uncertainty of the sample CDF.

4.2.1 Generic BB algorithm

The notation used to describe the BB algorithm in this subsection is generic and does not refer to the matrices defined elsewhere. Let X $(n \times k)$ be the source data matrix and Y $(s \times k)$ be the target data matrix. This means that we want to construct an $s \times k$ Bayesian bootstrap sample from an $n \times k$ matrix of source data. Each BB replicate ℓ is a unique $Y^{(\ell)}$.

1. Draw $n - 1$ random variables from $U(0, 1)$.
2. Sort u_i ascending and let $u_{(i)}$ denote the order statistics from lowest to highest. Define $u_{(0)} = 0$ and $u_{(n)} = 1$.
3. For $i = 1, \ldots, n$, let $\hat{p}_i = u_{(i)} - u_{(i-1)}$.
4. For $j = 1, \ldots, s$ sample with replacement from the rows X using $\hat{p}_i$ as the probability of selecting row i. Place the sampled row into Y_j.
5. Repeat from step 1 for as many BB replicates as desired.

In other words, beginning with a data matrix, X, that contains values for the k variables of interest, this process assigns a probability of choosing a given observation from X to provide data to a corresponding observation in Y for the k variables. The set of probabilities constitutes a non-parametric representation of the posterior distribution from which the sampling is done. In a conventional bootstrap, because of the assumption that the sample CDF is equivalent to the population CDF, each observation in X would be assigned probability $\frac{1}{n}$ of being chosen. There would be no

[5] An exception is the process used to create and synthesize the *ex post* weight, which is described in section 5. In that process the full matrix X was used.

[6] For a description of the Bayesian bootstrap see Rubin (1981). For a description of SRMI in its original application to missing data problems see Raghunathan et al. (1998).

uncertainty in what probability would be assigned to a given observation. However, the Bayesian bootstrap accounts for the fact that the sample CDF is not the population CDF and hence does not assign equal probability to each observation. To better understand this concept, consider the example of $k = 1$ where the variable of interest, x_1, is an indicator variable. Suppose that for 75% of the sample of individuals, $x_1 = 1$ and that $x_1 = 0$ for the remaining 25%. In a conventional bootstrap, with each individual assigned a probability of $\frac{1}{n}$ of being chosen, the CDF used for sampling would always give $x_1 = 1$ a 0.75 probability and $x_1 = 0$ a 0.25 probability. The resulting target matrix Y would not necessarily have a realized 75%/25% frequency distribution for the two values for x_1 but all the bootstrap samples would have been drawn from such a distribution. In a Bayesian bootstrap, when each source record is assigned a unique probability whose expected value is $\frac{1}{n}$, the CDF used for BB sampling might have 73% versus 27% probability of drawing $x_1 = 1$ or 0. The next BB might have 76% versus 24%. The variation in the BB probabilities reflects the fact that the sample proportion of 75% in X is an estimate of the probability that $x_1 = 1$.

4.2.2 Bayesian bootstrap application

Choose grouping variables such that the rows of (X, Y_{obs}) can be assumed to come from the same joint distribution within each group defined by the unique combinations of values of the grouping variables. Some collapsing of categories may be required and is described later under implementation details. What is required is the creation of G groups based on the values of the variables in the grouping variable list. It is essential to the success of the Bayesian bootstrap in accurately replicating statistical properties of the data that the observations in a given source (donor) group and a given target (donee) group be as homogenous as possible. Thus, ideally, a large list of grouping variables should be chosen initially. One of the main advantages of the Bayesian bootstrap is that the group sizes do not have to be as large as groups where parametric modeling is done. Another advantage that is described below is that groups of dependent variables can be done together. This method also helps to preserve the statistical properties of the data by keeping intact relationships among variables.

In the BB application, none of the grouping variables can contain missing data. There are no conditioning variables because no linear model is used. The dependent variables consist of all columns j of Y for which $R_{ij} = 0$ for some i. The ignored variable list consists of all variables that are neither grouping variables nor dependent variables. We first describe the application of BB to the missing data problem. This is complicated if the missing data pattern is non-monotone as defined in Rubin Rubin (1987). For the moment, assume that the missing data pattern is monotone. Then, proceed through the dependent variables in groups constructed as follows:

1. All dependent variables with missing data exactly comparable to the variable with the least missing data; *i.e.,* all j for which $R_{ij} = 0$ if and only if $R_{ij^*} = 0$, where j^* is the column index of the variable with the least missing data. This is dependent variable group 1.

2. Remove all variables from the dependent variable list that are already in a group. Let j^* represent the column index of the variable with the least missing data from among those dependent variables that remain. Group all dependent variables with missing data exactly comparable to the variable indexed by j^*; *i.e.,* all j for which $R_{ij} = 0$ if and only if $R_{ij^*} = 0$. This is dependent variable group h.

3. Increment h and repeat step 2 until no dependent variables remain.

This defines H dependent variable groups. Initialize the BB missing data algorithm by placing all dependent variables into the ignored variable list and setting $h = 1$.

1. Remove the variables in group h from the ignored variable list and place them in the dependent variable list.

2. For $g = 1, \ldots G$, BB the rows of Y_{mis} (target data matrix) using the rows of Y_{obs} as the source data matrix. Repeat the BB m times to get m imputations $Y_{mis}^{(\ell)}$.

3. Put the dependent variables in group h back into the list of ignored variables.

4. If $h < H$ then increment h and return to step 1; otherwise, stop.

The result is m completed data sets. When the missing data are not monotone, the BB algorithm can be used to get starting values for other algorithms described below, in particular, SRMI. The BB algorithm can also be used for synthesizing data. In this case, simply treat all observations as missing and use the above steps to find donors for every individual in the data.

4.3 Sequential Regression Multivariate Imputation

Sequential Regression Multivariate Imputation (SRMI) was first proposed as a general technique for multiple imputation of missing data by Raghunathan et al. (1998). Raghunathan, Reiter and Rubin (2003) extend the method to confidentiality protection. Abowd and Woodcock (2001) use the SRMI method for confidentiality protection combined with missing data imputation. Although the formulae for SRMI can be stated generically using joint probability distributions like $p\left(Y_{mis}|D\right)$, almost all applications assume that the entities that constitute the rows of (X, Y_{inc}) have been sampled independently. Nothing in the generic statement of the problem prohibits dependent sampling; however, as a practical matter, formalizing this dependence while implementing SRMI is complicated. Abowd and Woodcock (2001) illustrate these complications for the case of longitudinally linked employer-employee data. The algorithms are summarized below ignoring the complications associated with dependent sampling.

4.3.1 Definitions and general algorithm

In SRMI, the analyst cycles iteratively through the dependent variable list. In any given iteration, conditioning data may be taken from either the current or the previous iteration depending upon the location of the current dependent variable in the variable list. For missing data applications, the procedure is normally iterated until the effect of this conditioning has been minimized. In synthetic data data applications, the conditioning values are the same regardless of the position of the variable in the dependent variable list and so iteration is not required.[7]

Let Y_j denote the current dependent variable and let $Y_{\sim j}$ denote all other columns of Y. The general algorithm is most cleanly stated for the missing data case. The refinements for the partially synthetic data case will be noted below.

For each dependent variable, the analyst selects grouping variables, conditioning variables and ignored variables. The grouping variables stratify the estimation into G mutually exclusive and exhaustive groups as illustrated in section (4.2.2) . The conditioning variables may include all columns of $\left(X, Y_{\tilde{}j}\right)$, including columns that are created to allow for nonlinearities in the conditional relations. The ignored variables are all columns of $\left(X, Y_{\tilde{}j}\right)$ not included among the conditioning variables. We wish to generate m implicates $Y_{mis}^{(\ell)}$. SRMI is an iterative procedure. Denote the interim values of implicate ℓ as $Y_{mis}^{(\ell,s)}$. Initialize $\ell = 1$ and $s = 1$. Initialize $Y_{mis}^{(1,0)}$ using Bayesian bootstrap methods.

1. For $j = 1, \ldots, p$:

   (a) If $\ell = 1$ then estimate

   $$p\left(Y_j|X, Y_{obs,\tilde{}j}, Y_{mis,1}^{(\ell,s-1)}, \ldots, Y_{mis,j-1}^{(\ell,s-1)}, Y_{mis,j+1}^{(\ell,s)}, \ldots, Y_{mis,p}^{(\ell,s)}\right)$$

   (b) Fill $Y_{mis,j}^{(\ell,s)}$ with data sampled from

   $$p\left(Y_j|X, Y_{obs,\tilde{}j}, Y_{mis,1}^{(\ell,s-1)}, \ldots, Y_{mis,j-1}^{(\ell,s-1)}, Y_{mis,j+1}^{(\ell,s)}, \ldots, Y_{mis,p}^{(\ell,s)}\right)$$

2. If converged then

   (a) Set $Y_{mis}^{(\ell)} = Y_{mis}^{(\ell,s)}$.

   (b) Increment ℓ.

   (c) Reinitialize $Y_{mis}^{(\ell,0)} = Y_{mis}^{(\ell-1,s)}$

   (d) Reinitialize $s = 1$

[7] An exception to this statement occurs when the data to be synthesized have exact logical dependencies among the variables. In this case a parent/child tree is used to coordinate these dependencies. The conditioning data for a particular variable will include the results of the synthesis of variables that were antecedents in the parent/child tree (parents). Iterating this process, however, simply produces another synthetic implicate.

3. If $\ell \leq m$, go to 1.

The test for convergence is not formal. In practice s is often limited to 10 or less. The algorithm estimates the joint distribution $p(Y_{mis}|D)$ by iterating over each conditional distribution $p(Y_{mis,j}|D)$ and filling the "data matrix" with imputed values based on the previous iteration's estimate of $p(Y_{mis}|D)$. Once the estimation has converged, the implicates are all drawn from the same estimate of $p(Y_{mis}|D)$. However, the completion of D for each implicate results in different conditioning data for the draws. In the implementation of the algorithm, one cycles over the grouping variables $g = 1, \ldots, G$ performing the entire algorithm for each homogeneous group. In steps $1.a$ and $1.b$ only the conditioning variables appropriate for $Y_{mis,j}$ in conditioning group g are actually included in the conditioning set. The initial selection of these variables is dependent on the analyst. However after the variables are tentatively included in the model the Bayes Information Criterion (BIC) is used to reduce the variable list by eliminating variables that have a posterior odds ratio below a pre-specified level. The posterior odds ratio cutoff for the BIC value in this variable selection mechanism can be controlled by the analyst. See Abowd and Woodcock (2001) for details.

4.4 Summary of synthetic data production

We now provide specific details about the process used to create synthetic data for this project. The first step of the process was to multiply impute all missing data. Missing data in our sample are due to survey item non-response and to out-of-scope survey years. Failing to provide an answer to the question about whether an individual was born in the United States or a foreign country is an example of item non-response. Missing income in 1996 because the individual was surveyed in the 1990 SIPP panel, which ended before 1996, is an example of missing due to out-of-scope survey years. The goal of the first step is to impute values for every variable whenever it is missing due to item non-response or out-of-scope survey years. We call this "completing the data," because the result of this first step is a set of files that contain all the original data plus imputed values when the original data were missing. Each one of these files is then referred to as a "completed" data set.

Regular missing data, which we multiply impute, result from item non-response or an out-of-scope survey year. 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. Stucturally missing data still exist in our completed and synthetic data–every individual will not necessarily have a value for every variable. For example, an individual who was born in the United States will have structurally missing data for the variable that indicates which decade the person immigrated to the United States. For survey data, structurally missing values occur when the skip logic of the survey dictates that a question is not appropriate because of the response given to a prior question. Administrative record data have a similar, albeit implicit, structure. Statisticians usually call such values "structural zeroes." Structurally missing data are never completed (*i.e.*, imputed) because they do not represent missing information. In contrast, regular missing data, which we complete by multiple imputation, do represent a failure on the part of the survey or administrative records to capture certain information. In this report we use the term "missing" to mean "missing-to-be-completed" and will explicitly describe any other data that are missing as "structurally missing."

Completing data involves choosing a model for each variable with missing data. We used the SRMI methodology to impute missing values for most of the SIPP variables. The few exceptions are described in 4.5.5 where we give details about the modeling for each variable. We used the BB technique to handle missing data due to missing SSNs. When an individual failed to provide an SSN that could be validated, we could not link that individual to the administrative databases (PCF, SER, DER, and SSA benefits) and, as a result, several hundred administrative variables were missing. One approach to this problem would have been to use the SRMI methodology to model each individual administrative variable and impute missing values. However the magnitude of this task and concern about the need to preserve internal consistency among all the administrative variables, led us to choose the BB completion method for the SSN variable. This method allowed us to choose an appropriate donor record with a non-missing SSN which provided the complete set of administrative variables: PCF (birth date, death date), SER and DER (earnings), and MBR (SSA benefits). Once the SSN had been completed, we treated all administrative data as completed. If a validated SSN did not have a record in a particular administrative database, we treated these data as structurally missing. In other words, no Master Beneficiary Record meant the person had not received benefits from SSA under a program that would generate an MBR record and no DER job records meant the person had not earned federally taxable income since 1978. Once again, in the completed administrative data, an individual does not have a value for every variable. But individuals who were originally missing SSNs now have donated SSNs which link to administrative data that is either present or structurally missing.

One important feature of how we applied the BB is worth mentioning. When both members of a married couple were both missing SSNs, we chose a donor couple based on couple characteristics instead of two separate individual donors. In this way we hoped to preserve the important effects of marriage on SSA benefits. When only one member of a couple was missing an SSN, we also chose a donor couple based on couple characteristics but then only used the donated SSN for the couple member with the missing SSN. By using this method, we were able to choose a donor donor couple that resembled the couple with the single missing SSN and a donor spouse who looked liked the donee and was married to someone who looked like the donee's spouse.

The actual process of completing data is iterative. We begin with a base data set that contains only original, non-missing data. We then use the BB to complete the SSN and hence the administrative data. Donors are chosen on the basis of non-missing SIPP variables. This data set serves as the input for the SIPP data completion stage using SRMI. Models are estimated using originally non-missing dependent variables and any available non-missing explanatory variables from either the administrative or SIPP data. Variables are modeled beginning with the variable with the least missing data and progressing to the variable with the most missing data. As models are estimated and missing values are imputed, the data set is updated to include the imputed values. Hence, for the first variable modeled, almost all other SIPP variables will have missing values and hence a number of cases will be excluded from the estimation in this first round. As variables are completed and the data set is updated, there will be fewer and fewer missing values, and increasingly more cases available for model estimation. The end product of the SRMI process is a data set that contains completed administrative and SIPP variables.

We then iterate the process. We perform the Bayesian bootstrap again to complete the SSN, this time using the updated, completed SIPP variables from the end product of iteration 1. We then use SRMI to estimate models for the SIPP variables again. As in the first iteration, only originally non-missing dependent variables are used in model estimation. However beginning with the second iteration, the first variable to be modeled uses explanatory variables from the completed data that was the output of iteration 1. This prevents the exclusion of any cases due to missing data. 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 SRMI step, another completed data set is generated which has updated values for all the SIPP and administrative variables.

As part of the creation of version 3.0 of the preliminary public use file, we performed 8 iterations of missing data completion as described. As part of the creation of version 4.0 of the preliminary public use file, we performed one additional iteration of missing data completion. This was done for two reasons. First, our experience modeling variables over the past year led us to make many improvements that we wished to implement both in the data completion and data synthesis phases. Second, Yves Thibaudeau, from the Census Bureau Statistical Research Division (SRD), provided us with new 1996 SIPP data for home equity. These data were the result of an on-going research project at SRD, sponsored by SSA, to improve the imputation models for some of the variables collected in the wealth topical module in wave 3 of the 1996 panel. Our hope was that these improved starting data would lead to better models for our completion and synthesis of the wealth variables.

The SRMI method estimates the posterior distribution of the regression parameters (coefficients and variance of the error) and draws from this distribution to obtain parameters used to impute values. We impute multiple times, meaning we take multiple draws from the posterior distribution of the regression parameters. 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. For example, if total net worth was non-missing for 75% of the individuals in the sample, then 75% of the observations in each implicate file would have identical values for total net worth. The remaining 25% of the observations would have different values of total net worth across implicate files because of the multiple imputation. The implicates are generated by 4 separate SRMI processes, which is necessary because of the inter-related nature of the variables. Once a variable has been completed, its updated value is used as a right-hand-side variable in the imputation process for other variables. For example, once total net worth has been completed, its updated value will be used to impute a missing value for total income in 1990. Thus, in order to maintain internal consistency within an implicate file, each implicate must be generated separately. For version 4.0, we created four missing data implicates.

Because of the many iterations necessary to complete the data, the majority of the computing time spent creating a synthetic data set is actually spent dealing with missing values. Once the data are completed and contain no missing data except for structurally missing items, the final step of actually synthesizing all the data is takes much less time (*i.e.,* several weeks versus several months). Synthetic implicates are just like completed data implicates except that every individual has his or her values imputed, variable by variable, conditional on the completed data.[8] For example, in the case of total net worth, in the data completion phase 25% of individuals received imputed values to replace originally missing data. In the synthesizing phase, 100% of individuals received an imputed value to replace their original data, whether it was missing or not. Synthesizing data is in essence like doing one more iteration of missing data completion except everyone's data has to be completed.

The completed data from the appropriate 9th iteration implicate serve as the input for estimating the PPD used in the synthesizing phase. SRMI models are estimated using only originally non-missing dependent variables and completed explanatory variables. Explanatory variables thus contain either original non-missing data or imputed values from the 9th iteration.[9] We take a draw from the distribution of regression parameters and then impute a value based upon the most up-to-date synthetic data. This means that while the synthetic variables are not used in the model estimation, they are used to impute other synthetic values. For example, when estimating a model for total income in 1990, the values of total net worth used as explanatory variables would come from the 9th iteration completed data. However, when taking draws from the posterior predictive distribution for total income in 1990 in order to generate the synthetic total income 1990 variable, the synthetic value of total net worth would be used if this variable had been previously synthesized. Otherwise, the value from the 9th iteration of completed data was used.

Each one of the completed data implicates serves as the basis for creating four synthetic implicates. Since there are four completed data implicates, there are four separate input files to the synthesizing process. Each completed data implicate then has four distinct modeling steps and produces four separate draws for the regression parameters and four separate sets of synthesized values. This procedure preserves the internal consistency of each implicate file. In the end there are 16 synthetic implicates.

4.5 Modeling details

The actual implementation of either a Bayesian bootstrap iteration or an SRMI iteration is controlled by a SAS program that contains information about every variable and, based on this information, executes the appropriate modeling routines. The critical information that the analyst must provide for every variable is variable type, parent-child relationships, restrictions, level, and a set of grouping and conditioning variables to use in modeling. In this section we define these terms and explain how we assigned values in general. In the next section we list the specific values chosen for every variable.

4.5.1 Types of variables

The first information the analyst must provide about a variable to be completed and synthesized is the variable type. There are three major types of variables in the public use file: continuous, binary discrete, and categorical. The variable type determines which estimation routine will be used for the modeling step. We describe each in turn.

For continuous variables, the imputation model is a normal linear regression, which means that the posterior predictive distribution is based on the normal/inverted gamma posterior distribution for the parameters of a normal linear regression. Under an appropriate uninformative or conjugate prior, the posterior predictive distribution for the variable under study is normal (given the conditioning variables and the standard error of the equation). If the univariate

[8] Reiter (2004) distinguishes between the models used for the missing data imputation and those used for the synthesis, indicating that these models should not be the same if different conditions apply to the selection of values to be synthesized as compared to those that are missing. We fully implemented this distinction. Estimation and sampling from the posterior predictive distribution correctly reflects differences in the conditioning information. For example, to sample a synthetic birth date, we first estimated the PPD for birth date unconditional on range restictions. When we sampled from the birth data PPD, we imposed the range restrictions discussed below using accept/reject resampling from the unconditional PPD.

[9] This final step of model estimation in the synthesizing phase is in essence a repeat of the estimation done in the 9th iteration of missing data completion. This is because there is no updating of the explanatory variables. The explanatory variables always come from the completed data set that was generated in the 8th iteration of data completion. In fact if we had stored the parameter distribution results from the last round of data completion, we could skip this final model estimation step altogether and use the model results from the data completion phase. However, our programs are not set up to operate in this manner so this has been left for future research.

distribution of the variable we are trying to synthesize, y_k, differs greatly from conditional normality, the distribution of the synthetic values will differ from that of the confidential values. To handle this situation, 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.

The first step is to obtain an estimate of the unconditional distribution of y_k. Since the exact parametric distribution of y_k is unknown, we use a nonparametric estimator; namely, the kernel density estimate $\hat{K}$. For technical reasons, the kernel density estimator (KDE) is computed from a Bayesian bootstrap sample of y_k, not the exact Gold Standard copy of y_k. The KDE $\hat{K}$ is estimated separately for each set of grouping variable values. In addition, for each set of grouping variable values, the transformation is also estimated and applied to other continuous conditioning variables when appropriate (*e.g.*, if y_k is DER earnings this year and one of the conditioning variables is DER earnings next year, then both variables are transformed by an appropriate KDE estimate of each of their univariate distributions). Next we use the estimated KDEs to transform the actual dependent variable and any appropriate independent variables to normality. For each observation y_k, obtain the transformed value $y'_k = \Phi^{-1}\left(\hat{K}(y_k)\right)$, where Φ denotes the standard normal CDF. By construction, the y'_k have a standard normal distribution. Next, estimate the regression of y'_k on its (possibly transformed) predictor variables to get an estimate of the posterior predictive distribution of y'_k. Sample synthetic values $\tilde{y}'_k$ from this posterior predictive distribution. The synthetic values are normally distributed with conditional mean and variance defined by the regression model.[10] After standardizing the $\tilde{y}'_k$ to have zero mean and unit variance, compute the inverse transformation $\tilde{y}_k = \hat{K}^{-1}\left(\Phi\left(\tilde{y}'_k\right)\right)$. The imputed values $\tilde{y}_k$ are distributed according to $\hat{K}$, preserving the univariate distribution of the underlying confidential data. Further details of this procedure can be found in Woodcock and Benedetto (2006).

For binary discrete variables, the PPD is based on the asymptotic posterior distribution of the parameters of a logistic regression model. As described in section 4.3, we first split our sample of SIPP respondents into homogenous sub-groups using a set of grouping variables (sometimes called by-variables because they specify the subsets of observations that will be used for a particular model). Next, we estimated logistic regression models for each sub-group. We encountered problems with this approach when some sub-groups did not have enough variation to make the computation of a unique maximum likelihood estimate feasible. In other words, for some sub-groups, there were some combinations of right-hand-side variables that perfectly determined some value of the dependent variable. This problem, which is well known in the logistic regression literature see Albert and Anderson (1984), created a continuum of maximizers and prevented convergence in the algorithm used to maximize the likelihood function.[11] Because of this problem, known as quasi-separation, the results of the logistic regressions were sometimes unreliable and the coefficients had very large standard errors. The problem of partial ordering of the dependent variable in a logistic regression, which causes the log likelihood function not to have an interior maximum even though it is globally concave, is usually handled by respecifying the logistic regression. Failure to do so causes numerous problems with our synthesizer–in particular, the BIC-based automatic variable selection drops too many variables, if not all of them, and the draws from the posterior predictive distribution are extremely dispersed.

We believe that we used well-formulated logistical regression models and that none of the conditioning variables (sometimes called x-variables because they serve as right-hand side variables in the statistical models) had structurally determined relationships with the dependent variable. Hence, we believe that the quasi-complete separation problem was actually a sample size issue. Some of the sub-groups were simply too small. If we were to have large enough sub-group samples, every combination of x-variables and responses would eventually take on some positive probability for every sub-group. That is, we believe that the problem was sampling zeroes, not structural zeroes. Hence, we addressed this problem by using an informative prior on the logistic regression probabilities that is implemented using data augmentation; see Tanner (1996). The augmenting data matrix consists of one record for each potential combination of discrete conditioning variables and each discrete outcome. This imposes an informative Dirichlet prior on the space of outcomes of the logistic regression. The augmenting data provide the variation guaranteed to create a unique estimator for the posterior mode (equivalent to the maximum likelihood estimator in this case). However,

[10] This explanation is simplified. We take proper account of the inverted-gamma distribution on the standard error of the regression. Our procedure samples from the posterior distribution of the standard error of the regression, conditional on the sample error sum of squares and degrees of freedom. The sampled value of the regression equation standard error is used in the conditional normal posterior distribution of the regression coefficients.

[11] In the SAS logistic regression procedure, this error is reported as the warning for possible "QUASI-COMPLETE SEPARATION OF DATA POINTS."

the effects of the informative prior are dominated by the original data matrix in determining the parameter estimates except when one of the sampling zeroes occurs in a particular sub-group. Then, the prior distribution ensures a unique posterior mode.

For categorical variables, the PPD is based on the asymptotic posterior distribution of the parameters of a binary tree of logistic regression models that are used to model each level of the categorical variable successively as branches in the binary tree. The categorical variable modeling program looks for an equal split of individuals across categories, thus lumping some of the original categories together, and then models the probability that a person falls in either the first group or the second group. Then, within these two groups, another split is done and the probability that a person falls into one or the other of these subcategories is modeled. The binary tree continues until all the original categories have been modeled. Finally, the binary tree is synthesized and the synthetic values are used to recreate a synthetic value of the original categorical variable. For example, when the industry variable with four categories is modeled, the program might first split people into groups based on those with $ind\_4cat = \{1, 2\}$ and those with $ind\_4cat = \{3, 4\}$. It will then split the groups again in order to model $ind\_4cat = 3$ versus 4 and $ind\_4cat = 1$ versus 2. After the modeling is finished, a new synthesized ind_4cat variable is created that takes on values 1 to 4.

4.5.2 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 will 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. Instead, it will be set to SAS missing. An example of this type of relationship can be constructed from the variables $foreign\_born$ and $time\_arrive\_usa$. $Foreign\_born$ is the parent variable and takes a value of zero or one for everyone in the data set. It controls whether an individual is in scope to have a value for $time\_arrive\_usa$, the child variable. If a person was born outside the US, then that person should have a value for decade of arrival in the U.S. This value may be originally missing or not, but when $foreign\_born = 1$, the person is in scope to contribute data to the estimation of the model for $time\_arrive\_usa$ and will receive an imputed value for this variable that either replaces the missing data or synthesizes the original data. In this manner, we can prevent structurally missing data from skewing our modeling and we can also ensure that only the appropriate people receive a value for $time\_arrive\_usa$. In this example, the child variable is in-scope only when the parent variable takes a specific value ($foreign\_born = 1$). However, the method generalizes so the parent can take on a range of values. For instance, a person is in-scope to have a value for weeks worked part-time if weeks worked with pay is greater than or equal to one and less than or equal to five. In other words, as long as weeks worked with pay is positive, the person is in-scope to have a value for weeks worked part-time. If a person works a full month but never part-time, that person will have weeks worked with pay equal to four or five (depending on the month) and weeks worked part-time equal to zero. If a person does not work at all in a month, that person will have weeks worked with pay equal to zero and weeks worked part-time will be SAS missing.

Our second tool for handling relationships among variables is to place restrictions or constraints on some variables. Constraints do not restrict which observations are used in estimation nor do they restrict which observations receive an imputed or synthetic value. Instead, constraints specify a minimum and maximum value that restricts the range of draws from the posterior predictive distribution. For example, we synthesize birth date for every individual regardless of the value of any other variables. Thus, there is no parent variable for birth date. However the synthesized value for birth date must be consistent with the age requirements for any SSA benefits received by an individual. For example, if the individual began receiving retirement benefits in 1980, he or she must have been born by 1918 at the latest in order to be at least 62 years old by the time initial retirement benefit receipt. Thus, restrictions are imposed on birth date by the initial type of benefit and date of initial entitlement variables. Our programs impose these constraints by calculating what we term “utility variables” that contain these maximum and minimum values for every constrained variable. When we draw from the posterior predictive distribution for a constrained variable, the candidate sampled value is compared to the maximum and minimum for this individual and if the candidate draw falls outside the specified range, another draw is taken. This comparison and re-sampling is repeated until the candidate sampled value satisfies the constraints or 100 candidate draws have been performed–at which point the value is set equal to the closest boundary (*i.e.,* if the value is over the maximum on the 100th candidate draw, it is set equal to the maximum).

4.5.3 Levels of the parent/child tree

The implementation of the parent-child relationships and the imposition of exact restrictions are accomplished by assigning every variable a level in the binary tree representing the graph of the parent-child relations. Hence, this information must be provided by the analyst for every variable. The level governs the order in which the sequential regression imputation is done. If the variable does not depend (for any reason) on another variable being modeled first, then it is at the first level, the root of the graph representing the binary tree. Otherwise, a variable must be one level higher than the highest level of any variables on which it depends, so that estimation occurs when the algorithm reaches a node with a binary decision or a leaf of the tree (nodes which are not parents of any variable) where the child variable is not structurally missing. The dependence modeled in the binary tree can be either in the form of a parent-child relationship or constraints. The variable list is then sorted by level (ascending) and missingness (descending) so that all first level variables are imputed or synthesized in a given iteration prior to second level variables, *etc.* In most cases, any variable with either a parent or restrictions of some type will be either a level two variable or higher.

There are a few exceptions. If a parent variable or a variable imposing a restriction is never missing and will not be synthesized, then its child variable or constrained variable can still be at level one for purposes of the estimation.

At the outset of each iteration, the values of all parent variables are stored in a separate file, $orig\_parents$. Since a parent variable must be at a lower level in the tree than its associated child variables, in any given iteration, it will be imputed or synthesized before its children. Once a parent variable has been imputed or synthesizd, the current iteration file contains the most up-to-date parent values. The previous iteration's values of the parent variable are still in $orig\_parents$. However, at this point in the iteration cycle (after a parent has been imputed but before its children have been imputed), the previous iteration's parent values are the ones that correspond to the most up-to-date child variable values. Hence, when the programs reach the point at which they must estimate current iteration models for child variables, they use only observations where the value of the parent variable in $orig\_parents$ falls in the aforementioned range for the estimation.

At each level and for every variable, fresh model estimation is used to form the posterior predictive distribution. However, when actually imputing values (sampling from the PPD), the programs use the most-up-to-date parent variables to select the observations that will receive values for the children variables. Thus, when the iteration is finished, the parent and children variables all agree again. Child variables only take on values when their parent variables are in the appropriate range and all other observations are set to SAS missing to denote structural missingness.

4.5.4 Grouping and conditioning variables

Finally, as described in sections 4.2 and 4.3, the analyst chooses both grouping variables and conditioning variables. Grouping variables are chosen so that each group meets a minimum size requirement and at the same time contains people who are as similar as possible. In SRMI models, adding additional grouping variables is very costly in terms of computational time so the analyst must seek to make a parsimonious but effective list of variables 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 esitmate 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, the programs 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 along specified dimensions and then split into sub-groups again, using a list of grouping variables that is shorter and produces fewer groups. Hence, the analyst actually specifies multiple lists of grouping variables and conditioning variables for each model. Each set of grouping variables is defined by progressively fewer variables as variables are dropped in order to create sub-groups of larger sizes. As variables are dropped from the grouping variables list, they are added to the list of conditioning variables. Hence, each list of conditioning variables becomes progressively longer. For example, the analyst might originally use $black$, $male$, and age_cat_expand, an 11 category age variable, as grouping variables. This would produce 44 groups (2 categories for $black$, 2 categories for $male$, and 11 categories for age). The program would form these 44 sub-groups and check the sample size of each group against the minimum of 15 times the number of conditioning variables. If the analyst included 7 conditioning variables, each sub-group would need at least $\max(1000, 105) = 1000$ observations. If the analyst included 100 conditioning variables, then each sub-group would require at least $\max(1000, 1500) = 1500$ observations. Any sub-group that was large enough would be sent directly to the modeling step using the specified conditioning variables. All groups that were too small would be combined and then split again using a the next set of grouping variables specified by the analyst. In this case the analyst might use only $black$ and $male$ as grouping variables and then include age_cat_expand in the list of conditioning variables that corresponds to this second 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 and sent to the regression modeling step.

As with grouping variables, the initial selection of conditioning variables is dependent on 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). The analyst controls the criteria for determining the critical BIC (posterior odds ratio for the model including the variable versus the model excluding the variable) and can make the selection criterion stronger or weaker, depending on the need to keep fewer or more conditioning variables. In version 4.0, we have considerably weakened the critical BIC in order to ensure that important conditioning variables were not dropped from the right-hand side of models.

4.5.5 Specific variable details

We have created an Excel workbook with spreadsheets that give the details of the synthetic data creation procedure for every variable on the public use file.[12] The workbook is attached to this report and should be useful to analysts who need information about the methods used for any particular variable in the data completion and synthesis phases. We give the source of each variable (SIPP, IRS/SSA, SSA), whether it contained missing data, whether it was synthesized, what type of model was used to complete missing data, what type of model was used to create synthetic data, and the range of values. We list variables that serve as either parents or children and we specify restrictions, if any, imposed by other variables. We describe any post-processing requirements for the variable, including whether any additional variables need to be created for the final file. Finally, we provide a link to the set of grouping and conditioning variables used in the modeling. We also provide this information in Appendix Table A1 which is included with this report as a separate PDF file. In this section of the report, we describe groups of variables and the modeling techniques used for the group in both the completion and synthesis phases.

Unsynthesized variables Early discussions among committee members produced a list of variables that would not be synthesized: gender, race (black/African-American), three categories of education, marital status, three categories of age, and a link to the record of the spouse at the time of interview. The idea behind unsynthesized variables was that these would enhance the analytic validity of the synthetic file by preserving some basic individual characteristics. Unsynthesized variables, however, also provided a very effective matching strategy for anyone trying to link the new synthetic public use file to the original SIPP public use files. If the unsynthesized variables are used to stratify the sample and if some combinations produce very small groups of people in the Gold Standard file, then an intruder attempting to link synthetic data records to already public SIPP files could match these small groups and might be able to re-identify some individuals in the original SIPP public use files. Thus, this original list of unsynthesized variables was chosen to minimize the number of cells in the Gold Standard file with fewer than 10 people when cross-classified by all the unsynthesized variables.

During this final year of the project, the Census Bureau and SSA conducted lengthy discussions about the possibility of including unsynthesized SSA benefit variables on the file. Although these variables were administrative and hence did not have direct equivalents in the original SIPP survey files, the Census Bureau was concerned that adding more unsynthesized variables to the file would create even more small cells that would allow a user to link across

[12] See varlists_description_version_4_0.xls in the appendix to this report.

synthetic implicates. If the synthetic implicates were linked, they could be averaged and something resembling the original record could possibly be re-created. The Census Bureau felt that this possibility presented too much disclosure risk and preferred to keep the number of unsynthesized variables small enough to avoid large numbers of cells with fewer than 10 people.

Discussions between the two agencies produced the following compromise. Gender, marital status, and the spouse-link would remain unsynthesized. In addition, we would add two important SSA benefit variables to the unsynthesized list: type of benefit at time of initial benefit receipt and type of benefit in April 2000. These two categorical variables quantify fact of receipt as well as the reason and are hence the most fundamental of all the SSA benefit variables. Thus, the list of unsynthesized variables in the final version of the synthetic public use files is gender, marital status, initial type of benefit, type of benefit in 2000, spouse initial type of benefit and spouse type of benefit in 2000 (both created using the unsynthesized spouse link), and the spouse identifier variable.[13] The resulting configuration of unsynthesized variables creates no small cells using only the variables originating from Gold Standard SIPP variables. Furthermore, there are only approximately 130 cells with fewer than 10 individuals when stratifying using the full list, which includes the two SSA-provided type of benefit variables that are not present on any current SIPP public use file. See Table 1 for a full break down of small cells created by various configurations of unsynthesized variables.

The existence of unsynthesized variables requied the imposition of some constraints on other variables. In particular, receipt of certain types of benefits imposed constraints on an individual's age at a given point in time and marital status at the time of the survey imposed constraints on the marital history of an individual. We describe how we handled these restrictions in sections 4.5.5 and 4.5.5, respectively.

Birth date, death date, and dates of benefit receipt One of the most important variables in the file, from the perspective of both disclosure risk and usefulness in analyses, is $birthdate$.[14] It was essential that this variable be adequately protected yet synthesized well enough to reproduce appropriate age distributions for many sub-groups. We used the administrative value of the date of birth (from SSA administrative records) whenever we could. The administrative $birthdate\_pcf$ was missing in cases where the individual did not have a validated SSN and was completed using the couple-level Bayesian bootstrap described in 4.4. We modeled the variable in the data synthesis phase as a continuos variable with restrictions. If a person received benefits in April 1, 2000 ($tob\_2000 = \{1, 2, 3, 5, 100\}$), we forced the synthetic $birthdate$ to be such that the individual would be appropriately old enough for the benefit received. Individuals with retirement benefits had to be at least 62 by April 1, 2000, individuals with aged spouse benefits ($tob\_2000 = 3$) had to be at least 62, individuals with aged widow benefits ($tob\_2000 = 5$) had to be at least 60, and individuals receiving disability benefits ($tob\_2000 = 2$) could not be 65 years old or older. In addition, we restricted draws for the synthetic $birthdate$ such that it was no more than a year in either direction from the original administrative birth date ($birthdate\_pcf$). So that

$$birthdate\_pcf - 365 \leq birthdate \leq birthdate\_pcf + 365$$

where we note that date variables are measured in days.

Because tob_2000 and $tob\_initial$ are unsynthesized, further consistency restrictions were imposed on $birthdate$. If an individual's initial benefit types were retired or retired spouse ($tob\_initial = \{1, 3\}$) and unsynthesized $date\_initial\_entitle$ is before April 1, 2000, then $birthdate$ must be consistent with age at April 1, 2000 greater than 62. The reason for this restriction is that when $date\_initial\_entitle$ is synthesized, there will be support for a synthetic value that is consistent with these types of benefits starting before April 1, 2000. If an individual's initial benefit types were retired or retired spouse ($tob\_initial = \{1, 3\}$) and unsynthesized $date\_initial\_entitle$ is on or after April 1, 2000, then $birthdate$ must be consistent with the individual turning 62 (and thus being eligible for these types of benefits) before December

[13] We did make one change with respect to the gender variable that was necessitated by disclosure risk. The Gold Standard contained 5 married couples that had the same gender. Due to the unusual nature of these cases, we could not leave gender and marital status unchanged for these couples without ensuring a link between the synthetic data and the public use SIPP. Hence for these 5 couples, we randomly changed the gender of one of the spouses. We did so in a manner that allowed the weighted counts of males and females in the synthetic data files to remain close to what they were before the gender swaps.

[14] In the synthetic data files there is only one birth date variable: $birthdate$. In the Gold Standard file, there are two birth date variables: $birthdate\_pcf$, the administrative birth date, and $birthdate\_sipp$, the SIPP birth date. The SIPP birth date is only used during the disclosure avoidance analysis.

31, 2002.[15] The same process is repeated for aged widow benefits ($tob\_initial = 5$) using an age cut off of 60. Finally, for disabled benefits ($tob\_initial = 2$) we reverse this procedure to keep $birthdate$ consistent with being less than 65 years old when this type of benefit is collected. If $tob\_initial = 2$ and unsynthesized $date\_initial\_entitle$ is on or after April 1, 2000, then the minimum synthetic $birthdate$ is May 1, 1935 so that there is support for a synthetic $date\_initial\_entitle$ on or after April 1, 2000 and the individual would be age-eligible for disability benefits at that time.

The completed data, which are based on the Gold Standard file and which contain either the matching administrative data for the individual and his/her spouse or a complete administrative record (all dates, all earnings, and all benefit variables drawn from the same individual's administrative records and his/her spouse), exhibit some dating inconsistencies that are not due to either the missing data imputation or the synthesis. Because age eligibility restrictions have been imposed in the synthetic data, the synthetic data are cleaner than the completed data; that is, they do not display as many age-related eligibility anomalies as can be seen in the completed data.

The variable $deathdate$ was completed in a similar manner as $birthdate$, using the donor chosen in the couple-level Bayesian bootstrap. This variable was also modeled as a continuous variable during the synthesis phase; however, we also synthesized whether or not the individual died ($flag\_deathdate\_exist$). The construction of $flag\_deathdate\_exist$ used the existence of a date of death in the PCF as the indicator of death without modification.. The synthetic $flag\_deathdate\_exist$ is the parent to $deathdate$. $Deathdate$ was restricted such that the earliest possible year of death was 1990.

The following constraints on $deathdate$ obviously only pertain to the cases where the synthetic death indicator is in scope ($flag\_deathdate\_exist = 1$). In the cases where the completed $flag\_deathdate\_exist = 1$, we constrained the draw of synthetic $deathdate$ to be within 365 days of the completed $deathdate$. If benefits were received in the month of April 2000 ($tob\_2000 > 0$), then the minimum value of the synthesized $deathdate$ is April 1, 2000, since we do not want anyone receiving benefits after death. If there is no benefit amount reported for the entire month of April 2000 ($tob\_2000 =$ SAS missing), the initial benefit type is present ($tob\_initial > 0$), and the unsynthesized initial entitlement date is before April 1, 2000 ($date\_initial\_entitle <$ April 1, 2000), then $deathdate$ can be no later than March 31, 2000. Thus, if an individual dies and stops receiving benefits between the initial entitlement date and April 1, 2000, we ensure that the explanation for this loss of benefit is the date of death. If there is no benefit amount reported for the entire month of April 2000 ($tob\_2000 =$ SAS missing), the initial benefit type is present ($tob\_initial > 0$), and the unsynthesized initial entitlement date is on or after May 1, 2000 ($date\_initial\_entitle >=$ May 1, 2000), then the minimum value for $deathdate$ is May 1, 2000. We do this to create support for a draw of the synthetic date of initial entitlement that is consistent with receiving no benefits in the month of April 2000 and the synthetic date of death.

The final date variable that we completed and synthesized was year of initial entitlement to SSA benefits ($date\_initial\_entitle$). Both completion and synthesis were done in the same manner as the $birthdate$ and $deathdate$ variables. The restrictions on the initial entitlement variable were derived from the draws for the synthetic $birthdate$ and $deathdate$ as well as from the type of benefit variables. If initial type of benefit was retired worker ($tob\_initial = 1$), then year of initial entitlement had to be at least 62 years (actually 62×365.25 days) from the synthetic $birthdate$ value. For other types of initial benefits we imposed the following restrictions: at least 62 years old for aged spouses ($tob\_initial = 3$), at least 60 years old for aged widows ($tob\_initial = 5$), and less than 65 years old for disabled workers ($tob\_initial = 2$)). Date of initial entitlement had to be before $deathdate$ and before April 1, 2000 if type of benefit 2000 indicated benefit receipt at this point in time. If no benefits were received in April 2000, then date of initial entitlement had to be after April 2000. Hence, date of initial entitlement did not cross the April 2000 boundary. We made two additional restrictions. Because the MBR file did not provide benefit amounts prior to 1962, we did not allow date of initial entitlement to cross the January 1962 boundary. This allowed us to leave the monthly benefit amount variable missing for those with a synthetic (and original) date of initial entitlement prior to January 1962. Finally we restricted draws for $date\_initial\_entitle$ such that the synthetic value was forced to be no more than 2 years in either direction from the original value.

Administrative earnings After completing missing SER and DER data using the couple-level Bayesian bootstrap, the administrative earnings variables were synthesized in two parts. We first modeled whether the SIPP individual had positive earnings in a given year and then only modeled actual earnings for those with a positive earnings indicator.

[15] In Version 4.0 of the Gold Standard and SIPP/SSA/IRS-PUF the SSA MBR data end with calendar year 2002, even though the earnings data end with calendar year 2003. This separation is due to the schedule of extract updates maintained between the Census Bureau and SSA.

Thus, the earnings indicator was the parent variable and the actual earnings variable was the child. We synthesized the earnings indicators using a Bayesian bootstrap, done one year at a time. We used leads and lags for previous and future years as grouping variables as well as demographic variables and summary earnings measures. We began with SER earnings (capped at the FICA maximum) in 1951. Using the bootstrap, we created a synthetic value for every individual for the variable $ser\_posearn\_1951$. For those with ser_posearn_1951=1, we then used a bootstrap to create a synthetic value for whether each individual had reached the FICA taxable maximum in 1951 ($ser\_maxearn\_1951$). For those with $ser\_maxearn\_1951 = 1$, we automatically set $totearn\_ser\_1951$ equal to the maximum. For those with $ser\_maxearn\_1951 = 0$, we modeled earnings using our continuous variable techniques, including the two-sided KDE transform. After 1951 was completed, we moved to 1952 and repeated the process. When creating grouping variables for 1952, we used the new synthetic values for 1951 and the completed data for 1953 and after. We moved through the entire array in this manner until the year 1978.

The DER array of earnings begins in 1978. Beginning with this year, we synthesized total earnings. We used a similar process to the one used for the SER earnings except that we synthesized four separate time series: non-deferred total earnings at FICA covered jobs ($nondefer\_der\_fica\_\{year\}$), deferred total earnings at FICA covered jobs ($defer\_der\_fica\_\{year\}$), non-deferred total earnings ($nondefer\_der\_nonfica\_\{year\}$) at non-FICA covered jobs, and deferred total earnings at non-FICA covered jobs ($defer\_der\_nonfica\_\{year\}$). After each year of DER earnings was synthesized, we calculated SER earnings as the lesser of total non-deferred and deferred earnings at FICA covered jobs or the FICA taxable maximum:

$$\begin{aligned} totearn\_ser\_\{year\} &= \min(taxmax, nondefer\_der\_fica\_\{year\} + \\ &\quad defer\_der\_fica\_\{year\}) \end{aligned}$$

This process was continued until 2003, the last year of available earnings data.

One final constraint was imposed on the SER and DER earnings arrays. Earnings could only be positive in years where the individual was at least 15 years old and in years up to and including date of death.

Social Security benefits We synthesized two SSA benefit variables: monthly benefit amount for the month of initial entitlement and monthly benefit amount for April 2000. Each of these variables was the child of the corresponding type of benefit variables. Only individuals with a positive initial type of benefit received a synthesized value for the initial MBA ($mba\_initial$) and likewise for mba_2000. However since neither type of benefit variable was synthesized, the set of people with positive $mba\_initial$ and mba_2000 values was the same in the completed and synthetic data. Both MBA variables were synthesized using continuous variable methods and were restricted such that synthetic values had to be no more than \$50 less than or greater than the original values:

$$mba\_initial(\text{completed}) - \$50 \leq mba\_initial(\text{synthetic}) \leq mba\_initial(\text{completed}) + \$50.$$

and similarly for mba_2000.

Once the synthetic data files had been created, we created two additional variables that were direct derivatives of SER earnings: Average Indexed Monthly Earnings ($AIME$) or Average Monthly Wage (AMW) and Primary Insurance Amount (PIA). The $AIME$/AMW calculation is the method used to summarize a person's lifetime earnings in order to make OASDI benefit calculations. The $AIME$/AMW is used to calculate the PIA, which in theory tells what benefit a person receives. However, additional rules about spouses, children, family maximums, *etc.,* mean that the actual monthly benefit amount often differs from the PIA. The precise calculations for the $AIME$/AMW and the PIA depend on a person's gender, date of birth, type of benefit sought, and year of application. The rules governing these calculations are quite complicated (partly because they change a great deal over time) and depend on many things not necessarily observable in our data set. The PIA is an actual variable on the SSA Master Beneficiary File (MBR), but the decision was made by SSA and the Census Bureau not to synthesize this variable or include it on the file, primarily because of concerns that it would be inconsistent with the synthetic SER earnings array. Instead, it was decided that the $AIME$/AMW and the PIA would be calculated directly from the synthetic earnings using a simplified set of rules.

For individuals who reached age 62 before 1979, we calculated the AMW and for those who reached age 62 after 1979, we calculated the $AIME$. To compute the AMW, we first calculated the number of years between age 21 (or 1951 if later) and age 62, subtracted five years, and multiplied by 12 to get the number of months at risk. We then summed earnings between age 21 and age 62, dropping the five lowest years. Total summed earnings were then divided by the number of months at risk to give the Average Monthly Wage. There was one exception. For men (but not women) born before 1911, the calculation was performed using the years between age 21 and age 65 because the retirement age for men was three years older prior to 1973. The $AIME$ calculation was essentially the same as the AMW but earnings were indexed to the year in which the individual turned 60.

Once the $AIME$/AMW had been calculated, the PIA was determined by applying the cut-off points and percentages applicable for the year of initial entitlement to benefits. In a given year, $a\%$ of the first X dollars of the $AIME$ formed the initial portion of the PIA. The $b\%$ of the next Y dollars formed the next portion and $c\%$ of the next Z dollars formed the final portion. The sum of these three portions was the PIA. Prior to 1979, the cut-off points stayed constant across years and the percentages changed. Post 1979, the cut-offs changed every year while the percentages stayed constant. We used tables 2.A8, 2.A10, 2.A11, and 2.A16 from the SSA Statistical Supplement 2005 to make these calculations and consulted with Barbara Lingg at SSA to clarify details.

It is important to note that we calculated the $AIME$/AMW and the PIA for individuals based on the assumption that they were applying for retired worker benefits. We did not make separate calculations for individuals who received disability, spouse, or death benefits. Thus the $AIME$/AMW and PIA on the file will not correspond to the MBA for types of benefits other than retired worker. However, since the $AIME$/AMW and PIA do not contain any additional information and are direct calculations based on other variables in the file, any researcher interested in performing a different calculation may do so. We include these two variables solely for the convienence of retirement researchers.

SIPP time series arrays The synthetic data includes 13 time series of SIPP variables: weeks with pay, weeks part-time, total annual hours, family poverty cut-off, family total income, personal total income, personal total earnings, family welfare participation, family welfare income, private health program participation, private health program income, general health insurance coverage, employer-provided health insurance coverage. In addition, weeks with pay, weeks part-time, annual hours, family income, personal income, and personal earnings have corresponding arrays of indicator variables that serve as parent variables and tell whether the continuous variable takes on a value or not. We use a Bayesian bootstrap to complete and synthesize all the indicator arrays. We then use continuous variable methods to complete and synthesize the remaining variables with the indicators serving as parent variables.

Wealth variables In modeling the wealth variables (total networth, own home indicator, home equity, and non-housing wealth), we create a set of flags to indicate whether the three continuous variables are non-zero. We then use a Bayesian bootstrap to complete and synthesize these three flags together with the home ownership indicator. These four variables are bootstrapped as a group to ensure consistency. We then use the three flags as parents of the three continuous variables. Using our continuous variable techniques, individuals are modeled to have a value of each of the three wealth variables only if the the appropriate flag indicates a non-zero value.

Marital history variables The challenge in synthesizing the marital history variables was to ensure that the historical variables were consistent with the reported marital status and with each other. To accomplish this, we used a Bayesian bootstrap to both complete and synthesize marital history variables. We first bootstrapped a group of variables that summarized the history ($mh\_category$, number of marriages, number of divorces, and married at end of history) using marital status as one of the grouping variables. This guaranteed that individuals would receive donated values of $mh\_category$ and the three other summary variables only from other individuals with the same marital status so no inconsistencies would arise. We then used an additional Bayesian bootstrap for $flag\_mar4t$ with $mh\_category$ as one of the grouping variables. Once these variables had been modeled, we created a set of indicator flags that indicated whether the individual should have an age at time of first marriage, duration of first marriage, duration of end of first marriage, duration of second marriage, duration of end of second marriage, duration of third marriage, duration of end of third marriage, and duration of fourth marriage based on the events that occurred in his or her history. Individuals with at least one marriage in their history were modeled to have an age at time of first marriage. Individuals whose first marriage had ended were modeled to have a duration of first marriage and duration of first marriage end. Individuals with at least two marriages were modeled to have a duration of end of first marriage (*i.e.*, time between first and second marriages) and duration of second marriage and so on until the fourth marriage. The age and duration variables were modeled using our continuous variable techniques and were children of the indicator flags.

After the synthesizing was finished, we post-processed these data to create the $mh1$-$mh7$ flags that report the same information as $mh\_category$. We used the age at time of first marriage and the duration variables to create the ages at time of each marital history event. To accomplish this, we first summed all the synthetic duration variables to create a total duration and calculated what percentage of the total duration was accounted for by each particular spell. For example, if the individual had 2 marriages, with the second marriage on-going, we calculated what percentage of the total duration was made up of the first marriage duration, time between first and second marriage, and second marriage duration. We took the time period between age at time of first marriage and 2003 (end of our administrative data) and divided it into marital event intervals using the percentages. To continue our example, if age at time of first marriage was 25 and (based on $birthdate$) occurred in 1983, then the total time period was 20 years which would need to be divided between duration of first marriage, interval between first and second marriages, and duration of second marriage. If according to the modeled durations, the first marriage accounted for 50% of the time, the interval between marriages accounted for 25% of the time, and the second marriage accounted for 25% of the time, then age at time of first marriage ending would be 35 (1993) and age at time of second marriage would be 40 (1998).

5 Weight Creation and Synthesis

5.1 Introduction and background

The creation of a unique new public use file that combines SSA/IRS administrative data with extracts from five separate SIPP panels required many special efforts to insure that the final product would be analytically valid. One concern that arose early in the process was how to provide researchers with proper weights for a file that pooled survey respondents from five separate samples. There are design instructions that explain how to combine the official SIPP weights when using panels that contain overlapping years in order to produce estimates that are representative of a known universe at a specific date, but the existing SIPP public use files do not contain the information needed to create a weight that is appropriate for pooling all of the panels into a single analysis. When longitudinal administrative data are linked to these SIPP panels, every observation potentially contributes data to any time period; therefore, the problem of constructing an appropriate weight was integral to permitting these data to be used to make national estimates. In addition, because the different SIPP panels over-sample low income individuals and other targeted demographic groups at different rates, the pooled survey data can only be used to make estimates about the U.S. population if an appropriate weight is used in analyses. Thus, one of the stated objectives of the SIPP/SSA/IRS-PUF project was to create a weight for the five merged SIPP panels where each SIPP person's weight indicated how many persons in the reference population that individual represented. The designated reference population is all individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000.

In order to determine how many people in the reference population each SIPP person represented, we used the 1996 SIPP sampling plan as our guide and divided the Decennial reference population into the same strata (*i.e.*, groups) from which SIPP individuals were originally sampled. We then located each SIPP individual in the Decennial reference population. Once we knew how many SIPP people were in each stratum, the preliminary weight calculation was straight forward: each SIPP person's weight equals the number of Decennial persons in that particular stratum divided by the number of SIPP persons in the same stratum. For example, if the tenth stratum contained 100 Decennial persons and two SIPP sample individuals, then each SIPP person in the tenth stratum received a preliminary weight of 50=100/2. The final weight was calculated by raking the preliminary weight to match official U.S. civilian non-institutional population estimates as of April 1, 2000 based on the same control total categories used for the 1996 SIPP weights in the current public use files. The validity of the final weight was tested by computing univariate statistics for key SIPP and SSA variables and comparing them to independently derived estimates from other sources. The results of this testing are reported in Table 2.

In order to locate SIPP individuals in the Decennial reference population, we linked the two data files using the PIK, a unique Census person identifier that replaces the SSN, and which has been added via probabilistic record linking to the Census 2000 micro-data files. For about two-thirds of the individuals in the Gold Standard SIPP file, the PIK link was successful. For the remaining one-third of SIPP individuals, it was not possible to locate an exact match in the Decennial reference population. This occurred either because these SIPP individuals did not provide an SSN to the SIPP survey (and therefore had no PIK) or their PIK did not successfully match to an individual in the Census 2000 micro-data. Of the 263,793 individuals in version 4.0 of the Gold Standard file, 177,165 matched exactly to a Census 2000 reference person by PIK. The other 88,628 SIPP individuals were matched to a Census 2000 reference person using probabilistic record linking.

The strata from the SIPP sampling plan had several levels. The first stratification level (or grouping level) was Primary Sampling Units (PSUs), which were created by grouping geographic counties together. The SIPP Survey Design Branch (SIPPSDB) in the Demographic Statistical Methods Division (DSMD) provided us with a file that assigned geographic counties to PSUs. Large counties were assigned a unique PSU while smaller counties were grouped together to form a single PSU. The second stratification level was by stage-1-clusters, which were simply created by grouping PSUs together. Some PSUs were self-representing, meaning that they were the only PSU in their stage-1-cluster and were sampled with certainty. Other PSUs were non-self-representing, meaning they were grouped with other PSUs and were sampled with probability less than one. The SIPP Survey Design Branch provided us with a file that assigned the 1,928 PSUs to 217 stage-1 clusters. These stage-1 clusters were then used to select PSUs from which individuals would be sampled. Once PSUs were selected, individuals in high poverty strata were over-sampled in each selected PSU. Therefore, our final stratification level was defined by whether an individual was in the high poverty stratum or the low poverty stratum according to the definitions of high and low poverty in the SIPP Sampling Plan. The final stratification which combined the location of an individual in a stage-1-cluster and a poverty stratum was called a stage-2-cluster. The number of SIPP and Decennial persons in each stage-2 cluster was used to calculate the preliminary weight according to the above formula. Raking was then applied directly to the preliminary weight to create the final weight. Finally, a synthetic version of the weight was created for each of the synthetic implicates.

The rest of this subsection provides the details of this weight creation process. We begin by giving a summary of each of the seven main steps in the process. This summary is meant to give the reader a general idea of how the weight was created before we present the details. Following the summary, parts A-G give careful descriptions of exactly how each step was performed.

5.2 Summary of the weight creation process

Our method for creating an ex-post weight for the merged SIPP panels involved seven steps. Parts A and B describe the method of creating the Census 2000 reference population and dividing it into strata according to the 1996 SIPP Sampling Plan. Parts C and D do the same for the SIPP, describing the method by which the SIPP was divided into strata according to the 1996 SIPP Sampling Plan. Part E describes the method by which each SIPP person was located in the Decennial reference population. Part F describes the creation of the preliminary weight according to the formula mentioned above. Part G describes the creation of the final weight by raking (*i.e.,* adjusting) the preliminary weights to agree with official U.S. population control totals for the sex/age/race/ethnicity demographic breakdown of the reference population, as supplied by the Census Bureau's Population Estimates Division. The next two subsections (5.11 and 5.12) describe some geography and birth date issues that arose during the weight creation process. The next subsection (5.13) discusses the overall evaluation of the Gold Standard weight, and the final two subsections (5.14 and 5.15) describe the creation of the synthetic weight and discuss the results of the analytical validity testing of this weight.

5.2.1 Part A: Creation of poverty stratification variable for Census 2000 records

Part A describes the creation of a poverty stratification variable for Census 2000 records according to original 1996 SIPP stratification rules. Households were assigned to a poverty stratum based on either household income or household composition. For long form households (Sample Census Edited File, SCEF), an income variable was available and households/records were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold. For long form respondents for whom income data was not available and for short form respondents (Hundred percent Census Edited File, HCEF), household composition was used to proxy poverty status. A household was assigned to the high poverty stratum if it had any of six characteristics such as a black householder under age 18 or over age 64 (see 5.3 below for the full list of characteristics).

5.2.2 Part B: Creation of stage-2 clusters for Census 2000 records

Part B describes the methods by which counties were assigned to PSUs, PSUs were assigned to stage-1 clusters, and stage-2 clusters were created for the Census 2000 records. This section also describes the manner in which the Decennial reference population was created by only including decennial records that were in the civilian, non-institutionalized U.S. population ages 18 and older on April 1, 2000.

5.2.3 Part C: Creation of poverty stratification variable for SIPP records

Part C is analogous to Part A for the SIPP. It describes the creation of a poverty stratification variable for SIPP records according to the original SIPP stratification rules. Households were assigned to a poverty stratum in the same manner as they were for the Decennial records.

5.2.4 Part D: Creation of stage-2 clusters for SIPP records

Part D is analogous to Part B for the SIPP. It describes the methods by which counties were assigned to PSUs, PSUs were assigned to stage-1 clusters, and stage-2 clusters were created for the SIPP records.

5.2.5 Part E: Matching SIPP individuals to the Census 2000 records

Part E describes the methods by which SIPP persons were located in the Census 2000 reference population. There were 263,793 individuals in the SIPP Gold Standard file, 177,165 of which were matched exactly by PIK to a Decennial record. The remaining 86,628 SIPP records were matched by a probabilistic record linking method to an in-scope Census 2000 record (*i.e.*, a record determined to be in the reference population) in the following manner. Each SIPP person was first assigned a set of Decennial candidate records (candidates for a match) that agreed exactly with that SIPP record's values for each variable in a set of blocking variables. Then, one of the Decennial candidates was chosen as a match for the SIPP record based on how closely that Decennial record's values agreed with that SIPP record's values for each variable in a set of matching variables. There were two blocking passes through the data. The first blocking pass used 6 blocking variables and 7 matching variables (see 5.8 below for the complete list). Any SIPP record that had 30 or fewer Decennial candidates was considered unmatched and sent through the second blocking pass, which used 3 blocking variables and 10 matching variables.

5.2.6 Part F: Creation of a preliminary weight

Part F describes the calculation of the preliminary weight using Census 2000 stage-2 cluster counts and SIPP stage-2 cluster counts, and the formula above: preliminary weight equals the number of records in Decennial stage-2 cluster divided by the number of records in SIPP stage-2 cluster. This preliminary weight was the same for all SIPP records in a particular stage-2 cluster.

5.2.7 Part G: Creation of final weight

Part G describes the creation of the final weight by raking (*i.e.*, adjusting) the preliminary weights to agree with population control totals for the demographic breakdown of the reference population as provided by the Population Estimates Division. The reference date for the population control totals was April 1, 2000. The list of groups to which the weights were controlled (*e.g.*., black males ages 19-24, black males ages 25-29, *etc.*) was provided by SIPP Survey Design Branch and was the same as the list of population subgroup totals used for raking the original 1996 SIPP weights.

5.3 Part A: Creation of poverty stratification variable for Census 2000 records

We first created the variables that were needed to define poverty status of individuals and households in the SIPP unit frame. The SIPP had four other sampling frames in addition to the unit frame: Area, New Construction, Group Quarters, and Coverage Improvement. However, in the 1996 SIPP panel approximately 80% of records came from the unit frame. Therefore, due to the extraordinary amount of work involved in identifying the stratification rules for the other four sampling frames, we only created the poverty stratification variable according to the unit frame and assumed everyone came from the unit frame. Construction of the necessary poverty-defining variables was different depending on whether the individual completed the Census 2000 short or long form.

5.3.1 Data sources for short-form respondents

For individuals completing the short-form, we took relevant geographic and demographic information from two HCEF data files, namely a person-level file and a block-level file. From the person-level file we obtained indicators for householder, child of householder, spouse of householder, gender, black, Hispanic, age groups ($<$18, 18-64, $>$64; and $<$18, 18-62, $>$62), birth date, and geography (state, county, approximate tabulation geography). From the block-level file we obtained county, state, population count, housing count, and place code by $geocodecoll$ (unique collection block identifier). We then used the person-level data to create a housing-unit file that contained an indicator for family-type housing versus group quarters, a count of persons living in family-type housing, number of children under age 18, householder information (female, black, Hispanic, age: $<$18, 18-64, $>$64), and an indicator variable for households with a female householder and no spouse present. Also, in cases where no person was assigned to be the householder (*e.g.*, group quarters have no householder in the Decennial), we assigned the oldest person to be the householder. This file was then merged to each person's record.

5.3.2 Data sources for long-form respondents

Information about long-form individuals came from three SCEF data files: block-level, housing-unit level, and person-level files. From the person-level file we obtained the same demographic and geographic variables as from the short-form: indicators for householder, child of householder, spouse of householder, gender, black, Hispanic, age groups ($<$18, 18-64, $>$64; and $<$18, 18-62, $>$62), birth date, state, county, and approximate tabulation geography. In addition, we obtained information on education (college, some high school) and income (total annual personal income, 1999). From the block-level file we also obtained the same variables as from the short-form: county, state, population count, housing count, place code by $geocodefull$ (unique tabulation block identifier) as well as housing counts and population counts. Finally, from the housing-unit file we obtained an indicator for family-type housing versus group quarters, count of persons living in family-type housing, number of children under age 18, and an indicator for monthly rent below $300. We also used the person-level data to create some additional housing-unit information, in particular an indicator for family-type housing versus group quarters, count of persons living in family-type housing, number of children under age 18, householder information (female, black, Hispanic, age: <18, 18-64, >64), and an indicator variable for households with female householder and no spouse present. In cases where no person was assigned to be the householder (*e.g.*, group quarters have no householder in the Decennial), we assigned the oldest person to be the householder. Using the person-level income variable, we created a variable for total annual housing unit income in 1999. All household information was again attached to each person's record.

5.3.3 Data source for MSA variable

The Population Division provided us with a file that included an indicator for "Living in a central city (MSA)". This indicator was merged to the Census 2000 records by state, county, and Census place code. Accordingly, 82,249,968 persons lived in a central city and there were 636 unique central cities.

5.4 Poverty stratum assignment

Households were assigned to strata based on income and household composition. Long-form households for whom an income variable was available were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold for that household type. The following list gives the poverty thresholds for various household types.

- if one-person-housing-unit, age of householder <=64 years, and no children under 18 years then poverty threshold 1999(hhpov1999)=8667;
- else if one-person-housing-unit, age of householder >64, and no children under 18 years then poverty threshold 1999 =7990;
- else if two-person-housing-unit, age of householder <=64 years, and no children under 18 years then poverty threshold 1999 =11156;
- else if two-person-housing-unit, age of householder >64, and no children under 18 years then poverty threshold 1999 =10070;
- else if two-person-housing-unit, age of householder <=64 years, and 1 child under 18 years then poverty threshold 1999 =11483;
- else if two-person-housing-unit, age of householder >64, and 1 child under 18 years then poverty threshold 1999 =11440;
- else if three-person-housing-unit and no children under 18 years then poverty threshold 1999 =13032;
- else if three-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =13410;
- else if three-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =13423;
- else if four-person-housing-unit and no children under 18 years then poverty threshold 1999 =17184;
- else if four-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =17465;
- else if four-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =16895;
- else if four-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =16954;
- else if five-person-housing-unit and no children under 18 years then poverty threshold 1999 =20723;
- else if five-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =21024;
- else if five-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =20380;
- else if five-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =19882;
- else if five-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =19578;
- else if six-person-housing-unit and no children under 18 years then poverty threshold 1999 =23835;
- else if six-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =23930;
- else if six-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =23436;
- else if six-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =22964;
- else if six-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =22261;
- else if six-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =21845;
- else if seven-person-housing-unit and no children under 18 years then poverty threshold 1999 =27425;
- else if seven-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =27596;

- else if seven-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =27006;
- else if seven-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =26595;
- else if seven-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =25828;
- else if seven-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =24934;
- else if seven-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =23953;
- else if eight-person-housing-unit and no children under 18 years then poverty threshold 1999 =30673;
- else if eight-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =30944;
- else if eight-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =30387;
- else if eight-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =29899;
- else if eight-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =29206;
- else if eight-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =28327;
- else if eight-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =27412;
- else if eight-person-housing-unit and 7 children under 18 years then poverty threshold 1999 =27180;
- else if >=9-person-housing-unit and no children under 18 years then poverty threshold 1999 =36897;
- else if >=9-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =37076;
- else if >=9-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =36583;
- else if >=9-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =36169;
- else if >=9-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =35489;
- else if >=9-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =34554;
- else if >=9-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =33708;
- else if >=9-person-housing-unit and 7 children under 18 years then poverty threshold 1999 =33499;
- else if >=9-person-housing-unit and >= 8 children under 18 years then poverty threshold 1999 =32208;

When income data were not available for long-form households, household composition was used to proxy poverty status. A household was assigned to the high poverty stratum if it had any of the following characteristics:

1) Female householder with children under 18 and no spouse present;
2) Living in a central city of a MSA and renter with rent less than $300;
3) Black householder and living in a central city of a MSA;
4) Hispanic householder and living in a central city of an MSA;
5) Black householder and householder less than age 18 or greater than 64;
6) Hispanic householder and householder less than age 18 or greater than 64.

Since short form respondents did not report income, the available household composition was used to proxy poverty status.

1) Female householder with children under 18 and no spouse present;
2) Black householder and living in a central city of an MSA;
3) Hispanic householder and living in a central city of an MSA;
4) Black householder and householder less than age 18 or greater than 64;
5) Hispanic householder and householder less than age 18 or greater than 64.

There were a total of 285,230,516 Decennial records, 64,493,265 of which were placed into the high poverty stratum, and 220,737,251 into the low poverty stratum.

5.5 Part B: Creation of stage-2 clusters for Census 2000 records

In order to group all Decennial individuals into the same stage-2 clusters for SIPP sampling, we first added SIPP sampling frame information to all Census 2000 records. The SIPP Survey Design Branch provided us with several files and memos containing SIPP sampling information. These files assigned Primary Sampling Units (PSUs) to geographic entities (mostly counties, with smaller counties grouped together to form a PSU); determined which PSUs were in the same risk pool to be sampled (*i.e.*, in the same stage-1 cluster); and reported which PSUs were in actuality sampled. Thus, the SIPP sampling frame information allowed us to begin with state and county information from the Decennial file and assign every Decennial record to a stage-1-cluster. We then combined the stage-1 cluster with the poverty stratum created in Part A and created stage-2 clusters.

5.5.1 Creation of PSUs

The original file containing the mapping between state/county and PSUs had 3,141 unique state/county observations and 1,928 unique PSU values. However, at this point we encountered a problem caused by the fact that SIPP sampling for the 1990s panels was based on 1990 geography definitions. Since we were creating weights with a reference point of April 1, 2000 and were linking to Census 2000, we needed to extrapolate the 1990 SIPP sampling frame to the year 2000. We therefore needed to take account of the county changes between 1990 and 2000. During that time period several counties were deleted/added/changed in such a way that their geographic changes needed to be addressed.

· Alaska: Denali (02-068) was created from part of the Yukon-Koyukuk Census Area (02-290) and an unpopulated part of the Southeast Fairbanks Census Area (02-240) in December 1990. Given that there were very few people in the area that was taken from the Southeast Fairbanks Census Area, Denali was assigned the same PSU as Yukon-Koyukuk. Yukon-Koyukuk Census Area and Southeast Fairbanks Census Area had different PSUs, but were in the same stage-1-cluster.

· Alaska: Skagway-Yakutat-Angoon Census Area (02-231) was split to create the Skagway-Hoonah-Angoon Census Area (02-232) and Yakutat City and Borough (02-282) in September 1992. Both new counties were assigned the PSU value of Skagway-Yakutat-Angoon Census Area and its stage-1-cluster code.

· Florida: Dade County (12-025) was renamed as Miami-Dade County (12-086) in November 1997. The county codes just needed to be changed for 2000.

· Montana: Yellowstone National Park (30-113) was annexed to Gallatin (30-031) and Park (30-067) counties in November 1997. Park County and Yellowstone National Park were assigned in 1990 to the same PSU and stage-1-cluster, Gallatin was assigned to a different PSU and stage-1-cluster. Because most people were moving to Park County from Yellowstone National Park and only very few people were living in Yellowstone National Park, no changes were made to the sampling frame, except that the record for Yellowstone National Park was taken out.

· Virginia: South Boston City (51-780) changed to town status and was added to Halifax County (51-083) in June 1995. In 1990 both South Boston City and Halifax County belonged to the same PSU. Therefore the change in county status was irrelevant for the assignment of counties to PSUs. The county code just needed to be changed.

The changes outlined above resulted in 2 additional state/county records and in the deletion of 2 other state/county records.

5.5.2 Creation of stage-1 clusters

The SIPP Survey Design Branch provided us with a file that assigned the 1,928 PSUs to 217 stage-1-clusters that were used to select PSUs to be sampled. Memos given to us provided the information about the PSUs that were actually sampled from. We merged that information onto the Census 2000 data by PSU.

5.5.3 Creation of stage-2 clusters

The Census 2000 file now held information on the 217 stage-1-clusters and on poverty status. The poverty variable had two values, high and low, and, hence, our final grouping of Decennial records contained 434 different stage-2-clusters.

5.5.4 Dropping Census 2000 records that were out-of-scope for SIPP samples

Because of the differing nature of a census and a program survey, we recognized the need to exclude some Decennial records as out-of-scope to be sampled for the SIPP. The SIPP Quality Profile 1998, third edition states:

> The survey population for SIPP consists of persons resident in United States households and persons living in group quarters, such dormitories, rooming houses, religious group dwellings, and family-type housing on military bases. Persons living in military barracks and in institutions, such as prisons and nursing homes, are excluded ... The survey population for the SIPP consists of adults (ages 15 and older) of responding households at the first interview. Each original sample member is followed until the end of the panel or until the person becomes ineligible (by dying, entering an institution, moving to Armed Forces barracks, or moving abroad) or leaves the sample. (page 17)

Several groups of the U.S. population that were counted in the Decennial but were out-of-scope for the SIPP based on the above definition and therefore were not considered when calculating the final weight. Accordingly, the following groups were not counted in the strata for the Decennial files:

1. Residents of the commonwealth of Puerto Rico, and residents of the outlying areas under U.S. sovereignty or jurisdiction (principally American Samoa, Guam, Virgin Islands of the U.S., and the Commonwealth of the Northern Mariana Island). This restriction excluded 3,808,610 persons.

2. Residents living in institutional group quarters: persons residing in correctional and juvenile institutions and nursing homes. This restriction excluded an additional 4,059,039 persons.

3. Residents living in non-institutional group quarters: persons living in military quarters, crews of maritime vessels, and staff residents of military institutions. This restriction excluded an additional 361,815 persons.

4. Children under age 18 (born before April 1, 1982). This restriction excluded an additional 72,145,912 persons.

In total, we excluded 80,375,376 Decennial records because they were out-of-scope for the SIPP samples.

5.5.5 Census 2000 stage-2 cluster tabulations

After removing the Census 2000 records that were out-of-scope for the SIPP, we made the appropriate Decennial cell counts for the 434 stage-2-clusters explained above. The 204,885,140 Decennial in-scope observations translated into a 472,016.45 mean cell count. The largest cell contained 4,578,514 observations and the smallest cell contained 8,754 observations.

5.6 Part C: Creation of poverty stratification variable for SIPP records

The creation of the poverty stratification variable for each SIPP record involved similar steps to those undertaken for the Census 2000 records. We first created the necessary variables. When data were available, household income in 1999 was created by summing monthly household income across all twelve months for 1999. The following demographic variables were taken from the earliest wave of the SIPP panel in which they were available for each respondent: birth year, birth month, sex, race, and ethnicity. All other demographic variables used for creating the poverty status of a household or the final weight were taken from the year closest to 2000 for each panel, i.e., the last year of each panel. These variables were: dummy variables for female householder, black householder, and Hispanic householder, age of householder (age categories were $<$18, 18-64, $>$64), number of children under 18 in the household, and whether a spouse was present in the household.

We then created the poverty stratification variable for each SIPP record. Individuals were assigned to strata based on either household income or household composition. For Gold-Standard respondents surveyed in the 1996 SIPP panel, 1999 household income was available (81,409 respondents) and they were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold for their household type. Thresholds were defined according to criteria used for the Census 2000 records (see 5.3).

For SIPP respondents from the early 1990s SIPP panels or for individuals who were missing from the later waves of the 1996 panel because of attrition, 1999 income data were not available. Household composition was used to proxy poverty status. A household was assigned to the high poverty stratum if it had any of the following characteristics:

1) Female householder with children under 18 and no spouse present;
2) Black householder and householder less than age 18 or greater than 64;
3) Hispanic householder and householder less than age 18 or greater than 64.

It was not possible to assign SIPP respondents to the high poverty stratum based on whether they lived in the central city of an MSA (as was done for the Decennial respondents) because this variable depended upon knowing state, county, and Census place code information for each household, and we did not have Census place code on the internal SIPP file. The final stage of adjusting the weight to correct population control totals within sex, race, ethnicity, and geographic location (see 5.10) handled this problem.

Of the 263,293 total individuals in version 4.0 of the Gold Standard file, 33,868 of them were placed into the high poverty stratum, and 229,925 into the low poverty stratum.

5.7 Part D: Creation of stage-2 clusters for SIPP records

As with Census 2000 records, we used the information provided by the SIPP Survey Design Branch to assign Primary Sampling Units (PSUs) to geographic entities. To assign each SIPP individual to a PSU, we needed state and county information. Unfortunately, county level geography was very difficult to obtain for the 1990-1993 SIPP panels. Given the likelihood that an individual's county had changed between the early 1990s and 2000, we did not invest in obtaining SIPP county information for the early panels. Instead we used the state variable recorded for respondents during the last year of their panel and then randomly assigned county and the corresponding PSU. For respondents from the 1996 panel, state and county geography was available and PSUs were assigned as they were for Census 2000.

Once SIPP respondents were placed in PSUs, the creation of stage-1and stage-2 clusters proceeded as outlined in 5.5. At this point, we used the link between the Decennial and the SIPP to flag SIPP individuals who matched to a Decennial record that had previously been determined to be out-of-scope, as explained in 5.5. The Gold Standard version 4.0 file contained 263,793 people, 177,165 of which were matched by PIK (*i.e.*, replacement SSN) to a Decennial record. Of these 177,165 records, 2,229 were matched to a Census 2000 record that was out-of-scope for the SIPP, meaning that these SIPP records received a zero weight in the final weight calculation. The remaining 261,564 SIPP in-scope records were used in calculating the weight. The link between the Decennial and the SIPP essentially served to indicate when a person interviewed in the 1990s had experienced a life-change by 2000 that removed them from the reference population.

After removing the SIPP records that matched to out-of-scope Decennial records, we made the appropriate SIPP cell counts for the 434 stage-2-clusters. The 261,564 SIPP in-scope observations translated into a 602.68 mean cell count. The largest cell contained 4,210 observations and the smallest cell contained 2 observations. The strata with very small numbers of SIPP observations could have presented a confidentiality problem when the weight is used on the SIPP/SSA/IRS public use file. We addressed this issue by synthesizing the weight in the Preliminary PUF 4.0 (see 5.14.

5.8 Part E: Matching SIPP individuals to Census 2000 records

There were 263,793 total SIPP individuals in the Gold Standard file, 177,165 of which were matched by PIK to a Census 2000 record. Of these 177,165 records, 2,229 were matched to a Decennial record that was out-of-scope for the SIPP, meaning that these SIPP records received a zero weight in the final weight calculation. Of these 177,165 records, 4,695 were matched by PIK to more than one Decennial record, because sometimes two Decennial records had the same PIK.

5.8.1 Un-duplication of SIPP-Census 2000 matches

Two match scores were created for each Decennial record. The first match score checked whether the Census 2000 record's date of birth and gender matched exactly to the date of birth and gender for that PIK in the Numident data. The first match score also checked whether the Decennial record's date of birth, gender, and race matched exactly to the same variables in the SIPP record. The first match score went up by 1 anytime the Decennial record matched on a variable (either to the Numident data or to the SIPP record). The second match score checked whether the Decennial record's date of birth, gender, and race were allocated or imputed, and went up by 1 anytime one of these characteristics was not allocated or imputed. After creating these match scores, the Decennial record with the highest first match score was chosen as the correct match for the SIPP record. If the two Decennial records tied on the first match score, the one with the highest second match score was chosen. If they tied on the second match score, one Decennial record was chosen at random as the correct match for the SIPP record.

5.8.2 Matching SIPP to Decennial through probabilistic record linking

The remaining 86,628 SIPP records were matched by probabilistic record linking to an in-scope Decennial record. The first blocking pass used 6 blocking variables and 7 matching variables. The 6 blocking variables were

a. psu (as defined above)
b. poverty stratum (as defined above)
c. male (dummy variable)
d. black (dummy variable)
e. Hispanic (dummy variable)
f. birth year

The 7 matching variables were

a. birth month
b. children under 18 (dummy variable)
c. no spouse present (dummy variable)
d. female householder (dummy variable)
e. black householder (dummy variable)

f. Hispanic householder (dummy variable)
g. age of householder (<18,18-64,>64)

Each SIPP record was assigned a set of Decennial candidates which agreed with that SIPP record exactly on all six blocking variables. Any SIPP record that had 30 or fewer Decennial candidates was considered unmatched and sent through the second blocking pass, which used 3 blocking variables and 10 matching variables. There were 72,866 SIPP records who had at least 31 Decennial candidates, and these SIPP records were each matched to a Decennial record using the same 7 matching variables as above.

For each matching variable, conditional m and u probabilities were created using the definitions in Fellegi and Sunter (1969):

- m = conditional probability that a SIPP-Decennial match had values for the matching variable that agreed exactly, given that the match was correct;
- u = conditional probability that a SIPP record and a randomly chosen Decennial record within the same set of blocking variables had values for the matching variable that agreed exactly.

The m conditional probabilities were estimated using the 177,165 SIPP records who were already matched to a Decennial record by PIK, and the u conditional probabilities were estimated by randomly assigning to each SIPP record in the first blocking pass one of its Decennial candidates. For both blocking passes (using 6 and 3 blocking variables, respectively), m and u probabilities were first created within cells using 3 blocking variables: psu, poverty stratum, and male. The cells were defined by the complete cross-classification of the three blocking variables psu, poverty stratum, and male. If there was a cell which had at least one SIPP record in the probabilistic record link, but no SIPP records in the set already matched by PIK to Census 2000, then that cell had no m probability. For these cells, m probabilities were estimated using coarser cells, first by only 2 blocking variables: psu and poverty stratum, and finally using no blocking variables. In other words, if a cell created from the complete cross-classification of 2 blocking variables was still missing an m probability because there were no SIPP records in that cell which had already been matched by PIK to a Decennial record, then that cell was assigned an m probability using all the SIPP records that had already been matched by PIK to a Decennial record. Whenever an m probability was created using a coarser set of cells, the u probability was created using the same set of cells. In other words, some m and u probabilities were created within cells that used three blocking variables, some within cells that used only two blocking variables, and some with no blocking variables, but the number of blocking variables used to create the m and u probabilities for a particular SIPP record always agreed.

Once m and u probabilities were created for each matching variable and for each SIPP record, agreement and disagreement weights were created for each Decennial candidate as follows: agreement weight $= \ln(m/u)$ and disagreement weight = $\ln((1-m)\,/\,(1-u))$. These weights were used to create a matching score for each Census 2000 candidate based on whether the Decennial candidate agreed with the SIPP record on the value of each matching variable. Then, the Decennial candidate with the highest matching score for each SIPP record was chosen as the correct match for that SIPP record.

The matching score was created in the following manner: if a Decennial candidate agreed exactly with its SIPP record on the matching variable, that Decennial candidate's matching score went up by the agreement weight for that matching variable, and if a Decennial candidate disagreed with its SIPP record on the matching variable, that Decennial candidate's matching score went up by the disagreement weight (which was always negative) for that matching variable. A few SIPP records had $u-$probabilities that were greater than $m-$probabilities for a particular matching variable (which differed across SIPP records), causing that particular matching variable to have no matching power for that particular SIPP record. In this case, that matching variable was not used in creating the matching score, so the matching score went up by zero whether or not the Decennial candidate agreed with its SIPP record on the matching variable.

Once all Decennial records were assigned a matching score, the Decennial record with the highest matching score for each SIPP record was chosen as the match in the following manner: For each SIPP record that was alone in a cell created from the complete cross-classification of the blocking variables (created from 6 blocking variables in the first blocking pass and 3 in the second blocking pass), and hence had a unique set of Decennial candidates, the Decennial record with the highest matching score was chosen as the match. If two or more records had identical matching scores, one record was chosen at random as the match. For the SIPP records who shared cells (created from the complete cross-classification of blocking variables) with other SIPP records, it was possible that two SIPP records each had the same Decennial record chosen as the match because it had the highest matching score. When this happened, the Decennial record with the higher matching score was chosen (or chosen at random if the two had identical matching scores), and the SIPP record that had been matched to the Decennial record that was not chosen was sent back through to receive another Decennial record as its chosen match from the pool of Decennial records that had not yet been chosen as a match for any SIPP record. This process was repeated until each SIPP record was matched to a Decennial record, and each Decennial record that had been chosen as a match was unique.

The second blocking pass contained the remaining 13,762 SIPP records who had 30 or fewer Decennial candidates from the first blocking pass, and used 3 blocking variables: $psu90sip$, poverty stratum, and male, and 10 matching variables: black, Hispanic, birth year, birth month, children under 18, spouse present, male householder, black householder, Hispanic householder, and householder's age. m and u probabilities and matching scores were created as they were in the first blocking pass, and a Decennial match was chosen for each SIPP record in the same manner as well.

5.9 Part F: Creation of preliminary weight

After all SIPP records were matched to Decennial records, a preliminary weight was calculated. In order to calculate this weight, we used the Decennial stage-2 cluster counts from 5.5 and the SIPP stage-2 cluster counts from 5.7. The preliminary weight was calculated using the following formula:

$$\text{prelim\_weight} = \frac{\text{number of records in Decennial stage-2 cluster}}{\text{number of records in SIPP stage-2 cluster}}$$

This preliminary weight was the same for all SIPP records in a particular stage-2 cluster. The weights ranged from 163.39 to 23,643.67, with a mean of 783.19.

5.10 Part G: Creation of the final weight

After calculating the preliminary weights we calculated population totals for the newly-weighted SIPP for particular subgroups. Given the discrepancy between these totals and the corresponding totals in the Census 2000, the weights needed to be controlled by population totals. We used a method called iterative proportional fitting to adjust the preliminary weights to reflect correct population totals for certain subgroups. This is the same method used by other Census Bureau surveys to calculate final weights. The list of subgroups used was the same list of population subgroups used to adjust the original 1996 SIPP sampling weights, and was provided to us by Tracy Mattingly from the SIPP Survey Design Branch.

To get the population totals for each subgroup, we used the Population Estimates Base for the U.S. civilian non-institutionalized population ages 18 and older on April 1, 2000 as released on the following Population Estimates web site on June 9, 2005: http://www.census.gov/popest/national/asrh/2004_nat_ni.html. A file containing the population totals used from this web site for April 1, 2000, and a spreadsheets containing the population totals that we calculated for certain subgroups have been supplied as part of this final report.

The iterative proportional fitting the preliminary weights to the population subgroup totals for the following demographic breakdown. We first divided the SIPP into four separate tables by race (black/non-black) and ethnicity (Hispanic/non-Hispanic). Then within each table, the rows of the table were the appropriate ages for that subgroup (provided by Tracy Mattingly) and the columns were male/female. The iterative proportional fitting raked the weighted SIPP tables (weighted by the preliminary weights) to the Population Estimates tables, where the numbers to rake to were both the row and column totals from the Population Estimates tables. The output was a set of adjusted tables. For each age/sex cell in a table, the ratio of the adjusted count for that cell to the unadjusted count for that cell was the factor which was multiplied by the preliminary weight to create the final weight for each individual. The final weights ranged from 30.90 to 32,625.69, with a mean of 780.64.

5.11 Geography issues

5.11.1 Different geography concepts on HCEF and other Census 2000 files

In order to establish the poverty indicator we used information from the HCEF and from other files that were merged onto the HCEF either through person IDs or geography (*e.g.*, central city indicators, PSUs). Merging by person IDs did not pose any problems, but merging by geography did. We were working with an internal HCEF file that had not yet been converted to the "tabulation" geography concept that the SCEF (as well as the other files) used. Our HCEF file had geography that was on collection geography level, which made it easier for the enumerators to perform the interviews. The SCEF file we used (as well as the files that we received from the pop-division and the SIPP Survey Design Branch) had tabulation geography (which was the geography concept that all the Census 2000 tabulations on the web used, for example). While state information was the same for "collection" and "tabulation" geography, county information could be different. Merging therefore was not straightforward. On our internal version of the HCEF was another geography variable (Current geography). This was not "Tabulation geography" either but matches it reasonably well. We used this variable to merge by geography.

5.11.2 Changing geography boundaries between the 1990s and 2000

Geography and, especially, county boundaries changed between 1990 and 2000. There were boundary changes as well as the deletion and creation of new counties during that time frame. This affected the use of SIPP-sampling units, because the SIPP sampling units for the 1991-1996 panels were created using the 1990 Census, and the SIPP sampling units for the 1990 SIPP sample used geography from before the 1990 Census. Also, people moved across county boundaries and therefore were counted in different PSU units in 2000 compared to the time they started participating in the SIPP survey. The changes that were made to accommodate the additions and deletions of counties were written down in detail in 5.5. We have not made any changes for counties that purely changed boundaries.

5.12 Birth date issue

Several people claimed to have been born on February 29, 1900. SAS did not accept this date as a leap year. We therefore changed the birth date for these people to February 28, 1900.

5.13 Overall evaluation of Gold Standard weight

Our method of creating an ex-post weight for the SIPP-SSA public use file utilizes our link between Census 2000 and the SIPP samples of the 1990s to determine how many people in the U.S. population each SIPP individual should represent. This weight will be a key component of the proposed public use product and will allow researchers to confidently represent the U.S. population as of April 1, 2000.

Table 2, Columns B and C presents the results of testing of the Gold Standard weight. We chose several selected statistics from the 2001 SSA Annual Statistical Supplement and calculated these same statistics using our weighted Gold Standard data. Our weighted Gold Standard file reproduces all of these selected statistics fairly closely. In particular, the number of workers receiving retirement benefits in December of 2000 in the Gold Standard data is lower than the number reported by SSA by only one million. The number of widows and widowers receiving benefits in the Gold Standard is lower than the corresponding SSA statistic by only 300,000, and the number of disabled receiving benefits is higher by only 800,000. The average monthly benefit received by these various sets of workers falls within 3%, 7%, and 6% of the SSA reported average monthly benefit for retired workers, widows and widowers, and disabled workers, respectively. The number of permanently insured individuals in December 2000 in the Gold Standard data falls within 1% of the corresponding number reported by SSA, and the number of wage and salary workers with taxable earnings for 2000 falls within 3% of the SSA reported number. The DER average earnings for 2000 in the Gold Standard is about $3,000 higher than the DER average earnings reported by SSA, and the SER average earnings for 2000 in the Gold Standard is about $1,400 higher than the SER average earnings reported by SSA. In general, we believe our Gold Standard weight does a particularly good job of reproducing these selected statistics from the 2001 SSA Annual Statistical Supplement.

5.14 Synthesizing the weight

The weights on the sixteen synthetic implicates were quite similar across implicates, allowing many observations to be identified across implicates by the value of their weight. Thus, we decided to create a synthetic weight for each synthetic implicate. We created synthetic weights by taking draws from a Dirichlet distribution to obtain the probabilities of having each possible value of the weight for each person in the data.

The theory for sampling from the Dirichlet distribution is described in Tanner (1996), Gelman et al. (2000) and Minka (2003). Suppose that each observation in the data can take on one of k possible outcomes. Let y be the vector of counts of the number of observations that take on each outcome. The multinomial distribution describes this data as follows:

$$p(y|n;\theta) \propto \Pi_{j=1}^{k} \theta_j^{y_j},$$

where θ_j is the probability of taking on the jth outcome category; these probabilities sum to one ($\Sigma_{j=1}^{k}\theta_j = 1$). The total number of observations is $\Sigma_{j=1}^{k} y_j = n$. The conjugate prior distribution for this multinomial distribution is known as the Dirichlet,

$$p(\theta|\alpha) \propto \Pi_{j=1}^{k} \theta_j^{\alpha_j - 1},$$

where the θ_j's are all nonnegative and again sum to one. The posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$. We call $a = \Sigma_{j=1}^{k}\alpha_j$ the "prior sample size" and we call $n = \Sigma_{j=1}^{k} y_j$ the likelihood component, or the "data sample size."

In our application, each person in the data can take on one of 55,552 possible values for the weight.[16] The sum of the weights played the role of the "data sample size, and equaled 204,044,727. We used a noninformative prior distribution by spreading additional observations evenly across the 55,552 cells; this was the "prior sample size."[17] The sum of the "data sample size" and the "prior sample size," is called the "posterior sample size" in the posterior Dirichlet distribution for the cell probabilities.

In practice, we replaced the likelihood counts, y_j, with their expected values. We used the SAS procedure PROC CATMOD to model the expected counts for each of the possible 55,552 cells created by the six strata variables (stage-1 cluster, poverty stratum, male, black, Hispanic, and age category). This procedure performs categorical data modeling of data that can be represented by a contingency table. We supplied the procedure with the weighted cell count data from the completed data, where each observation was a cell in the contingency table created by the complete cross-classification of the six strata variables, and each cell count was the weighted sum of the number of persons in that cell. The procedure used maximum likelihood analysis to estimate a log-linear model and calculate the predicted cell frequencies. We computed the maximum likelihood estimates using an iterative proportional fitting algorithm rather than the usual Newton-Raphson algorithm because it allowed us to obtain the predicted cell frequencies without performing time-consuming parameter estimation. The log-linear model included all six main effects (one for each stratum variable), all two-way interaction effects, and a single three-way interaction effect between the poverty stratum, black, and Hispanic variables.

We took four draws from the Dirichlet distribution for each input contingency table coming from one of the four completed data implicates, giving us a total of sixteen draws, one for each synthetic implicate. Each draw provided us with a vector of 55,552 posterior probabilities (which summed to one) for belonging to each of the 55,552 cells. We then multiplied these probabilities by the "data sample size," 204,044,727, to obtain the final weight value for each cell as a whole, and finally divided by the number of SIPP observations in each cell to obtain the final synthetic weight value for each person in that cell.

5.15 Evaluation of the synthesized weight

Table **2**, Columns C and D present the results of comparing the weighted completed data to the weighted synthetic data for the same published SSA statistics as were chosen for the testing of the Gold Standard weight. The results from the synthetic data very closely match those from the completed data. Column E shows that the percentage difference between these statistics for the two types of data is very small, ranging from no difference in the number of disabled workers receiving benefits to 4.4% difference in the number of widows and widowers receiving benefits. More specifically, the estimated number of individuals in the reference population receiving retirement benefits in December of 2000 in the synthetic data is lower than the estimated number in the completed data by 700,000. The estimated number of widows and widowers receiving benefits in the synthetic data is lower than the corresponding statistic in the completed data by only 200,000, and the number of disabled receiving benefits is exactly the same in the synthetic and completed data. The average monthly benefit received by workers in the synthetic data falls within 1% of the average monthly benefit in the completed data for all three types of workers. The number of permanently insured individuals in December 2000 in the synthetic data falls within 2% of the corresponding number in the completed

[16] This number of different possible values for the weight comes from the fact that the weight differed only by the values of the following variables: stage-1 cluster, poverty stratum, male, black, hispanic, and age category. There were 217 stage-1 clusters, 2 values for poverty straturm, male, black, and hispanic, and 16 age categories, resulting in 217*2*2*2*2*16 = 55,552 possible unique values for the weight.

[17] By agreement with the Census Bureau Disclosure Review Board, we do not disclose the prior sample size when a Dirichlet prior is used for confidentiality protection.

data, and the number of wage and salary workers with taxable earnings for 2000 falls within 1% of the corresponding statistic in the completed data. The DER average earnings for 2000 in the synthetic data is about $1,400 higher than the DER average earnings in the completed data, and the SER average earnings for 2000 in the synthetic data is about $800 higher than the SER average earnings in the completed data. Overall, we have shown that our weighted synthetic data does a very good job of matching our weighted completed data on these selected statistics from the 2001 SSA Annual Statistical Supplement.

6 Analytical Validity

Of primary importance to the success of any synthetic data set is the ability to preserve the univariate distributions of variables and to maintain relationships among variables. In this sense, the modeling done to create synthetic data is different than modeling done in order to predict future outcomes or to analyze cause and effect relationships that are important to policy makers. In creating synthetic data, the analyst's goal is to refrain from imposing prior beliefs about the relationships amongst variables and instead to allow the data themselves to determine the nature of these relationships. Thus, when modeling a particular variable, all other variables can potentially be used as explanatory variables, even when such a relationship might not seem sensible to a social science researcher. In practice, due to feasibility issues, the analyst must choose some subset of variables to go on the right hand side of the predictive regressions but the goal remains to impose as few prior beliefs as possible.

Once the synthetic data are created, however, a different kind of analysis becomes necessary, where prior beliefs become important. Standard economic and demographic models must be tested using the synthetic data and analysts with experience evaluating such results must determine whether the synthetic data are statistically valid. We define statistical validity according to Rubin (1996) as:

> First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands averaging over the sampling and posited nonresponse mechanisms. ... Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited nonresponse mechanisms. (p. 474)

This definition should be modified to include the phrase "confidentiality protection mechanisms" wherever "nonresponse mechanisms" appears.

Thus in order to assess the quality and usefulness of synthetic data, an analyst must determine what statistics are of interest, calculate these statistics, average them over the implicates of synthetic data, and then compare them to the best estimate of the same statistics from the completed Gold Standard data, which we will euphemistically call the "truth" since it is the best available comparison data. If the estimates are unbiased and the variances of the estimates are such that inferences drawn about the estimates are similar to the inferences in the completed Gold Standard (*i.e.*, "true") data, then the data are statistically valid.

6.1 Complete data estimation

Interest focuses on a complete data estimand Q which is a function of (X, Y) and has dimensions $(c \times 1)$. This estimand can be any computable, vector-valued function of the data. For example, it could be the average value of Y, many moments of Y, conditional moments of Y, given X, parameters of a model relating columns of (X, Y), percentiles of the distribution of Y, and so on. The essential feature of Q is that it is computable from complete data on the population and, therefore, is not random. To help clarify the ideas of this section, we will use the example of average income in 1990. If we had complete income data on every individual in the United States, *i.e.*, if we knew every element of Y $(N \times p)$ associated with the column representing 1990 income, we could calculate the national average with certainty.

Estimates of Q are random because they are based on D, which involves sampling from the finite population and incomplete observation of Y in the sample. We can only calculate an estimate of the average 1990 income because of the sampling involved with the SIPP and because not all SIPP individuals provided 1990 income data. When all sampled individuals provide data on all p variables, there are no item missing data. However, an estimator of Q is still random because of the sample design embodied in I. Even if all SIPP individuals in our sample reported 1990 income, the sample design of the SIPP would still make the average 1990 income a random variable. We will call the complete data estimator $q(D)$ and its variance estimator $u(D)$. Notice that because of the definition of complete data, q and u depend only on (X, Y_{obs}, I) and not on R. The analyst is assumed to have an inference system for $q(D)$ and $u(D)$. In particular, complete data inference can be based on $(q(D) - Q) \sim N(0, u(D))$, which may be exact or an approximation but is assumed to be appropriate in what follows.

6.2 Inference frameworks using multiple imputation

6.2.1 Missing data only

In the classic Rubin (1987) missing data application, Y_{mis} is imputed m times by sampling from $p(Y_{mis} | D)$, the posterior predictive distribution of Y_{mis} given D. The completed data consist of m sets $D^{(\ell)} = \left\{D, Y_{mis}^{(\ell)}\right\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw from $p(Y_{mis} | D)$ and is called the ℓ^{th} implicate. Continuing the example of 1990 income, we estimate the posterior predictive distribution of missing 1990 income conditional on everything else we observe about the individual (1991 income, gender, race, marital status, *etc.*). We sampled four times and created four implicates $D^{(1)}$, $D^{(2)}$, $D^{(3)}$, and $D^{(4)}$, each of which consists of original non-missing 1990 income data (D) and imputed 1990 income ($Y_{mis}^{(1)}...Y_{mis}^{(4)}$). Inference is based on the following formulae:

$$\text{statistic calculated on each implicate file: } q^{(\ell)} = q\left(D^{(\ell)}\right).$$

In our example the function q is the average of 1990 income across all individuals in the sample. This average is calculated separately for each implicate and then averaged across implicates as the next formula indicates:

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

The statistic $\bar{q}_m$ is the new quantity of interest and will serve as the basis for comparison with the synthetic data. Analytic validity requires that synthetic data reproduce $\bar{q}_m$, on average, and that inferences made about $\bar{q}_m$ remain the same, as expressed by the confidence interval associated with $\bar{q}_m$. In order to draw proper inferences, the correct variance measure must be used. The variance of $\bar{q}_m$ has two parts. The first part is commonly referred to as the "between-implicate" variance, defined by the following formula:

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

The measure b_m tells how much variation has been introduced by the multiple draws from the posterior predictive distribution. The second component of the overall variance of $\bar{q}_m$ is calculated by averaging the within implicate variance across implicates. We define the variance of $q^{(\ell)}$ for each implicate ℓ and the average across implicates as follows:

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

and

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

In our continuing example of 1990 income, $u^{(\ell)}$ is the sampling variance of average income (defined as $\frac{s^2_{income}}{N}$) for each implicate ℓ. The total variance of 1990 income is then calculated as a weighted sum of the between implicate variance and the average within implicate variance, defined as follows:

$$\text{total variance of the average statistic across implicates: } T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

When n and m are large, inference is based on $(\bar{q}_m - Q) \sim N(0, T_m)$. When m is moderate and the estimator $\bar{q}_m$ is univariate (*i.e.,* $c = 1$), 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

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

Proofs and further details can be found in Rubin (1987, 1996).

6.2.2 *Missing and partially synthetic data*

In order to analyze synthetic data that were created from data that originally contained some missing values, the missing data imputation and the synthetic data sampling must be done sequentially. First, complete m versions of D by sampling from $p\left(Y_{mis}|D\right)$. Denote the m completed data sets as $D^{(\ell)} = \left\{X, Y_{obs}, Y_{mis}^{(\ell)}, I, R\right\}, \ell = 1, \ldots, m$. Let the vector Z $(n \times 1)$ denote entities i for which any values of Y_{obs} have been synthesized. So, $Z_i = 1$ if any of the values of $Y_{obs,i}$ have been synthesized. Partition Y_{obs} into Y_{nrep} containing the rows where $Z_i = 0$ and Y_{rep} containing the rows where $Z_i = 1$. Then, for each completed data set, partially synthesize r implicates by sampling from $p\left(Y_{rep}|D^{(\ell)}, Z\right)$. Denote the r completed partially synthetic data sets as $D^{(\ell,k)} = \left\{X, Y_{nrep}^{(\ell)}, Y_{rep}^{(\ell,k)}, I, R, Z\right\}, k = 1, \ldots, r$ and where $Y_{nrep}^{(\ell)}$ corresponds to the rows of $\left(Y_{obs}, Y_{mis}^{(\ell)}\right)$ for which $Z_i = 0$ and $Y_{rep}^{(\ell,k)}$ corresponds to the rows of $\left(Y_{obs}, Y_{mis}^{(\ell)}\right)$ for which $Z_i = 1$. Note that $Y_{nrep}^{(\ell)}$ contains no synthetic data but may contain missing data imputations whereas $Y_{rep}^{(\ell,k)}$ may contain both missing data implicates (an element of $Y_{rep}^{(\ell,k)}$, say ij, for which item j is missing for entity i but not synthesized; entity i is in this set because $Z_i = 1$ whenever any element of Y_{inc} is synthesized) and synthetic data (an element of $Y_{rep}^{(\ell,k)}$, say ij, for which item j is missing for entity i and is synthesized; entity i is in this set because $Z_i = 1$ and element j element of $Y_{inc,i}$ is synthesized).

As with the case of missing data only, a statistic of interest is calculated for each implicate and averaged across implicates. However, because of the data structure that resulted from first completing missing data and then creating synthetic data, the averaging must account for the different types of implicates. Consider the continuation of the example of average 1990 income. Suppose there are 4 missing data implicates and that 2 synthetic implicates per missing data implicate were generated. In the notation used above, $m = 4$ and $r = 2$, which results in 8 unique data sets. We first calculate average income for each of the 8 implicates:

$$\text{statistic calculated on each implicate file: } q^{(\ell,k)} = q\left(D^{(\ell,k)}\right).$$

Then, we average across the 2 synthetic implicates that correspond to a given missing data implicate creating $\bar{q}^{(1)}$, $\bar{q}^{(2)}$, $\bar{q}^{(3)}$, $\bar{q}^{(4)}$ according to the formula:

$$\text{average of the statistic across the synthetic implicates: } \bar{q}^{(\ell)} = \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{r}$$

Finally, we average across all 8 implicates to create $\bar{q}_M$. This final average can then be compared to the $\bar{q}_m$ created from the missing data implicates only:

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

The variance calculations for data that have been completed and synthesized must also account for the additional source of variation that comes from synthesizing. Thus, we calculate the "between synthetic implicate" variance using the following formula:

$$\text{variance of the statistic due to variation in synthetic implicates: } b^{(\ell)} = \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{r-1}.$$

This formula quantifies the variation introduced by differences between two synthetic implicates that were generated from the same missing data implicate, *i.e.,* deviations of the synthetic implicate from the average across both synthetic implicates $q^{(\ell,k)} - \bar{q}^{(\ell)}$. We then average this variance over the missing data implicates:

$$\text{average of } b^{(\ell)} \text{ over missing data implicates: } b_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{m\left(r-1\right)} = \sum_{\ell=1}^{m} \frac{b^{(\ell)}}{m}.$$

The next source of variation comes from the multiple implicates due to missing data completion. This variance is calculated using the deviations of the average for a missing data implicate from the overall average, *i.e.,* $\bar{q}^{(\ell)} - \bar{q}_M$. This is the "between missing data implicate" variance:

$$\text{variance of the statistic due to variation in missing data implicates: } B_M = \sum_{\ell=1}^{m} \frac{\left(\bar{q}^{(\ell)} - \bar{q}_M\right)\left(\bar{q}^{(\ell)} - \bar{q}_M\right)'}{m-1}.$$

Finally, the last source of variance comes from the within implicate variance, which is averaged across the synthetic implicates for a given missing data implicate and then averaged across all the implicates according to the formulae:

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

$$\text{average variance of the statistic across synthetic implicates: } \bar{u}^{(\ell)} = \sum_{k=1}^{r} \frac{u^{(\ell,k)}}{r}$$

and

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

The total variance is, once again, a weighted sum of the difference sources of variation–between synthetic implicate, between missing data implicate, and within implicate:

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

T_M is the variance used to draw inferences about $\bar{q}_M$ and variation introduced by the synthetic and missing data implicates must not be so large that the inferences will be substantially different from those drawn using $\bar{q}_m$ and T_m.When n, m and r are large, inference is based on $(\bar{q}_M - Q) \sim N(0, T_M)$. When m and r are moderate and the estimator $\bar{q}_M$ is univariate (*i.e.,* $c = 1$), 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

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

Proofs and details can be found in Reiter (2004).

6.3 Application to the SIPP/SSA/IRS-PUF

Version 4.0 of the public use file consists of 16 implicates. We created four implicates in the missing data completion phase and then created four synthetic implicates per missing data implicate, thus $m = 4$ and $r = 4$. We chose to focus on two types of statistics–regression coefficients and univariate statistics (means, variances and percentiles) because these are most likely to be of interest to the potential users of our public use file. When showing regression results, we report $\bar{q}_m$ and $\bar{q}_M$ as vectors of regression coefficients. To calculate $\bar{q}_m$ we run the same regression on each of the four missing data implicates and then average the coefficients across implicates. To calculate $\bar{q}_M$ we run the same regression on each of the 16 synthetic implicates and then average the coefficients across these implicates. We also report the variance associated with each average coefficient in the form of vectors that contain the diagonal elements of the covariance matrices T_m and T_M. In the same format we report the standard error (square root of diagonal elements of T_m and T_M), $t-$ratio (each coefficient divided by the standard error), degrees of freedom (calculated using formulae above), and upper and lower bounds of the 95 percent confidence interval. To show the effect of the two types of implicates on the total variance calculation, we also report the component pieces of the overall variance: diagonal elements of $B_M, b_M, \bar{u}_M$ for the synthetic data, and b_m and $\bar{u}_m$ for the missing data. Univariate statistics are reported in the same manner except the results are scalars instead of vectors.

6.4 Results

6.4.1 General interpretation

When comparing results from completed data to results from synthetic data, there are a number of things to consider. First, and most obvious, is how closely to the point estimates correspond to each other. Regression coefficients and moments of the univariate distribution should be similar between the two data sources. However, this leads to the obvious question: "How similar is similar enough?" To answer this question it is important to compare the confidence intervals surrounding the point estimates. In an ideal situation, the point estimates are very close and the confidence intervals completely overlap, presumably with the synthetic confidence interval being slightly larger because of the increased variation due to synthesizing. Results like this give us confidence that the point estimates really are very similar and that inferences drawn about the coefficients will be the same whether one uses synthetic or completed data. In cases where the point estimates are somewhat further apart, the confidence intervals give us some idea of how far off we are. If there is still some overlap, then the synthetic and completed analyses are not so radically different. In cases where there is no overlap of the confidence intervals, the synthetic variable will need to be carefully examined to determine what might have caused the discrepancy.

Even in cases where the synthetic confidence interval contains the entire completed data confidence interval, we might still be concerned with the relative size of the synthetic interval. If the synthetic point estimate is in the middle of a very large interval, then inferences drawn using synthetic data may be too weak. This could happen because the variables being synthesized cannot be well-modeled and, therefore, each synthetic implicate introduces considerable variation into the analyses that involve those variables. This problem can be improved by the creation of more synthetic implicates. Higher numbers of r implicates would reduce the between $r-$implicate variance, b_M, and tighten the confidence intervals. It would also solve another potential problem. If b_M is too large in the synthetic data, the overall variance T_M can become negative because the b_M term is subtracted in the total variance formula. A large between $r-$implicate variance swamps other sources of variation and makes the synthetic total variance undefined. When we have cases like this in our results, we revert to the asymptotic formulae (based on $r = \infty$), and note this in the tables. Essentially we calculate T_M as the weighted sum of the between $m-$implicate variance and the within variance and do not subtract the between $r-$implicate variance. Then we treat the coefficients as if they were normally distributed and calculate the confidence intervals using the appropriate critical points from the normal distribution instead of from the $t-$distribution. In the tables we create an indicator called $flag\_dfnotexist$ which indicates that we could not calculate degrees of freedom for a $t-$distribution. In cases where the degrees of freedom are less than or equal to two, we also indicate that degrees of freedom do not exist and use the asymptotic (in r) normal distribution to calculate the confidence interval.

It is important to note one more detail about the univariate and regressions results we present here. We have used the weight that we created by matching individuals in our sample to the Census 2000 micro-data. Hence, in both the completed data and the synthetic data, all the statistics we report are weighted and should be interpreted as representative of individuals from the civilian non-institutional U.S. population age 18 or older as of April 1, 2000.

6.4.2 Summary statistics for OASDI beneficiaries

Tables 3-18 give results comparing means of important earnings and benefits variables by demographic group and type of benefit for individuals who became OASDI beneficiaries during the time period covered by these data (*i.e.,* had date of initial entitlement between 1951 and 2002). Tables 3-10 show results for SER work indicators (positive FICA covered earnings in a year) and SER earnings (total FICA covered earnings up to the maximum). As in version 3.1, the percentage of individuals who worked in a given year is very close, on average, for all the groups and across all the years and the confidence intervals overlap. In addition, average earnings are now much closer for all the groups. For example in 1995 average earnings for white males who retire at some point were \$10,347 in the synthetic data and \$11,012 in the completed data. For white females who retire the correspondence is even closer: \$5,495 versus \$5,566. Particularly strong improvement was made for black males. In 1995, black males who retire at some point earned \$8,856 on average in the synthetic data and \$8,564 in the completed data and there is almost complete overlap in the confidence interval. Synthetic earnings data for this group was particularly problematic in earlier versions so this result represents a significant step forward in our modeling. Charts 1 and 2 show the time trend for labor force participation for the four main demographic groups for individuals who retire at some point. Charts 3 and 4 show the same time trend for earnings for the same groups. Labor force participation and earnings trends are the closest for white women, followed by black women and black men. White men have a slightly higher discrepancy between synthetic and completed earnings in 1985. Still the trend is the same and other years have closer correspondence.

As shown in Tables 11-12, Total SER earnings summed over all years 1951-2003 and total number of years with positive earnings are also very close for most groups. White females who retire at some point earned on average $192,468 over this time period according to the synthetic data compared to $198,303 in the completed data and they worked a total of 26.17 versus 26.69 years. None of the individuals who retire or receive disability benefits have total years off by even a full year when comparing the synthetic and completed data. Total earnings differ by between $1,000 (black males) and $25,000 (white males). Table 13 shows that patterns of work are also similar between the synthetic and completed data. This table reports on balances in a "personal account" created by taking 2% of earnings annually from age 21 or 1951 (which ever was later) and compounded it annually at 5% interest until the date of entitlement. If individuals worked in predominantly different decades in the synthetic versus the completed data, we would expect the accumulated totals to be very different. Instead these totals are quite similar, even for white male retirees (less than $1000 difference).

Tables 14-18 compare the synthetic and complete OASDI benefit variables. Table 14 shows that synthetic year of initial entitlement is very close to the completed value, with the differences being less than 6 months for retirees and disability recipients in every demographic group. The confidence intervals are very tight for this variable and the overlap between synthetic and completed is high. Initial MBA in Table 15 is equally well synthesized. The +/-$50 restriction helps to ensure that the averages for every category in this table are very close. Retired white males received $716 in benefits on average in the synthetic data, compared to $730 on average in the completed data. Table 16 shows similar results for average monthly benefit amount in April 2000. Finally tables 17 and 18 show the average $AIME/AMW$ and PIA for the various demographic/beneficiary groups. We remind users of the data that these variables were not taken from the Master Beneficiary File but were calculated according to basic retiree benefit formulae for both the synthetic and complete data. Comparisons of these variables between the two data types is simply another way of summarizing the SER earnings data and judging how well we synthesized the history. On average the $AIME/AMW$ is very close: $725 synthetic versus $765 completed for white retired females and $1,693 versus $1,789 for white retired males. The PIA is $411 versus $434 and $749 versus $776 for the same groups. These results provide us some level of confidence that our synthetic earnings history will produce valid benefit calculations.

6.4.3 Summary statistics for all workers

Tables 19-28 show comparisons between some of the same synthetic and completed variables described above but for all workers instead of just OASDI beneficiaries. Specifically, Tables 19-26 show percentages of individuals who worked and average earnings for the years 1965, 1975, 1985, and 1995. These comparisons show very close correspondence between the synthetic and completed data. Average earnings for white males in 1995 is $17,047 using synthetic data and $17,241 using completed data. Of the white males in our sample, 67.1% had positive FICA covered earnings in 1995 according to the synthetic data versus 67.5% according to the completed data. These results are consistent across years and demographic groups. Charts 5-8 show these trends graphically. For whites, the synthetic and completed time trends lie on top of each other. For blacks, there are a few more differences, in particular earnings for black males seem to diverge a bit in 1995, but generally the time trends are close and show the same pattern.

Tables 27 and 28 compare total earnings and years worked from 1951-2003. The group with the closest correspondence on average between the synthetic and complete data is white females (total earnings of $211,817 versus $212,751 and total years 17.76 versus 17.99). The group with the largest difference is black males ($257,525 versus $240,933 and 18.99 versus 18.41 years). None of the groups differ by more than half a year in the total number of years worked and both black females and white males differ by less than $10,000 in total earnings.

6.4.4 Summary statistics by education categories

We next consider means of several important variables stratified by race, gender, and education category. In our analyses of version 3.1, we found that the relationship between education and other variables had not always been well preserved in the synthetic data. In this version of synthetic data, we find some improvements in this respect. Tables 29-34 show means for monthly benefit amount in April 2000, uncapped non-deferred earnings from all FICA-covered jobs in the year 2000, total FICA covered earnings up to the maximum in the year 2000, and percentages of foreign-born, Hispanic, and disabled individuals by race, gender, and education. Both mba_2000 and $totearn\_ser\_2000$ show very close correspondence between the synthetic and complete data. For example, calculated using the synthetic data, white male college graduates received, on average, \$786 in monthly benefits in April 2000 compared to \$812 calculated from the completed data. For this same group, the confidence interval also overlaps \$758 to \$814 in the synthetic data versus \$796 to \$828 in the completed data. The same group has FICA covered earnings of \$34,103 on average in the synthetic data compared to \$35,830 in the completed data with complete overlap in the confidence interval. Synthetic and completed average FICA covered earnings are particularly close in the "some college" and "college degree" categories for every demographic group, differing by no more than \$500 in almost every group. Total non-deferred earnings ($nd\_der\_fica\_2000$) have greater differences between the synthetic and completed data where the discrepancies range between \$1,000 and \$5,000 on average. The confidence intervals calculated from the synthetic data for the "high school degree," "some college," and "college degree" categories overlap the confidence intervals from the completed data for almost every demographic group. The confidence intervals for the "graduate degree" and "no high school degree" categories are generally very large in the synthetic data and sometimes do not overlap the completed data. These categories contain far fewer individuals and, hence, are more difficult to model; consequently, the synthetic data display more uncertainty.

The percentages of individuals who are foreign-born and Hispanic are also very close for the demographic and education sub-groups. Again the three middle education categories show particularly close correspondence between the synthetic and completed data. For example 9.3% of white males with some college are foreign-born according to the synthetic data compared to 9.4% in the completed data. The synthetic and completed data both give 9.0% of individuals as being Hispanic for the same group. In both cases there is complete overlap in the confidence intervals. In past versions, Hispanic was a particularly difficult variable to synthesize but these results seem to indicate that we have made significant progress modeling this variable. Percentages of individuals who report being disabled in the SIPP are also relatively consistent between the synthetic and completed data. White males are the closest across all education categories (%disabled synthetic - %disabled completed $< 1\%$ for all groups except graduate degrees) and black males are the most different (but still %disabled synthetic - %disabled $<2\%$ for all groups except graduate degrees), but in all cases there is significant overlap in the confidence intervals.

6.4.5 Summary statistics by foreign born

We also consider means of the three administrative variables discussed in the previous section, stratified by race, gender, and foreign born in Tables 35-37. Both uncapped and capped earnings from FICA covered jobs are relatively close for non-foreign-born and foreign-born workers in the year 2000. For uncapped earnings (DER), the confidence intervals for white females and white males completely overlap, although the intervals for the foreign-born workers are significantly larger in the synthetic data, as one would expect for the smaller sub-group. For black females the confidence intervals overlap but are shifted up slightly in the synthetic data and the same is true for foreign-born black males, while non-foreign-born black males have a large synthetic confidence interval that completely overlaps. A similar pattern holds for capped earnings (SER). Average monthly benefit amounts in April 2000 show some differences when stratified by foreign-born. The difference in the average benefit between the synthetic and completed data is about \$60 for black female, white male, and black male foreign-born individuals. Even though synthetic mba_2000 is always within \$50 of the completed value for this variable at the individual level, different individuals end up in a given sub-group because black and foreign-born are synthesized. Hence multiple individuals in the completed data who were not foreign-born may have been synthesized to be foreign-born, and hence may move from one sub-group to another, changing the overall average monthly benefit amount in that sub-group by more than \$50, even though as individuals their personal monthly benefit amounts did not change by more than \$50. For these three groups of foreign-born individuals, there is some overlap in the confidence intervals, although the synthetic data interval is somewhat lower than the completed data interval.

6.4.6 Summary statistics for marital histories

Table 38 shows means and confidence intervals for six marital history variables: number of marriages, percent ever divorced, percent ever widowed, duration of 1st marriage, duration of 2nd marriage, and age at first marriage. The first three variables are nearly indistinguishable on average between the synthetic and completed data, clearly the result of a successful Bayesian bootstrap of $mh\_category$. The durations are shorter in the synthetic data than in the completed data for both the point estimates and the confidence intervals by 2-3 years. Age at first marriage is approximately 23 years in both data types. The consistent synthesis of these marital history variables is another major step forward given that past versions of the synthetic data contained synthetic values that did not even meet minimum consistency standards with the unsynthesized and other synthesized variables.

6.4.7 Age at time of retirement

Of particular interest when considering the synthesis of birth date and year of initial entitlement is whether these two variables are consistent enough with each other to produce an expected distribution of retirement ages. Table 39 gives both weighted and unweighted counts of individuals who retired (*i.e.,* had $tob\_initial = 1$) at different ages and Charts 9-10 graph the weighted and unweighted distributions respectively. The first important thing to note is that the completed data have some discrepancies between recorded retirement age and legal retirement age. There are almost 5,000 individuals in our sample whose original administrative birth date and year of initial entitlement imply that they retired between age 61 and age 62. It also appears that in the completed data there are large numbers of individuals retiring at age 62 and at age 64. We had expected the spike at age 62 but thought the later spike would be at age 65. In our synthetic data, we attempt to impose the restriction that retirees must be at least 62 and are successful in all but a few cases. Hence the group retiring between ages 61 and 62 vanishes in our synthetic data. The synthetic data also have a high point at age 62 but then taper off more uniformly across ages 63, 64, and 65. Ideally the counts of individuals retiring at age 63 in the synthetic data might have dropped off more quickly. However the modeling is difficult here because the completed data are not entirely as expected and we are forcing some data consistency that does not exist in the original data. Given our careful modeling of date of initial entitlement and its close correspondence on average between the synthetic and completed data, more research is needed to determine the exact cause of the differences in these distributions.

6.4.8 Selected regression results

We begin our discussion of regression results with Tables 40-43 where the dependent variable is the log of total DER earnings (sum of deferred and non-deferred at FICA and non-FICA jobs) in the year 2000. We ran four separate regressions for each of the major demographic groups: white males, black males, white females, and black females. The closest correspondence between the synthetic and completed regression coefficients is in the education variables which always have the same sign and generally have significant overlap in the confidence intervals. The exceptions for overlapping confidence intervals are usually the graduate degree indicator, not surprising given the results in the means presented earlier. The demographic group with the closest synthetic and completed education coefficients is white males. The coefficient on high school degree only in the synthetic data regression is .214 compared to .230 using the completed data, and for some college, the coefficients are .400 and .431 respectively. In both these cases the confidence intervals in the synthetic data contain the confidence intervals in the completed data. In comparison the high school degree only coefficients for black females are .263 and .347 for synthetic and completed data respectively and for some college the coefficients are .494 and .587. The confidence intervals overlap to a great extent but not completely.

The other SIPP demographic variables, Hispanic, disabled, and foreign-born, are not as consistently similar between the synthetic and completed data but they have improved significantly compared to prior versions of the synthetic data. Foreign-born and disabled always have the same sign and Hispanic has the same sign in the regressions for white males and black females. For white males and females the confidence intervals for foreign-born and disabled overlap, and for black males and females the confidence intervals for all three variables overlap. The magnitudes of the coefficients differ but the confidence intervals give reason to be hopeful that the synthetic data are not producing estimates that are entirely different from the completed data.

The right hand side variables with the most discrepancies in these regressions are the experience coefficients (years of positive SER earnings, with squared, cubed, and quartic terms). While the signs are generally the same and the point estimates of the higher order terms are sometimes similar in magnitude, the confidence intervals do not usually overlap, meaning that the synthetic and completed coefficients are significantly different. Using the synthetic data provides a lower return to experience than using the completed data. For example the coefficient on years of experience for white males is .173 in the synthetic data regression versus .275 in the completed data. For black males the difference is .173 versus .388.

Tables 44-47 show regression results where the dependent variable is log of total SER earnings (capped at FICA maximum) in the year 2000. These results follow the same pattern as the DER earnings. The education coefficients are quite similar between the synthetic and completed data and the confidence intervals overlap to quite a large extent.

The signs for disabled, foreign-born, and Hispanic agree in all four regressions and the confidence intervals for black males and females and white males overlap for Hispanic and foreign-born. In addition they overlap for disabled for black males and females and Hispanic for white females. The experience coefficients are again significantly different in the synthetic and completed data. Not surprisingly, the coefficients on marital status indicators are very similar between the synthetic and completed data. Since marital status was not synthesized and was a grouping variable in the modeling of earnings, this is to be expected.

Table 48 shows results for a regression of the log of the $AIME/AMW$ variable on various demographic characteristics. The results for this summary measure of earnings generally show point estimates that are quite close between the completed and synthetic data. The race/gender interaction terms have overlapping confidence intervals except for black females and even in this case the point estimates and the intervals are not very different (-.928 versus -.995). The education coefficients all have overlapping confidence intervals with the exception of graduate degree. The Hispanic and marital status indicators are all very close both in terms of confidence intervals and point estimates. Only disabled shows significant bias. The age coefficients are slightly different between the synthetic and completed data but the confidence intervals do overlap.

Tables 49-52 show results using the log of the monthly benefit amount, first at the date of initial entitlement for retired and then disabled workers, followed by the April 2000 amount for the same two groups of beneficiaries. We believe that the monthly benefit amounts are some of the most accurately synthesized variables and these regression results support this belief. For the initial MBA for retirees, all the confidence intervals overlap with the exception of log of total net worth. Even the disability variable performs well, with the synthetic data coefficient being -.048 and the completed data coefficient being -.039 and very close overlap of the confidence intervals. We also include as a right-hand side variable the percentage of years an individual worked from age 15 to time of retirement/death/end of data, whichever was first. This variable also performs similarly in the synthetic and complete data, causing us to be optimistic about the preservation of relationships between years worked and benefits received. The results are equally encouraging for beneficiaries receiving disability payments. Again almost all the confidence intervals overlap (with the exception of graduate degree) and the point estimates are very close. Particularly reassuring is the fact that age at time of initial entitlement and year of initial entitlement have extremely close coefficients in the synthetic and completed data regressions.

The April 2000 monthly benefit amount regressions in Tables 51-52 continue to show close correspondence between the synthetic and completed data. For retirees, the education coefficients are particularly similar between the two data types with overlapping confidence intervals for all four education indicators. Age in the year 2000 and the race/gender indicators are also essentially statistically the same. Even Hispanic and disabled, which have greater differences in the point estimates, have some overlap in the confidence intervals. Log of total net worth continues to be the largest discrepancy between the two regressions. The results for recipients of disability payments are similar. The point estimates are slightly further apart but there is still significant overlap in the confidence intervals.

Tables 53-54 consider comparisons between regressions results using log of total family income and log of total personal income in the year 1999. These SIPP variables have generally been difficult to model because of the large amounts of missing data due to the combination of five SIPP panels. Income variables for 1999 had to be completed for all individuals in the 1990-1993 panels. Thus the majority of completed data is imputed and not taken from survey responses. This fact contributes to difficulties in modeling these variables for synthesis. For many of the coefficients in these regressions, the synthetic degrees of freedom are missing, either because the total variance was negative or because the degrees of freedom calculated using the appropriate formula was not strictly greater than 2. It appears that the between synthetic implicate variance relative to the between completed implicate variance is too high with respect to these variables. In spite of this fact, the point estimates are often quite close between the synthetic and completed data total family income regressions, especially for marital status indicators, type of benefit receipt indicators, race/gender indicators, year of birth, and the first two education categories. Using the asymptotic formula for the confidence intervals in the synthetic data when the degrees of freedom do not exist, there is overlap with the completed data intervals for almost all variables (college only, graduate, disabled, and total number of kids in the family are the only exceptions). Results for total personal income are similar and even show some improvements over total family income. There is overlap in the confidence intervals for the college only and graduate degree indicators so that now only two sets of coefficients are statistically different - disabled and total number of kids in the family.

Table 55 shows results from a logistic regression of a pension indicator variable on various demographic and economic control variables. These regression coefficients have similar problems to those in the income regressions, namely the synthetic degrees of freedom often do not exist. Again, in spite of this, the coefficients from the synthetic data regression and the completed data regression generally agree in sign and have overlapping confidence intervals that confirm that the magnitudes are not entirely disparate. Some notable exceptions to this are age and age squared in the year 2000 and the industry and occupation indicators where the confidence intervals have no overlap.

Table 56 shows results from another logistic regression where the dependent variable was an indicator for positive weeks with pay in 1999, taken from the SIPP survey. The results here are somewhat mixed. The race/gender coefficients have very similar point estimates and overlapping confidence intervals. The education coefficients diverge somewhat from the completed and synthetic data regressions. Only some college has an overlapping confidence interval. Total number of kids, foreign-born, and Hispanic all have overlapping confidence intervals and quite close point estimates. An indicator for positive benefit receipt in 2000 does not have an overlapping confidence interval. The coefficient on average log real earnings from the DER, calculated using earnings from 1978-2003, has a different sign and non-overlapping confidence interval between the synthetic and completed data regressions. However the quartile indicators calculated using the distribution of this variable have similar coefficients in the two regressions and overlapping confidence intervals. For percentage of eligible years worked in the SIPP, the point estimate appears to disagree between the synthetic and completed data regressions but the confidence interval does overlap and again the quartile indicators are much closer in magnitude with overlapping confidence intervals. The coefficients on the marital status indicators are consistent between the two data types but coefficients on marital history variables that indicate whether the individual has ever been divorced ($divorced1$) or ever been widowed ($widowed1$) are less consistent, with non-overlapping confidence intervals.

The final four tables of regression results 57-60 use wealth variables from the SIPP as the dependent variables. Table 57 begins with log of total net worth for married couples, with one observation per married couple. Control variables for both members of the couple have been included on the right-hand side of the regression. The results here are not as encouraging as in other regressions. The education indicators for both members of the couple have significantly different effects in the synthetic and completed data regressions. Total number of kids, foreign-born, foreign-born-spouse, Hispanic, and Hispanic-spouse do have overlapping confidence intervals but an indicator for owning a home does not. The indicator for receiving retirement benefits also does not have an overlapping confidence interval although all the other benefit indicators do. Table 58 show results from a regression of log of total net worth for single individuals on characteristics of the individual. Again there is often not overlap between in the confidence intervals. Only the high school degree only indicator has any overlap among the education coefficients.

Tables 59-60 show results for the home equity variables. There is some overlap in the confidence intervals for the education coefficients for the main individual in the couple (chosen as whichever individual had the first sorted person identifier variable, basically random) but the spouse education coefficients are significantly different. There is also overlap for the quartile of average log real DER earnings for both the individual and the spouse. Some of the coefficients on the SSA benefit indicators have overlapping confidence intervals and some do not. Year of birth is very similar in the synthetic and completed data. The results for single individuals are similar.

6.4.9 Univariate distributions of continuous variables

Table 61 examines univariate distributions for all the continuous variables in our sample. The continuous variable synthesis techniques used in this project generally did a very good job of modeling the overall univariate distributions of a variety of variables. The percentiles of the synthesized variables match closely with the percentiles of the corresponding completed variables, capturing the general shape of the distribution; although, very sudden spikes and cliffs in the distributions do get smoothed out a bit. Some of the variables had their synthetic draws restricted to rather narrow windows making the close match not too surprising, but even the variables whose synthesis was unrestricted resulted in very similar univariate distributions.

The three date variables were all restricted to be close to the unsynthesized values (when in scope). Synthetic $birthdate$ (restricted to be within one year of administrative $birthdate\_pcf$) and synthetic date of initial entitlement (restricted to be within 2 years) are extremely close to their completed counterparts with all the percentiles within a couple months of each other. Synthetic $deathdate$ appears to struggle a little bit on the lower end of the distribution, but it turns out that this is do to a quirk in the synthetic weight. Unweighted, the synthetic and completed distributions of $deathdate$ are also very similar, but the completed files give zero weight to the people who die before the year 2000. The construction of the synthetic weight did not preserve this characteristic, thus making the weighted percentiles at the lower end of the synthetic $deathdate$ distribution seem significantly lower than the completed $deathdate$.

The MBA variables–MBA in the initial month of benefit receipt ($mba\_initial\_real$) and MBA in April 2000 (mba_2000)–were restricted to be within $50 of the original amounts, thus it is no surprise that the univariate distributions and means were preserved nearly perfectly.

The continuous marital history variables were synthesized without any constraints. As one can see, this did not affect the quality of the age at first marriage synthesis. The synthetic distribution lies almost exactly on the completed distribution. The duration variables which measure the lengths of all applicable events in the marital history–length of first marriage if ever married ($duration\_mar1$), length of single spell after first marriage if the first marriage ended ($duration\_end1$), length of second marriage if there was a second marriage ($duration\_mar2$), *etc.*–exhibit some of the smoothing that can take place in the synthesis when extremely sharp changes occur in the density of the completed variable. For example, $duration\_end1$ has an extremely dramatic rise somewhere between the 50th and 75th percentiles in the completed data. The synthetic data matches the 25th and 75th percentiles well, but overestimates the median because it has smoothed this spike out a bit. It is also worth noting that some of these duration variables for second, third, and fourth marriages have very small sample sizes which makes synthesis a little less accurate. Nevertheless, the synthetic and completed distributions for these variables match quite closely except for a little smoothing here and there.

The wealth variables have some of the toughest distributions to synthesize. They are highly skewed and have extreme outliers on the high end of the distribution. For both $homeequity$ and $nonhouswealth$, the synthesized variables tend to under-estimate the lower end of the distribution and over-estimate the upper end of the distribution, while $totnetworth$ also underestimates the lower end of the distribution but matches the upper end of the distribution. The means, however, look very good, and the general shape of the distribution is preserved.

The DER earnings arrays present some of the same challenges as the wealth variables only to a lesser degree. They also have some very large outliers and are heavily skewed. As a result, the synthetic values display some of the same problems as the synthetic wealth variables, but again, to a lesser degree. The lower ends of the distributions tend to be slightly underestimated and the upper ends slightly overestimated. The deferred earnings arrays have extremely small sample sizes and struggle a lot more than the non-deferred earnings arrays, but once again, the means and general shape of the distributions are preserved very well for all the years.

The SER earnings are capped and, therefore, take away one challenge of extreme outliers and introduce a new problem of a truncated distribution. The cap was modeled by introducing another binomial parent variable indicating whether an individual earned equal to or more than the cap in a given year. If not, the amount was modeled with our continuous variable techniques and the draws were restricted to lie between $0 and the cap. For the most part, these distributions look very good putting about the same amount of weight at the cap and matching the lower percentiles quite closely.

Finally the continuous SIPP arrays all look quite good at the overall univariate level. Although these variables sometimes exhibited analytical difficulties in multivariate analyses, the general approach used for transforming and modeling continuous variables has done an excellent job of matching the percentiles for almost all of these variables. The weeks worked variables are constrained to lie between 0 and 52, but otherwise the synthesis for all the SIPP arrays was unconstrained.

6.4.10 Counts and percentages of categorical variables

Finally, Table 62 shows weighted and unweighted counts and percentages of some of the basic demographic and benefit variables in the synthetic and completed data. Included variables are: $male$, $black$, $Hispanic$, marital status, $tob\_initial$, tob_2000, home ownership, foreign-born, education category, age category in 1990, age category at time of initial entitlement, and age category at time of retirement. We include these as a help to those seeking to do basic comparisons between the synthetic and completed data.

7 Assessing Disclosure Risk

7.1 Overview

The link between administrative earnings, benefits data and SIPP data adds a significant amount of information to an already very detailed survey and could pose potential 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 prevent a link between these new public use files and the original SIPP public use files, which are already in the public domain. In addition, the synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information. Our disclosure avoidance research 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, which were used to create the SIPP component of our Gold Standard file.

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted two distinct matching exercises between the synthetic data and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, 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.

It is important to remember that for an actual re-identification of any of the records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required. This additional step consists of making another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc*. Hence, the results from our matching process are a very conservative estimation of re-identification risk.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. 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, the ratio of true matches to the total number of matches (true and false) should be close to one-half. We have performed two types of matching exercises, probabilistic and distance-based. This section describes the results from both exercises and gives an assessment of the risk of disclosure associated with the synthetic data files.

7.2 Matching based on probabilistic record linking

We begin with the probabilistic record linking experiment. Since the public use files consist of 16 different implicates, one must consider the risk associated with each file. In previous runs of this matching process, similar results were found on the different implicates. The evaluation of disclosure risk described here centers on the risk presented by the publication of one single implicate file (the first synthetic implicate that matches to the first missing data implicate, *i.e.* $m = 1$ and $r = 1$). In view of the results that are described below, we expect that similar results would be obtained for the other implicate files individually. We will, however, need to conduct research to evaluate the disclosure risk presented by the release of all 16 implicate files. In particular, we will evaluate the disclosure risk presented by the file obtained by averaging the variables across all the implicate files. The analysis of the averaged file is currently being conducted.

Probabilistic matching requires cashing a set of blocking and matching variables that are common to both files. We implemented one blocking strategy using the unsynthesized variables for blocking. For married individuals we use the unsynthesized variable $male$ for each member of the couples. For unmarried individuals we use the two unsynthesized variables, $male$ and $maritalstat$. The latter can be either widowed, divorced/separated, or never married ($maritalstat = \{2, 3, 4\}$). In other words, for two records to be a match, they must necessarily have identical values for marital status and gender since these two variables were not synthesized. After this has been determined to be the case, other variables can be compared to determine the probability that two records represent the same person.

The probabilistic record linking was performed using the Census Bureau's internal record linking software, which is maintained by the Statistical Research Division. The discussion in this section describes the technical settings used for that software. We set the blank filter flag equal to 0 so that if the variable is missing, the record will automatically be considered to agree on that field. Matching for the two groups, married and unmarried, was done separately. Blocking variables help to reduce the number of records used for comparison; however, in any given run all records in the same blocking group of the synthetic implicate and the Gold Standard files are compared. Thus, record linking computation is quadratic with run times dominated by the size of the largest block. In this latest version of the SIPP/SSA/IRS-PUF, the block sizes are very large. For this reason, the matching is done within corresponding segments of the Gold Standard and PUF files. Internally we know when segments of the Gold Standard and PUF files (single implicate) correspond to the same individuals, because we make use of the common artificial person identifier ($personid$) that is on both files. Without the information contained in $personid$ (which is not on the actual PUF), an intruder would have to compare many more record pairs to find true matches and would not find any more true matches (the true match is guaranteed to be in the blocks being compared) and would almost certainly find more false matches. For this reason our approach leads to a conservative measure of the disclosure risk.

When the SIPP/SSA/IRS-PUF is finally publicly released there will be no link between the Gold Standard data and the synthetic implicate files. However for testing purposes, we have maintained this link by keeping the common person identifier on the Gold Standard file and the PUF implicate files. Thus, by naming this person identifier in the sequence field of the record linking software, we can check which matched record pairs with a given score are correct matches and which are false matches by comparing this person identifier. When the person identifier is the same, the matching algorithm was successful in finding the person in the Gold Standard file to whom the synthetic data record belonged. When the person identifier is different, the matching algorithm was unsuccessful. This technology is also used for the distance matching discussed in section 7.3.

Automatic searches for matches occur only within those records sharing the same values on the blocking variables. Matches agree exactly on values for the blocking variables and, additionally, they agree on values for the matching variables. An input file to the matching software specifies the agreement criterion for each of the matching variables. Two numbers have to be specified for each of the matching variables. The first number represents the conditional probability that the two records agree on the matching field value given that the two records represent a match, called the m probability. The second number represents the conditional probability that the two records agree on the matching field value given that the two records do not represent a match, called the u probability. This technology was also used in creating the weight; see 5.8.2.

From the agreement criterion, the software computes a score. The agreement score for a match on a particular variable from two comparison records is based upon $\ln(m/u)$. A larger ratio implies a stronger distinguishing power for that matching field. Presumably, the ratio $m/u > 1$. When using Census Bureau matching software for the un-duplication of a file, one is trying to identify specific duplicate pairs, so more precise probability estimates may be helpful. However, when using this software for extracting subsets of plausible matches from a large file, the conditional agreement probabilities can be rough general estimates. To lean towards a more conservative assessment of disclosure risk, we obtained the best possible m and u estimates by using the $personid$ variable that is common between the files. Given that the public will not have access to this variable, an intruder trying to match the two files cannot possibly obtain better results using matching software that is at least as efficient as the Census Bureau software.

It is easy to calculate the conditional agreement probabilities $m = \Pr(agreement \mid match)$ for each matching field, if one knows when true matches occur. This is just the relative frequency of the fields on the Gold Standard and PUF files being equal, call this f_0. It is also easy to calculate the unconditional probability $\Pr(agreement)$ for each matching field that has a categorical variable. If, for example, X is a categorical variable that can take on 3 possible values, x_1, x_2, x_3 then we obtain the distributions of X in the Gold Standard (GS) and PUF files (implicate 1) and calculate

$$\Pr(agreement) = \sum_{i=1,2,3} \Pr(X = x_i \mid GS) \Pr(X = x_i \mid PUF).$$

Next it is clear that $\Pr(match) = \frac{1}{N}$, with N being the common size of both the GS and the PUF files, since for each GS record there is only one PUF record representing the same person. Therefore $\Pr(nonmatch) = \frac{N-1}{N}$, so given $m = \Pr(agreement \mid match) = f_0$, we have

$$\Pr(agreement) = \frac{f_0}{N} + \frac{\Pr(Agreement \mid nonmatch)(N-1)}{N}$$

and can solve for $u = \Pr(Agreement \mid nonmatch)$.

The agreement and disagreement conditional probabilities for those variables used for matching individuals with spouses are shown in Table 63. All matching fields were assigned the exact matching comparison type. This caused the program to assign full agreement/disagreement scores according to whether the fields agree or disagree. The corresponding agreement probabilities for single individuals are just slightly different and are shown in Table 64.

These probabilities are used to calculate the scores given to this variable when it agrees or disagrees. The agreement score is defined as $\ln(\frac{m}{u})$. The disagreement score is defined as $\ln(\frac{1-m}{1-u})$. For example, the full agreement score for a "c-match" on Hispanic is $\ln(\frac{0.888222038}{0.817697432}) \approx 0.08$. The disagreement score is $\ln(\frac{1-0.888222038}{1-0.817697432}) \approx -.50$.

The software compares each matching field, decides whether the field agrees or not, and then assigns the appropriate score to the field based on the user supplied m and u probabilities. Next, a cumulative match score is calculated by summing the scores across all the matching variables. This cumulative score is used to decide whether two records match. It is compared to the cutoff values provided by the user and if it passes the stated threshold, a match is declared. The influence of a one variable relative to another on this cumulative score is controlled by the relative matching and non-matching agreement probabilities specified by the user, but in this case based on actual calculations from the relevant files. The non-matching agreement probability essentially tells how often a field will agree at random across two files. A high value for this probability will reduce the importance of this variable in the matching by causing the agreement score to be lower. This is desirable because if the field is likely to agree at random, any match in values between two files is less likely to signify a true match. At the same time, a high non-matching agreement probability causes the disagreement score to be less negative or smaller, meaning that the penalty for not matching on this variable is not as high. In contrast, the relative matching agreement probability tells the importance of this variable compared to other variables in determining whether two records are a match. A high matching agreement probability means that a match on this field is crucial to determining an overall match between two records. Thus a high value for m produces a high agreement score. It also produces a more negative or higher disagreement score, more severely penalizing non-matching in this field. Consider the example of the variable $flag\_mar4t$, which is used to identify individuals who reported more than three marriages. When two records agree on this variable, and they are a match, the cumulative matching score increases by 5.317686217. If the records are not a match, but agree on this variable, then the cumulative score decreases by -4.609063992.

The output cutoff flag for the cumulative matching score provides the comparison points for the matching score. In our testing we declare any pair of records with a cumulative score between -20 and 20 to be a potential match. That is, we consider matching two records whenever their agreement score exceeds -20 even though most applications of probabilistic record linking use a positive cut-off for the automatic selection of potential matches. Thus, we declare records to be candidate matches based on an aggressive matching strategy. From either Table 63 or 64 we can see that the total matching scores cannot be outside of this range. Essentially, we allow every record in the synthetic file to have candidate matches in the Gold Standard. The output files are sorted by decreasing cumulative agreement score; then, the best two matches are kept. Finally, the proportions of true matches and the ratio of true to false matches are obtained.

The number and proportions of false and true matches, for each of the segments of the file, are given in Tables 65 and 66. The number of true and false matches in each segment are reported in column 3 and sum to equal the total number of records in each segment. The ratio of true to total matches and false to total matches gives the percentage of true matches and false matches in each segment and is reported in column 4. In Table 65, there are no data segments that have a true match rate over 1% and the ratio of true to false matches is extremely low. In Table 66, the percentages of true matches are slightly higher but the highest value is still just over 1% (1.18% for segment 2).

7.3 Distance matching

Distance-based record linking is another common approach to estimating the risk of disclosure in micro data. In recent work, 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) the more commonly used probabilistic methods. Their work suggests that re-identification exercises should also include distance based methods. The broader the selection of methods used, the more informed the analyst is of the risk of disclosure. In particular, it is important to understand which methods pose the largest threat. Domingo-Ferrer et al. (2006) conduct similar comparisons of distance-based and probabilistic record linking methods.

Our tests consider the case of an intruder who uses distance-based re-identification to match the source records from the Gold Standard to synthetic SIPP/SSA/IRS-PUF observations. Such re-identification methods calculate the distance between a given record in the Gold Standard and every record in the synthetic implicate. The j closest records are then declared potential candidates for a match to the source record. In our analysis we consider $j = 3$.

Our distance-based re-identification proceeds in two stages. First we split both the Gold Standard and the first synthetic implicate ($m = 1$ and $r = 1$) into groups based on the unsynthesized variables. In this case, marital status and $male$ are 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 files 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. This is the same assumption as we used in section 7.2 to segment the comparison files in that analysis. The segmentation of the blocks uses our prior knowledge of which records are 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 the true $personid$. After splitting the data into blocking groups and segments, we then calculate the distance between a given Gold Standard record and every record in the synthetic file in its corresponding blocking group and segment using a set of 163 matching variables. The three closest records are then declared possible matches.

We use four distance metrics. Each metric is a special case of either Mahalanobis or Euclidian distance. Before formally defining the distance, 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 163 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 163 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.

Tables 67-68 shows the results of the re-identification exercises for each of the four metrics. Table 67 shows the results using the Mahalanobis distance measures and Table 68 shows the results for the Euclidian distance measures. For each metric there are six columns. Match rate 1 (closest two records in A and B), match rate 2 (second closest two records in A and B), ratio of 2/1, match rate 3 (third closest two records in A and B), ratio of 3/2, and ratio (3+2)/1. Match rate j is calculated as the number of successful matches within a blocking group based on the jth closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages). For example, match rate 2 is calculated as the number of successful matches within a blocking group and segment based on the second closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages).

We first note that match rate 1 finds the highest rate of re-identifications. This implies that choosing the closest record using the indicated distance metric is more likely to find true match than choosing the second or third closest record. We further note that the highest match rate among all blocking groups is only 2.91%. Thus, an intruder who defined the closest- distance record as a match would correctly link 1.09% of records overall in the synthetic files and less than 3% in the worst-case sub-group.

The three ratio columns give us a sense of how much better the closest match does than the second and third best matches. Ideally, we want to ensure that if an intruder looked at the top three matches, he or she would face sufficient uncertainty about which one was the correct match. If the second closest record is exactly as likely to be the correct match as the closest record, then the ratio of match rate 2 to match rate 1 would be unity. If this ratio is less than one, then the closest record is more likely to be the correct match. If this ratio is greater than one, then the second closest record is more likely to be the correct match. The other ratio columns have the same interpretation. For the $MAHA1$ metric, the column Ratio (3+2)/1 ranges from 0.79 to 1.12. This suggests that the 2nd or 3rd closest matches are almost as likely to be correct as the closest match. The totals in the last row are essentially weighted averages of each column where the weights are the percentage of records in each group.

As expected, the $MAHA1$ metric produces the highest match rates. The highest match rate for the $MAHA2$ metric, perhaps the most likely to be used by an intruder, is 2.2% and the ratio of (3+2)/1 is very close to unity for every sub-group. The Euclidian metrics are very similar to the $MAHA2$ metrics with the overall match rate not exceeding 1.2%, the highest sub-group match rate less than 2.4%, and the ratio of (3+2)/1 generally being very close to or slightly higher than unity.

8 Using Synthetic Data

Many potential users may be concerned about how to begin using synthetic data and multiple implicate files. In this section we give some suggestions and advice for using these data sets to perform analyses and apply the formulae described in 6.

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 on the proper combining formulae. 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.

Each synthetic implicate has two variables that control the relationship between implicate files. The variable $m\_implicate$ tells which completed data implicate served as the starting basis for this particular synthetic data implicate. The variable $r\_implicate$ gives the synthetic implicate number for the file. There are four completed data implicates, so the variable $m\_implicate$ ranges from 1 to 4. There are four synthetic implicates per completed data implicate so the variable $r\_implicate$ ranges from 1 to 4 also. The first synthetic implicate will have $m\_implicate = 1$ and $r\_implicate = 1$, the second synthetic implicate will have $m\_implicate = 1$ and $r\_implicate = 2$, and so on, until the fifth synthetic implicate, which will have $m\_implicate = 2$ and $r\_implicate = 1$. In this manner a user can tell which synthetic implicates stemmed from the same completed data implicate. This information is necessary in order to apply the combining formulae.

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 average earnings in a given year, he or she should calculate the average in each of the 16 implicates using standard methods and then calculate the simple average these 16 separate means to get one grand mean. If the researcher wants to know the variance of earnings in a given year, he or she should follow the same procedure: calculate the variance in each implicate and then calculate the simple average these 16 statistics to get one grand variance. Note, and this is very important, the grand mean of the variances is just one component of the estimated total variance required to compute a confidence interval for average earnings. The complete formula is contained in section 6. 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. In addition to the statistic of interest, the user should save the estimated sampling variance of this statistic for each of the 16 implicates. For example, if calculating one mean per implicate, the user should calculate the sampling variance of the mean once per implicate.[18] The within-implicate sampling variances are then averaged to estimate the average within-implicate variance, one component of the total variance. The user must then make use of the $m\_implicate$ and $r\_implicate$ variables to calculate the between-completed-data-implicate variance and the between-synthetic-data-implicate variance according to the formula in 6.2.2. The user first calculates the variance of the statistic across the four r implicates associated with a particular m implicate. There will be four of these variances: one per completed data implicate. These four variances are then averaged to give the overall between-synthetic-data-implicate variance. The user then calculates the mean of the statistic of interest for all the synthetic implicates associated with a particular completed data implicate. Again, there will be four of these means. The between m implicate variance is then calculated as the average of the squared deviations of these four means from the overall grand mean. If the statistics

[18]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.

of interest are saved in a data set, these calculations can be easily performed. The variance pieces are then combined to create the total variance and calculate degrees of freedom. 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 result 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 the manner described above in order to create a comparison basis.

9 Conclusion

Given the length and scope of this project, it is perhaps beneficial at this point to consider what has been accomplished. This collaboration between four government agencies has produced several new data products and advanced the body of knowledge on missing data imputation, assessing the validity of automated statistical modeling, disclosure avoidance techniques, and disclosure risk analysis. In the past six years, we have produced a highly useful compilation of SIPP data that combines five separate panels with edited administrative data from IRS and SSA, a weight to allow meaningful analysis of these combined panels, a set of files that multiply impute all missing data, and a set of synthetic data files that meet disclosure standards of the Census Bureau, the Internal Revenue Service, and the Social Security Administration. For the first time in 30 years, it appears that it will be possible to release lifetime earnings histories taken from administrative records, an accomplishment that will be of enormous benefit to the research community and the general population. This project has been a model for what inter-agency cooperation can accomplish by pooling the expertise of researchers from the Census Bureau, IRS, SSA, and CBO.

When we began this project, there was a great deal of uncertainty over whether synthesizing techniques could produce micro-data that would preserve relationships among variables and mitigate disclosure risk. In fact, almost none of the enhanced theory or experience with these methods required to complete the project existed. Based on the results at this point, we feel that both these questions can be answered in the affirmative. It is now imperative that outside users be given a chance to test these synthetic data and that the agencies involved develop a system for validating outside results using the Gold Standard in order to promote general confidence in the methods and to permit quality improvements. This process will help us to discover remaining flaws in the synthetic data and improve the synthesizing process, both of which will enable the collaborators to provide useful future updates to this data product, as funding resources permit.

Bibliography

Abowd, John M. and Simon D. Woodcock. 2001. Disclosure Limitation in Longitudinal Linked Data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, ed. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North Holland pp. 215–277.

Albert, A. and J.A. Anderson. 1984. "On The Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71:1–10.

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

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

Fellegi, Ivan P. and Alan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64:1183–1210.

Gelman, A. B., J. S. Carlin, H. S. Stern and D. B. Rubin. 2000. *Bayesian Data Analysis*. Chapman and Hall.

Little, Roderick J.A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9(2):407–426.

Minka, T. P. 2003. Bayesian inference, entropy, and the multinomial distribution. Technical report Microsoft, Inc. available at http://research.microsoft.com/ minka/papers/minka-multinomial.pdf (October 30, 2006).

Raghunathan, T.E., J.P. Reiter and D.B. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19(1):1–16.

Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk and Peter Solenberger. 1998. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." Survey Research Center, University of Michigan.

Reiter, Jerome P. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology* 29:181–188.

Reiter, Jerry P. 2004. "Simultaneous use of multiple imputation for missing data and disclosure limitation." *Survey Methodology* 30:235–242.

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *The Annals of Statistics* 9:130–134.

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

Rubin, Donald B. 1993. "Discussion of Statistical Disclosure Limitation." *Journal of Official Statistics* 9(2):461–468.

Rubin, Donald B. 1996. "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91(434):473–489.

Tanner, Martin A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributiosn and Likelihood Functions*. Third ed. Springer-Verlag.

Woodcock, Simon D. and Gary Benedetto. 2006. Distribution-preserving statistical disclosure limitation. LEHD technical paper tp-2006-04 U.S. Census Bureau. http://lehd.dsd.census.gov/led/library/techpapers/tp-2006-04.pdf (October 31, 2006).

A Appendix

The appendices to this report have been included as separate files. The list is provided here.

- Technical Description of the Creation of the SIPP/SSA/IRS Gold Standard Files and the SIPP-SSA-IRS PUF Version 4.0 (PDF file)
- Appendix Table A1 Detailed Variable Information (PDF file)
- Varlists_description_version_4.0.xls (Excel workbook)
- CompletionSynthesisTables1.xls (table 1) - Completion and Synthesis (Excel workbook)
- WeightsTables1.xls (table 2) - Weights (Excel workbook)
- AnalyticValidityTables1.xls (tables 3-18) - Analytical Validity (Excel workbook)
- AnalyticValidityTables2.xls (tables 19-28) - Analytical Validity (Excel workbook)
- AnalyticValidityTables3.xls (charts 1-8) - Analytical Validity (Excel workbook)
- AnalyticValidityTables4.xls (tables 29-34) - Analytical Validity (Excel workbook)
- AnalyticValidityTables5.xls (tables 35-37) - Analytical Validity (Excel workbook)
- AnalyticValidityTables6.xls (table 38) - Analytical Validity (Excel workbook)
- AnalyticValidityTables7.xls (table 39, charts 9-10) - Analytical Validity (Excel workbook)
- AnalyticValidityTables8.xls (tables 40-60) - Analytical Validity (Excel workbook)
- AnalyticValidityTables9.xls (table 61) - Analytical Validity (Excel workbook)
- AnalyticValidityTables10.xls (table 62) - Analytical Validity (Excel workbook)
- DisclosureTestingTables1.xls (tables 63-66) - Disclosure Testing (Excel workbook)
- DisclsoureTestingTables2.xls (table 67-68) - Disclosure Testing (Excel workbook)

Table 1: Small Cell Consequences of Selected Combinations of Non-synthesized Variables

1) Current: V3.0	2) Current0 Add MBA initial0	3) Current0 Add TOB initial0	4) Drop 6 Current0 Add MBA,TOB initial	5) Drop 4 Current Add TOB initial0	6) Drop 6 Current Add TOB initial0	7) Drop 6 Current0 Add TOB initial and 20
blackh maleh educ_3cath maritalstath age_cath black_spouseh male_spouseh educ_3cat_spouse age_cat_spouseh - - - -	blackh maleh educ_3cath maritalstath age_cath black_spouseh male_spouseh educ_3cat_spouse age_cat_spouseh MBA initialh MBA initial spouse - -	black maleh educ_3cath maritalstath age_cath black_spouseh male_spouseh educ_3cat_spouse age_cat_spouseh - - tob_initialh tob_initial_spouseh	- maleh - maritalstath - - male_spouseh - - MBA initialh MBA initial spouseh tob_initial tob_initial_spouse	- maleh educ_3cath maritalstath - - male_spouseh educ_3cat_spouse - - - tob_initialh tob_initial_spouseh	- maleh - maritalstath - - male_spouseh - - - - tob_initialh tob_initial_spouseh	- maleh - maritalstath - - male_spouseh - - - - tob_initialh tob_initial_spouseh tob_2000h tob_2000_spouseh
SMALL CELL ANALYSIS AT PERSON-LEVEL0						
Number of small person-level cells (<10 individuals):0						
DBV: 159h **EBV**: 159h	**DBV**: 16048h **EBV**: 17240h	**DBV**: 1952h **EBV**: 1967h	**DBV**: 8092h **EBV**: 7952h	**DBV**: 196h **EBV**: 194h	**DBV**: 26h **EBV**: 26h	**DBV:** 129h **EBV:** 133h
Number of individuals in small person-level cells:0						
DBV: 472h **EBV:** 472h	**DBV:** 37611h **EBV:** 38191h	**DBV:** 5448h **EBV:** 5387h	**DBV:** 20251h **EBV:** 19920h	**DBV:** 672h **EBV:** 648h	**DBV:** 76h **EBV:** 72h	**DBV:** 373h **EBV:** 373h
SMALL CELL ANALYSIS AT HOUSEHOLD-LEVEL0						
Number of small household-level cells (<10 households):0						
DBV: 93h **EBV**: 93h	**DBV**: 8724h **EBV**: 9333h	**DBV**: 1074h **EBV**: 1081h	**DBV**: 4238h **EBV**: 4171h	**DBV**: 113h **EBV**: 112h	**DBV**: 19h **EBV**: 19h	**DBV**: 85h **EBV**: 86h
Number of households in small household-level cells:0						
DBV: 280h **EBV:** 280h	**DBV:** 21143h **EBV:** 21437h	**DBV:** 3048h **EBV:** 3020h	**DBV:** 10807h **EBV:** 10650h	**DBV:** 379h **EBV:** 367h	**DBV:** 54h **EBV:** 52h	**DBV:** 246h **EBV:** 245h
MBA initial and MBA initial spouse are rounded to nearest $50h tob_initial, tob_initial_spouse, tob_2000, and tob_2000_spouse are put in categories of 1,2,3,5, and otherh						

Table 2: Analytic Validity of SIPP-PUF Weights -- Weighted Counts of Benefit ecipients

	SSA eportsR	Average across completed data using completed weightsR	Average across synthetic data using synthetic weightsR	Percentage difference between columns C and DR
Number of retired workers receiving benefits in Dec. 2000 (in millions)R	28.50	27.10	26.40	-2.58
Average monthly benefit for retired workersR	845.00	820.00	824.00	0.49
Number of widows and widowers receiving benefits in Dec. 2000 (in millions)R	4.70	4.50	4.30	-4.44
Average monthly benefit for widows and widowersR	810.00	752.00	753.00	0.13
Number of disabled receiving benefits in Dec. 2000 (in millions)R	5.00	5.90	5.90	0.00
Average monthly benefit for disabledR	786.00	736.00	738.00	0.27
Number of permanently insured individuals in Dec. 2000 (in millions)R	140.70	131.40	133.70	1.75
DER average earnings for 2000R	31,213.00	33,331.00	34,751.00	4.26
Number of wage and salary workers w/taxable earnings for 2000 (in millions)R	145.00	128.00	129.00	0.78
SER average earnings for 2000R	26,081.00	27,360.00	28,196.00	3.06

Table 3: SER work indicator for year 1965

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM SyntheticM	CompletedM
white femalesM	own retirementM	0.567M	0.563M	0.562M	0.572M	0.557M	0.569M	0M	0.00001M	0.00001M
	disabilityM	0.374M	0.374M	0.361M	0.386M	0.360M	0.388M	0M	0.00005M	0.00007M
	aged spouseM	0.133M	0.130M	0.122M	0.144M	0.114M	0.147M	0M	0.00004M	0.00008M
	aged widowM	0.225M	0.210M	0.216M	0.233M	0.200M	0.219M	0M	0.00002M	0.00003M
	otherM	0.057M	0.052M	0.053M	0.062M	0.048M	0.056M	0M	0.00001M	0.00001M
black femalesM	own retirementM	0.658M	0.693M	0.642M	0.674M	0.668M	0.717M	0M	0.00009M	0.00019M
	disabilityM	0.347M	0.317M	0.316M	0.378M	0.291M	0.344M	0M	0.00030M	0.00024M
	aged spouseM	0.255M	0.225M	0.209M	0.301M	0.185M	0.264M	0M	0.00066M	0.00057M
	aged widowM	0.300M	0.359M	0.262M	0.339M	0.313M	0.405M	0M	0.00047M	0.00069M
	otherM	0.044M	0.046M	0.034M	0.054M	0.036M	0.056M	0M	0.00003M	0.00003M
white malesM	own retirementM	0.895M	0.895M	0.889M	0.900M	0.891M	0.898M	0M	0.00001M	0.00000M
	disabilityM	0.563M	0.556M	0.548M	0.579M	0.542M	0.570M	0M	0.00007M	0.00007M
	aged spouseM	0.161M	0.166M	0.070M	0.253M	0.093M	0.240M	0M	0.00260M	0.00191M
	aged widowM	0.414M	0.456M	0.292M	0.535M	0.334M	0.579M	0M	0.00510M	0.00550M
	otherM	0.009M	0.007M	0.007M	0.011M	0.005M	0.009M	0M	0.00000M	0.00000M
black malesM	own retirementM	0.858M	0.865M	0.847M	0.868M	0.852M	0.879M	0M	0.00004M	0.00007M
	disabilityM	0.465M	0.448M	0.435M	0.496M	0.417M	0.479M	0M	0.00028M	0.00032M
	aged spouseM	0.202M	0.079M	0.068M	0.335M	-0.027M	0.184M	0M	0.00291M	0.00408M
	aged widowM	0.296M	0.253M	0.110M	0.483M	0.051M	0.456M	0M	0.01189M	0.01462M
	otherM	0.006M	0.003M	0.002M	0.009M	0.000M	0.005M	0M	0.00000M	0.00000M

Table 4: SER earnings in year 1965

DemographicM	Type of	eanM		Confidence IntervalM		Confidence IntervalM		SyntheticM	Total VarianceM	
GroupM	BenefitM	SyntheticM	CompletedM	SyntheticM		CompletedM		DF Not ExistM	SyntheticM	CompletedM
white femalesM	own retirementM	1,461.89M	1,523.24M	1,456.39M	1,467.40M	1,496.23M	1,550.25M	1M	10.49M	252.06M
	disabilityM	701.12M	731.24M	675.21M	727.04M	693.82M	768.65M	0M	242.88M	484.23M
	aged spouseM	148.21M	164.41M	123.88M	172.55M	139.32M	189.51M	0M	175.13M	204.31M
	aged widowM	350.18M	323.48M	339.50M	360.85M	303.76M	343.20M	1M	39.45M	143.18M
	otherM	70.24M	71.31M	62.63M	77.85M	62.39M	80.24M	0M	20.30M	28.00M
black femalesM	own retirementM	1,370.90M	1,522.66M	1,234.12M	1,507.69M	1,442.74M	1,602.57M	0M	3,937.78M	2,091.20M
	disabilityM	486.02M	483.09M	410.19M	561.85M	435.22M	530.96M	0M	1,287.51M	837.98M
	aged spouseM	279.02M	296.24M	166.64M	391.41M	228.57M	363.91M	0M	3,385.90M	1,691.88M
	aged widowM	373.80M	451.48M	278.31M	469.28M	348.49M	554.48M	0M	1,953.18M	3,319.55M
	otherM	40.04M	37.21M	27.44M	52.65M	25.58M	48.84M	0M	53.31M	47.73M
white malesM	own retirementM	3,678.21M	3,840.81M	3,655.68M	3,700.75M	3,821.64M	3,859.97M	0M	122.39M	135.54M
	disabilityM	1,815.13M	1,843.00M	1,764.99M	1,865.28M	1,801.70M	1,884.29M	0M	796.70M	627.67M
	aged spouseM	364.09M	462.62M	128.70M	599.49M	214.68M	710.56M	0M	17,818.49M	21,517.48M
	aged widowM	1,075.53M	745.42M	694.18M	1,456.89M	194.92M	1,295.91M	0M	52,089.95M	93,876.78M
	otherM	8.78M	8.77M	5.09M	12.46M	5.57M	11.98M	0M	4.63M	3.74M
black malesM	own retirementM	3,019.57M	3,067.00M	2,912.03M	3,127.10M	2,982.02M	3,151.98M	1M	4,001.08M	2,491.36M
	disabilityM	1,254.48M	1,205.29M	1,113.37M	1,395.59M	1,066.37M	1,344.21M	0M	3,927.03M	5,661.00M
	aged spouseM	341.05M	115.63M	26.88M	655.23M	-38.88M	270.15M	0M	23,592.61M	8,822.80M
	aged widowM	698.37M	422.35M	-361.61M	1,758.35M	15.54M	829.16M	0M	332,083.19M	59,957.88M
	otherM	3.45M	1.25M	-0.02M	6.92M	-1.44M	3.93M	0M	4.41M	2.64M

Table 5: SER work indicator for year 1975

DemographicM GroupM	Type of BenefitM	eanM		Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM	
		SyntheticM	CompletedM						SyntheticM	CompletedM
white femalesM	own retirementM	0.675M	0.690M	0.666M	0.684M	0.682M	0.698M	0M	0.00002M	0.00002M
	disabilityM	0.554M	0.548M	0.532M	0.576M	0.530M	0.566M	0M	0.00012M	0.00010M
	aged spouseM	0.158M	0.140M	0.148M	0.167M	0.130M	0.151M	0M	0.00003M	0.00004M
	aged widowM	0.250M	0.242M	0.241M	0.259M	0.225M	0.260M	0M	0.00003M	0.00008M
	otherM	0.217M	0.211M	0.208M	0.225M	0.204M	0.219M	0M	0.00003M	0.00002M
black femalesM	own retirementM	0.722M	0.742M	0.704M	0.739M	0.722M	0.762M	0M	0.00008M	0.00014M
	disabilityM	0.573M	0.599M	0.554M	0.591M	0.577M	0.621M	0M	0.00012M	0.00018M
	aged spouseM	0.268M	0.297M	0.231M	0.306M	0.248M	0.346M	0M	0.00047M	0.00086M
	aged widowM	0.303M	0.303M	0.273M	0.333M	0.255M	0.352M	0M	0.00032M	0.00074M
	otherM	0.179M	0.189M	0.166M	0.191M	0.172M	0.205M	0M	0.00005M	0.00010M
white malesM	own retirementM	0.865M	0.879M	0.860M	0.870M	0.875M	0.883M	0M	0.00001M	0.00001M
	disabilityM	0.705M	0.706M	0.696M	0.713M	0.697M	0.716M	0M	0.00003M	0.00003M
	aged spouseM	0.135M	0.106M	0.076M	0.193M	0.051M	0.160M	0M	0.00124M	0.00109M
	aged widowM	0.366M	0.392M	0.238M	0.494M	0.244M	0.539M	0M	0.00548M	0.00733M
	otherM	0.194M	0.191M	0.183M	0.205M	0.182M	0.201M	0M	0.00004M	0.00003M
black malesM	own retirementM	0.790M	0.814M	0.764M	0.815M	0.794M	0.833M	0M	0.00015M	0.00013M
	disabilityM	0.648M	0.661M	0.626M	0.669M	0.636M	0.686M	0M	0.00015M	0.00022M
	aged spouseM	0.136M	0.309M	-0.042M	0.314M	-0.139M	0.757M	0M	0.00841M	0.04861M
	aged widowM	0.252M	0.000M	0.035M	0.469M			0M	0.01382M	0.00000M
	otherM	0.159M	0.147M	0.130M	0.187M	0.128M	0.165M	0M	0.00020M	0.00012M

Table 6: SER Earnings for year 1975

DemographicM	Type of	eanM		Confidence IntervalM		Confidence IntervalM		SyntheticM	Total VarianceM	
GroupM	BenefitM	SyntheticM	CompletedM	SyntheticM		CompletedM		DF Not ExistM	SyntheticM	CompletedM
white femalesM	own retirementM	3,948M	4,144M	3,655M	4,242M	4,078M	4,211M	0M	10,404M	1,493M
	disabilityM	2,274M	2,279M	2,137M	2,411M	2,168M	2,391M	0M	4,464M	3,918M
	aged spouseM	359M	315M	300M	418M	280M	350M	0M	1,016M	451M
	aged widowM	854M	830M	798M	911M	760M	900M	0M	1,022M	1,494M
	otherM	543M	541M	462M	625M	511M	571M	0M	1,245M	327M
black femalesM	own retirementM	3,878M	4,308M	3,635M	4,121M	4,158M	4,458M	0M	13,942M	8,311M
	disabilityM	2,211M	2,435M	1,970M	2,452M	2,287M	2,582M	0M	14,500M	7,967M
	aged spouseM	581M	628M	404M	759M	462M	794M	0M	10,334M	9,718M
	aged widowM	1,033M	1,040M	889M	1,177M	814M	1,265M	0M	7,260M	16,326M
	otherM	419M	459M	317M	520M	392M	526M	0M	2,787M	1,578M
white malesM	own retirementM	9,241M	9,836M	9,197M	9,285M	9,770M	9,902M	1M	663M	1,564M
	disabilityM	5,446M	5,574M	5,329M	5,563M	5,465M	5,683M	1M	4,730M	4,371M
	aged spouseM	340M	180M	131M	548M	72M	288M	0M	15,455M	4,330M
	aged widowM	3,032M	2,644M	1,640M	4,425M	1,161M	4,127M	0M	643,749M	747,378M
	otherM	630M	589M	570M	690M	546M	631M	0M	586M	661M
black malesM	own retirementM	6,990M	7,500M	6,782M	7,199M	7,296M	7,704M	1M	15,086M	15,303M
	disabilityM	3,934M	4,037M	3,691M	4,176M	3,768M	4,306M	0M	18,273M	24,563M
	aged spouseM	401M	164M	-263M	1,064M	-92M	420M	0M	117,515M	18,317M
	aged widowM	1,602M	0M	-970M	4,174M			0M	2,061,954M	0M
	otherM	367M	339M	262M	471M	271M	407M	0M	3,056M	1,648M

Table 7: SER work indicator for year 1985

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM Synthetic	CompletedM
white femalesM	own retirementM	0.547M	0.557M	0.531M	0.563M	0.539M	0.574M	0M	0.00006M	0.00007M
	disabilityM	0.594M	0.597M	0.574M	0.614M	0.579M	0.615M	0M	0.00011M	0.00010M
	aged spouseM	0.150M	0.159M	0.116M	0.183M	0.129M	0.189M	0M	0.00025M	0.00021M
	aged widowM	0.210M	0.209M	0.194M	0.226M	0.192M	0.226M	0M	0.00007M	0.00008M
	otherM	0.419M	0.429M	0.407M	0.432M	0.418M	0.440M	0M	0.00005M	0.00004M
black femalesM	own retirementM	0.572M	0.558M	0.524M	0.619M	0.533M	0.583M	0M	0.00051M	0.00020M
	disabilityM	0.605M	0.593M	0.568M	0.641M	0.544M	0.643M	0M	0.00037M	0.00064M
	aged spouseM	0.165M	0.133M	0.118M	0.213M	0.082M	0.183M	0M	0.00068M	0.00079M
	aged widowM	0.224M	0.222M	0.189M	0.259M	0.190M	0.254M	0M	0.00040M	0.00037M
	otherM	0.366M	0.353M	0.344M	0.387M	0.334M	0.372M	0M	0.00016M	0.00014M
white malesM	own retirementM	0.673M	0.679M	0.660M	0.686M	0.668M	0.690M	0M	0.00004M	0.00003M
	disabilityM	0.634M	0.640M	0.621M	0.646M	0.625M	0.656M	0M	0.00005M	0.00007M
	aged spouseM	0.189M	0.206M	0.104M	0.274M	0.137M	0.275M	0M	0.00245M	0.00177M
	aged widowM	0.301M	0.293M	0.151M	0.451M	0.150M	0.437M	0M	0.00704M	0.00680M
	otherM	0.459M	0.466M	0.441M	0.478M	0.452M	0.480M	0M	0.00009M	0.00007M
black malesM	own retirementM	0.630M	0.607M	0.597M	0.663M	0.556M	0.658M	0M	0.00023M	0.00062M
	disabilityM	0.593M	0.572M	0.566M	0.621M	0.548M	0.595M	0M	0.00024M	0.00020M
	aged spouseM	0.075M	0.071M	-0.040M	0.189M	-0.081M	0.222M	0M	0.00438M	0.00713M
	aged widowM	0.250M	0.101M	0.051M	0.450M	-0.023M	0.226M	0M	0.01399M	0.00570M
	otherM	0.380M	0.387M	0.346M	0.413M	0.358M	0.417M	0M	0.00036M	0.00030M

Table 8: SER Earnings for year 1985

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM SyntheticM	CompletedM
white femalesM	own retirementM	7,101M	7,200M	6,945M	7,258M	6,927M	7,472M	1M	8,438M	18,878M
	disabilityM	6,069M	5,859M	5,636M	6,502M	5,673M	6,045M	0M	30,118M	12,792M
	aged spouseM	960M	892M	565M	1,355M	501M	1,282M	0M	29,660M	31,332M
	aged widowM	1,842M	1,724M	1,661M	2,023M	1,510M	1,938M	0M	8,060M	12,058M
	otherM	3,656M	3,582M	3,527M	3,784M	3,375M	3,790M	1M	5,725M	12,547M
black femalesM	own retirementM	6,915M	7,057M	6,116M	7,714M	6,657M	7,457M	0M	131,457M	55,966M
	disabilityM	5,655M	5,796M	5,311M	6,000M	5,277M	6,314M	0M	31,031M	84,628M
	aged spouseM	985M	903M	638M	1,333M	457M	1,348M	0M	39,258M	60,710M
	aged widowM	1,694M	1,692M	1,391M	1,996M	1,349M	2,035M	0M	32,491M	43,063M
	otherM	2,665M	2,576M	2,434M	2,897M	2,352M	2,800M	0M	17,039M	18,195M
white malesM	own retirementM	15,254M	16,091M	15,007M	15,500M	15,705M	16,477M	1M	20,992M	39,152M
	disabilityM	10,186M	10,573M	9,655M	10,718M	10,140M	11,005M	0M	67,302M	54,135M
	aged spouseM	1,385M	1,420M	501M	2,269M	506M	2,334M	0M	266,884M	275,395M
	aged widowM	4,318M	4,981M	1,339M	7,296M	1,119M	8,843M	0M	2,880,355M	4,575,582M
	otherM	5,966M	6,003M	5,724M	6,208M	5,742M	6,264M	0M	16,686M	23,163M
black malesM	own retirementM	11,381M	11,409M	10,926M	11,836M	10,155M	12,664M	1M	71,721M	386,020M
	disabilityM	7,870M	7,563M	7,127M	8,613M	7,093M	8,033M	0M	140,174M	81,435M
	aged spouseM	368M	907M	-387M	1,123M	-656M	2,470M	0M	206,113M	871,879M
	aged widowM	1,647M	524M	-755M	4,049M	-176M	1,223M	0M	1,728,091M	178,634M
	otherM	3,821M	3,582M	3,360M	4,283M	3,194M	3,970M	0M	69,510M	53,807M

Table 9: SER work indicator for year 1995

DemographicM GroupM	Type of BenefitM	eanM		Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM	
		SyntheticM	CompletedM						Synthetic	CompletedM
white femalesM	own retirementM	0.338M	0.343M	0.329M	0.347M	0.336M	0.349M	0M	0.00002M	0.00002M
	disabilityM	0.452M	0.452M	0.441M	0.463M	0.440M	0.463M	0M	0.00004M	0.00005M
	aged spouseM	0.109M	0.112M	0.079M	0.140M	0.080M	0.144M	0M	0.00020M	0.00022M
	aged widowM	0.157M	0.154M	0.149M	0.165M	0.145M	0.162M	0M	0.00002M	0.00003M
	otherM	0.606M	0.613M	0.593M	0.619M	0.598M	0.627M	0M	0.00005M	0.00006M
black femalesM	own retirementM	0.373M	0.362M	0.356M	0.391M	0.343M	0.382M	0M	0.00010M	0.00014M
	disabilityM	0.460M	0.456M	0.417M	0.503M	0.429M	0.482M	0M	0.00046M	0.00025M
	aged spouseM	0.122M	0.107M	0.060M	0.183M	0.066M	0.147M	0M	0.00097M	0.00054M
	aged widowM	0.173M	0.186M	0.138M	0.207M	0.148M	0.223M	0M	0.00036M	0.00045M
	otherM	0.597M	0.576M	0.564M	0.631M	0.552M	0.600M	0M	0.00029M	0.00020M
white malesM	own retirementM	0.422M	0.444M	0.413M	0.431M	0.436M	0.451M	0M	0.00002M	0.00002M
	disabilityM	0.439M	0.457M	0.409M	0.469M	0.434M	0.481M	0M	0.00020M	0.00014M
	aged spouseM	0.095M	0.107M	-0.006M	0.196M	0.023M	0.190M	0M	0.00267M	0.00214M
	aged widowM	0.273M	0.284M	0.179M	0.366M	0.164M	0.403M	0M	0.00320M	0.00507M
	otherM	0.701M	0.691M	0.678M	0.724M	0.670M	0.712M	0M	0.00013M	0.00012M
black malesM	own retirementM	0.435M	0.413M	0.388M	0.481M	0.393M	0.433M	0M	0.00032M	0.00015M
	disabilityM	0.443M	0.450M	0.422M	0.463M	0.425M	0.475M	0M	0.00015M	0.00023M
	aged spouseM	0.159M	0.233M	-0.094M	0.412M	-0.181M	0.646M	0M	0.01659M	0.04149M
	aged widowM	0.147M	0.235M	-0.016M	0.310M	0.052M	0.418M	1M	0.00919M	0.01221M
	otherM	0.628M	0.618M	0.597M	0.659M	0.584M	0.652M	0M	0.00013M	0.00037M

Table 10: SER Earnings for year 1995

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM SyntheticM	CompletedM
white femalesM	own retirementM	5,495M	5,566M	5,423M	5,566M	5,421M	5,710M	1M	1,764M	7,686M
	disabilityM	6,135M	6,084M	5,911M	6,359M	5,757M	6,411M	0M	12,002M	35,478M
	aged spouseM	1,106M	1,095M	678M	1,533M	466M	1,724M	0M	40,062M	80,651M
	aged widowM	1,849M	1,760M	1,732M	1,966M	1,566M	1,954M	0M	4,632M	11,808M
	otherM	9,044M	8,969M	8,808M	9,281M	8,597M	9,342M	1M	19,351M	42,286M
black femalesM	own retirementM	5,630M	5,501M	5,183M	6,078M	5,028M	5,973M	0M	55,905M	79,479M
	disabilityM	5,590M	6,032M	4,840M	6,341M	5,383M	6,681M	0M	150,415M	137,388M
	aged spouseM	1,258M	1,108M	428M	2,088M	605M	1,611M	0M	186,677M	87,164M
	aged widowM	2,003M	1,975M	1,526M	2,480M	1,488M	2,463M	0M	78,477M	87,268M
	otherM	7,341M	6,706M	6,903M	7,779M	6,275M	7,138M	0M	66,509M	67,843M
white malesM	own retirementM	10,347M	11,012M	10,003M	10,691M	10,751M	11,274M	0M	21,142M	23,637M
	disabilityM	8,756M	9,173M	7,735M	9,776M	8,271M	10,075M	0M	213,915M	187,909M
	aged spouseM	811M	951M	-520M	2,142M	-427M	2,328M	0M	480,470M	520,068M
	aged widowM	3,971M	3,855M	1,017M	6,924M	1,170M	6,540M	0M	3,007,166M	2,639,770M
	otherM	15,624M	15,661M	14,583M	16,664M	15,105M	16,216M	0M	224,431M	98,502M
black malesM	own retirementM	8,856M	8,564M	7,853M	9,858M	7,883M	9,245M	1M	347,683M	163,956M
	disabilityM	6,569M	5,915M	5,751M	7,387M	5,388M	6,441M	0M	175,129M	101,723M
	aged spouseM	1,691M	3,838M	-2,110M	5,492M	-5,596M	13,272M	0M	3,074,159M	20,071,412M
	aged widowM	1,012M	730M	-1,743M	3,767M	-2,054M	3,513M	0M	2,685,640M	1,947,740M
	otherM	9,757M	9,309M	9,093M	10,421M	8,614M	10,003M	0M	153,850M	176,484M

Table 11: Total SER earnings 1951-2003

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM SyntheticM	CompletedM
white femalesM	own retirementM	192,468M	198,303M	189,034M	195,902M	195,018M	201,589M	0M	3,635,902M	3,582,412M
	disabilityM	159,975M	160,721M	155,261M	164,689M	153,560M	167,882M	0M	7,290,556M	14,186,930M
	aged spouseM	31,945M	32,601M	20,290M	43,600M	18,207M	46,996M	0M	27,166,130M	40,572,680M
	aged widowM	52,794M	51,821M	49,392M	56,195M	47,184M	56,459M	0M	3,602,536M	5,853,323M
	otherM	187,463M	187,956M	182,067M	192,860M	180,912M	194,999M	0M	9,256,167M	14,299,967M
black femalesM	own retirementM	191,274M	195,617M	180,979M	201,570M	187,629M	203,605M	0M	27,671,924M	23,120,206M
	disabilityM	145,353M	151,265M	137,949M	152,757M	140,394M	162,136M	0M	18,055,260M	35,532,660M
	aged spouseM	36,723M	36,296M	25,665M	47,782M	25,237M	47,356M	0M	36,155,536M	39,974,037M
	aged widowM	56,721M	57,379M	48,964M	64,478M	48,166M	66,592M	0M	22,018,177M	31,222,865M
	otherM	152,606M	146,146M	144,237M	160,975M	138,596M	153,697M	0M	23,731,960M	20,555,483M
white malesM	own retirementM	417,976M	442,503M	413,684M	422,268M	438,563M	446,442M	0M	5,837,279M	5,662,728M
	disabilityM	276,091M	288,266M	254,564M	297,618M	268,521M	308,011M	0M	94,001,775M	82,880,905M
	aged spouseM	33,447M	32,596M	9,986M	56,908M	12,442M	52,749M	0M	143,099,719M	111,939,873M
	aged widowM	126,429M	134,014M	67,633M	185,225M	71,111M	196,917M	0M	1,227,132,456	1,441,472,756M
	otherM	315,302M	319,194M	300,010M	330,593M	306,393M	331,994M	0M	59,030,061M	45,981,602M
black malesM	own retirementM	330,958M	331,280M	311,262M	350,654M	317,708M	344,852M	1M	134,237,271M	63,661,959M
	disabilityM	204,902M	197,208M	186,983M	222,821M	185,899M	208,516M	0M	82,954,046M	43,726,354M
	aged spouseM	48,022M	66,377M	-25,930M	121,974M	-46,152M	178,906M	0M	1,289,053,431	2,882,382,428M
	aged widowM	56,265M	29,515M	5,535M	106,994M	-2,984M	62,014M	0M	841,272,477M	309,581,707M
	otherM	200,003M	194,732M	186,450M	213,556M	182,929M	206,534M	0M	63,536,541M	51,252,661M

Table 12: Total years worked in SER (i.e. positive FICA earnings)

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed						Synthetic	Completed
white females	own retirement	26.174	26.693	25.881	26.466	26.448	26.939	0	0.02243	0.01731
	disability	21.679	22.076	21.387	21.972	21.776	22.376	0	0.02815	0.02983
	aged spouse	8.047	8.099	7.189	8.904	7.172	9.026	0	0.15682	0.18089
	aged widow	10.614	10.353	10.349	10.879	10.096	10.609	0	0.02540	0.02425
	other	15.050	15.459	14.771	15.328	15.208	15.710	0	0.02286	0.01998
black females	own retirement	27.847	28.428	26.900	28.794	27.931	28.925	0	0.21083	0.08429
	disability	20.915	21.431	20.094	21.735	20.653	22.208	0	0.17740	0.17340
	aged spouse	10.317	9.974	8.972	11.663	8.792	11.156	0	0.55231	0.47391
	aged widow	12.594	13.320	11.495	13.694	12.293	14.346	0	0.38192	0.36726
	other	13.800	13.947	13.466	14.134	13.532	14.362	0	0.04064	0.06085
white males	own retirement	35.779	36.477	35.638	35.920	36.346	36.609	0	0.00654	0.00632
	disability	26.184	26.610	25.517	26.851	26.094	27.127	0	0.10068	0.06774
	aged spouse	8.108	8.506	6.575	9.641	6.615	10.398	0	0.86016	1.26191
	aged widow	15.958	15.778	11.908	20.008	11.962	19.593	0	5.75307	5.37211
	other	15.907	16.243	15.403	16.410	15.844	16.642	0	0.06124	0.04372
black males	own retirement	33.902	33.791	33.230	34.574	33.284	34.298	1	0.15613	0.09151
	disability	23.579	23.571	23.040	24.119	22.771	24.371	0	0.10190	0.19847
	aged spouse	10.429	9.296	7.422	13.437	4.335	14.257	0	1.19750	5.85639
	aged widow	14.820	10.428	7.940	21.701	6.120	14.735	0	15.52250	6.22151
	other	13.783	13.979	13.202	14.365	13.505	14.452	0	0.11294	0.08209

Table 13: Personal Account: 2% of earnings compounded annually at 5% interest from 1951 until date of initial entitlement

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM SyntheticM	CompletedM
white femalesM	own retirementM	7,177M	7,532M	6,975M	7,379M	7,348M	7,715M	0M	9,859M	9,142M
	disabilityM	4,976M	5,140M	4,774M	5,179M	4,993M	5,287M	0M	11,836M	7,725M
	aged spouseM	702M	692M	412M	991M	366M	1,018M	0M	15,763M	20,017M
	aged widowM	1,726M	1,710M	1,643M	1,808M	1,615M	1,805M	0M	2,455M	3,175M
	otherM	1,187M	1,242M	1,113M	1,261M	1,150M	1,334M	0M	1,995M	2,966M
black femalesM	own retirementM	7,247M	7,656M	6,871M	7,623M	7,356M	7,956M	0M	33,289M	32,941M
	disabilityM	4,465M	4,849M	4,261M	4,670M	4,531M	5,167M	0M	14,657M	33,688M
	aged spouseM	707M	664M	475M	938M	408M	919M	0M	13,710M	17,379M
	aged widowM	2,038M	2,256M	1,746M	2,330M	1,857M	2,656M	0M	31,310M	57,392M
	otherM	1,139M	1,282M	944M	1,334M	1,087M	1,477M	0M	13,535M	14,062M
white malesM	own retirementM	16,789M	17,985M	16,505M	17,074M	17,743M	18,227M	0M	17,721M	17,344M
	disabilityM	9,321M	9,945M	9,023M	9,618M	9,724M	10,167M	0M	25,702M	17,573M
	aged spouseM	1,284M	1,401M	643M	1,925M	663M	2,138M	0M	115,566M	146,144M
	aged widowM	5,767M	6,209M	3,201M	8,333M	3,324M	9,095M	0M	2,314,225M	3,036,071M
	otherM	1,529M	1,802M	1,160M	1,898M	1,382M	2,222M	0M	49,990M	65,210M
black malesM	own retirementM	13,730M	13,975M	12,802M	14,658M	13,305M	14,646M	1M	298,247M	139,161M
	disabilityM	6,835M	6,803M	5,811M	7,859M	6,426M	7,180M	0M	233,733M	51,875M
	aged spouseM	1,436M	1,187M	1,024M	1,848M	48M	2,326M	1M	58,685M	284,640M
	aged widowM	2,495M	1,240M	-106M	5,096M	485M	1,996M	0M	2,180,277M	204,410M
	otherM	1,409M	1,883M	771M	2,047M	869M	2,897M	0M	148,969M	370,626M

Table 14: Year of initial entitlement

Demographic Group	Type of Benefit	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic DF Not Exist	Total Variance	
		Synthetic	Completed						Synthetic	Completed
white females	own retirement	1990.495	1990.365	1990.366	1990.625	1990.233	1990.497	0	0.00521	0.00584
	disability	1991.684	1991.793	1991.491	1991.877	1991.537	1992.049	0	0.01357	0.02242
	aged spouse	1989.963	1989.739	1989.637	1990.290	1989.352	1990.126	0	0.03601	0.04880
	aged widow	1985.799	1985.683	1985.596	1986.002	1985.442	1985.924	0	0.01483	0.01989
	other	1980.468	1980.537	1980.201	1980.735	1980.334	1980.741	0	0.02160	0.01468
black females	own retirement	1990.550	1990.641	1990.197	1990.903	1990.141	1991.141	0	0.02547	0.07539
	disability	1992.005	1992.477	1991.659	1992.351	1991.923	1993.031	0	0.04377	0.09590
	aged spouse	1985.726	1984.274	1984.200	1987.252	1982.849	1985.698	0	0.75069	0.73582
	aged widow	1986.729	1987.628	1986.062	1987.397	1986.814	1988.442	0	0.15422	0.23381
	other	1980.272	1980.228	1979.910	1980.634	1979.721	1980.735	0	0.04777	0.08356
white males	own retirement	1991.261	1991.415	1991.114	1991.409	1991.317	1991.513	0	0.00318	0.00340
	disability	1990.464	1990.757	1990.189	1990.738	1990.514	1991.001	0	0.02214	0.01963
	aged spouse	1991.398	1991.777	1989.431	1993.365	1989.445	1994.108	0	1.17039	1.56484
	aged widow	1990.582	1991.390	1987.960	1993.204	1988.920	1993.861	0	2.23807	2.20803
	other	1978.964	1979.042	1978.706	1979.222	1978.812	1979.272	0	0.02174	0.01876
black males	own retirement	1992.160	1992.315	1991.870	1992.450	1992.003	1992.627	1	0.02911	0.03499
	disability	1990.428	1990.793	1989.913	1990.944	1990.335	1991.250	0	0.08794	0.07576
	aged spouse	1993.600	1994.522	1991.709	1995.492	1992.364	1996.680	0	1.01994	1.50382
	aged widow	1988.757	1988.541	1984.973	1992.541	1985.856	1991.227	0	4.72418	2.58058
	other	1979.383	1979.345	1978.910	1979.856	1978.894	1979.796	0	0.07976	0.07508

Table 15: Initial onthly Benefit Amount

DemographicM GroupM	Type of BenefitM	eanM SyntheticM	CompletedM	Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM SyntheticM	CompletedM
white femalesM	own retirementM	441.09M	444.99M	437.29M	444.89M	441.19M	448.79M	0M	4.88M	5.05M
	disabilityM	498.25M	495.18M	490.27M	506.24M	485.83M	504.53M	0M	21.96M	28.29M
	aged spouseM	330.97M	337.90M	325.31M	336.62M	332.40M	343.40M	0M	11.04M	10.63M
	aged widowM	384.91M	388.70M	374.64M	395.17M	377.27M	400.13M	0M	31.14M	37.17M
	otherM	169.50M	171.87M	162.97M	176.02M	165.89M	177.85M	0M	11.27M	9.97M
black femalesM	own retirementM	430.14M	439.66M	416.91M	443.36M	427.77M	451.54M	0M	51.22M	49.65M
	disabilityM	482.48M	495.58M	467.29M	497.68M	474.45M	516.71M	0M	75.18M	130.05M
	aged spouseM	255.94M	241.41M	230.75M	281.14M	223.13M	259.70M	0M	190.50M	122.32M
	aged widowM	381.71M	389.41M	358.37M	405.05M	357.10M	421.72M	0M	169.10M	324.52M
	otherM	136.92M	130.13M	127.00M	146.84M	120.37M	139.90M	0M	28.09M	27.63M
white malesM	own retirementM	716.69M	730.33M	710.73M	722.66M	725.39M	735.28M	0M	9.39M	8.31M
	disabilityM	680.11M	685.26M	672.84M	687.39M	676.86M	693.65M	0M	19.34M	25.20M
	aged spouseM	236.14M	238.09M	210.16M	262.13M	212.11M	264.06M	0M	246.03M	246.42M
	aged widowM	449.02M	457.16M	311.18M	586.86M	326.82M	587.49M	0M	5,745.09M	5,608.83M
	otherM	156.91M	158.97M	153.71M	160.12M	155.47M	162.46M	0M	3.75M	4.44M
black malesM	own retirementM	630.27M	629.62M	606.77M	653.78M	615.87M	643.37M	1M	191.17M	68.41M
	disabilityM	598.74M	589.93M	572.92M	624.57M	575.24M	604.61M	0M	182.97M	79.67M
	aged spouseM	278.44M	319.22M	208.73M	348.15M	206.15M	432.29M	0M	1,646.79M	3,656.73M
	aged widowM	334.20M	331.05M	203.67M	464.74M	191.72M	470.38M	0M	5,382.22M	5,362.25M
	otherM	129.48M	123.13M	122.55M	136.41M	116.54M	129.72M	0M	16.31M	16.01M

Table 16: onthly Benefit Amount April 2000

DemographicM	Type of	eanM		Confidence IntervalM		Confidence IntervalM		SyntheticM	Total VarianceM	
GroupM	BenefitM	SyntheticM	CompletedM	SyntheticM		CompletedM		DF Not ExistM	SyntheticM	CompletedM
white femalesM	own retirementM	583M	563M	573M	594M	558M	569M	0M	27M	12M
	disabilityM	581M	598M	564M	598M	591M	605M	0M	59M	16M
	aged spouseM	560M	598M	551M	569M	590M	606M	0M	24M	22M
	aged widowM	542M	584M	530M	554M	567M	602M	0M	48M	98M
	otherM	594M	643M	581M	607M	626M	660M	0M	61M	109M
black femalesM	own retirementM	485M	469M	472M	498M	457M	481M	0M	54M	53M
	disabilityM	442M	445M	429M	456M	431M	459M	0M	55M	69M
	aged spouseM	430M	448M	400M	460M	418M	477M	0M	229M	254M
	aged widowM	444M	450M	414M	474M	408M	492M	0M	305M	634M
	otherM	507M	620M	415M	598M	558M	682M	0M	1,891M	1,313M
white malesM	own retirementM	715M	709M	682M	749M	700M	717M	0M	214M	25M
	disabilityM	719M	739M	685M	753M	731M	747M	0M	205M	23M
	aged spouseM	708M	745M	682M	735M	734M	755M	0M	138M	40M
	aged widowM	786M	812M	758M	814M	796M	828M	0M	200M	91M
	otherM	844M	886M	796M	893M	869M	904M	0M	421M	106M
black malesM	own retirementM	598M	581M	563M	633M	562M	599M	0M	229M	118M
	disabilityM	538M	522M	489M	587M	501M	544M	0M	472M	167M
	aged spouseM	514M	490M	487M	540M	462M	518M	0M	228M	276M
	aged widowM	567M	584M	519M	616M	490M	678M	0M	837M	2,682M
	otherM	701M	650M	574M	828M	583M	717M	0M	3,801M	1,652M

Table 17: Average Indexed   onthly Earnings or Average   onthly Wage

DemographicM	Type of	eanM		Confidence IntervalM		Confidence IntervalM		SyntheticM	Total VarianceM	
GroupM	BenefitM	SyntheticM	CompletedM	SyntheticM		CompletedM		DF Not ExistM	SyntheticM	CompletedM
white femalesM	own retirementM	725M	760M	707M	743M	744M	776M	0M	80M	71M
	disabilityM	896M	907M	877M	916M	883M	932M	0M	138M	204M
	aged spouseM	112M	116M	90M	135M	86M	145M	0M	114M	179M
	aged widowM	195M	193M	187M	204M	183M	203M	0M	25M	35M
	otherM	1,184M	1,180M	1,157M	1,210M	1,149M	1,210M	0M	249M	313M
black femalesM	own retirementM	724M	773M	693M	756M	746M	801M	0M	248M	276M
	disabilityM	833M	870M	784M	881M	812M	927M	0M	670M	986M
	aged spouseM	151M	159M	123M	179M	116M	201M	0M	263M	582M
	aged widowM	229M	252M	199M	260M	211M	294M	0M	340M	614M
	otherM	978M	949M	914M	1,041M	903M	996M	0M	1,269M	800M
white malesM	own retirementM	1,693M	1,789M	1,666M	1,720M	1,771M	1,807M	0M	161M	103M
	disabilityM	1,621M	1,691M	1,546M	1,695M	1,619M	1,762M	0M	1,237M	1,196M
	aged spouseM	152M	157M	101M	204M	92M	221M	0M	904M	1,220M
	aged widowM	651M	646M	368M	934M	350M	942M	0M	28,080M	31,743M
	otherM	2,016M	2,031M	1,954M	2,078M	1,969M	2,093M	0M	1,156M	1,198M
black malesM	own retirementM	1,373M	1,399M	1,293M	1,453M	1,348M	1,450M	1M	2,227M	866M
	disabilityM	1,244M	1,213M	1,153M	1,336M	1,159M	1,267M	0M	2,184M	1,026M
	aged spouseM	146M	99M	106M	186M	30M	167M	1M	547M	1,289M
	aged widowM	292M	143M	-29M	613M	55M	231M	0M	32,478M	2,800M
	otherM	1,336M	1,297M	1,242M	1,431M	1,225M	1,369M	0M	2,917M	1,906M

*AIME for individuals who reach age 62 after 1979, otherwise AMW

Table 18: Primary Insurance Amount

DemographicM GroupM	Type of BenefitM	eanM		Confidence IntervalM SyntheticM		Confidence IntervalM CompletedM		SyntheticM DF Not ExistM	Total VarianceM	
		SyntheticM	CompletedM						SyntheticM	CompletedM
white femalesM	own retirementM	422M	434M	415M	430M	428M	440M	0M	14M	11M
	disabilityM	498M	502M	489M	508M	495M	510M	0M	27M	22M
	aged spouseM	100M	103M	81M	118M	80M	126M	0M	70M	104M
	aged widowM	178M	181M	170M	185M	173M	188M	0M	17M	19M
	otherM	409M	406M	404M	414M	400M	412M	0M	10M	13M
black femalesM	own retirementM	424M	441M	412M	436M	430M	452M	0M	31M	46M
	disabilityM	478M	493M	459M	498M	471M	515M	0M	107M	147M
	aged spouseM	118M	113M	96M	139M	88M	137M	0M	134M	169M
	aged widowM	184M	197M	166M	203M	169M	224M	0M	124M	250M
	otherM	381M	371M	368M	393M	360M	382M	0M	53M	47M
white malesM	own retirementM	749M	776M	740M	758M	770M	782M	0M	17M	11M
	disabilityM	710M	729M	684M	735M	707M	752M	0M	139M	118M
	aged spouseM	149M	162M	108M	190M	113M	211M	0M	544M	742M
	aged widowM	395M	366M	261M	530M	209M	523M	0M	6,414M	8,824M
	otherM	490M	490M	480M	500M	481M	500M	0M	31M	30M
black malesM	own retirementM	656M	661M	625M	686M	646M	676M	1M	322M	81M
	disabilityM	603M	593M	573M	633M	574M	611M	0M	241M	119M
	aged spouseM	138M	110M	109M	168M	57M	162M	1M	303M	820M
	aged widowM	199M	123M	33M	365M	45M	201M	0M	9,143M	2,171M
	otherM	434M	424M	410M	458M	402M	445M	0M	163M	147M

Table 19: SER work indicator for year 1965

DemographicC	Mean		onfidence Interval		Confidence IntervalC		SyntheticC	Total VarianceC	
GroupC	Synthetic	CompletedC	Synthetic		ompletedC		DF Not ExistC	Synthetic	CompletedC
white femaleC	0.178C	0.176C	0.176C	0.180C	0.174C	0.178C	0C	0.000001C	0.000001C
black femaleC	0.143C	0.148C	0.139C	0.148C	0.143C	0.154C	0C	0.000007C	0.000010C
white maleC	0.280C	0.276C	0.277C	0.283C	0.273C	0.278C	0C	0.000003C	0.000002C
black maleC	0.212C	0.199C	0.203C	0.221C	0.190C	0.207C	0C	0.000018C	0.000023C

Table 20: SER Earnings for year 1965

DemographicC GroupC	Mean		onfidence Interval		Confidence IntervalC		SyntheticC	Total VarianceC	
	Synthetic	CompletedC	Synthetic		ompletedC		DF Not ExistC	Synthetic	CompletedC
white femaleC	354C	375C	347C	362C	369C	381C	1C	18C	12C
black femaleC	218C	243C	191C	244C	231C	256C	0C	120C	55C
white maleC	916C	940C	905C	927C	931C	950C	0C	35C	30C
black maleC	589C	548C	575C	603C	520C	577C	1C	67C	256C

Table 21: SER work indicator for year 1975

DemographicC GroupC	Mean Synthetic	Mean CompletedC	onfidence Interval Synthetic		Confidence IntervalC ompletedC		SyntheticC DF Not ExistC	Total VarianceC Synthetic	Total VarianceC CompletedC
white femaleC	0.326C	0.326C	0.324C	0.328C	0.324C	0.329C	0C	0.000001C	0.000003C
black femaleC	0.285C	0.301C	0.278C	0.292C	0.294C	0.307C	0C	0.000014C	0.000015C
white maleC	0.441C	0.442C	0.438C	0.443C	0.440C	0.445C	0C	0.000003C	0.000003C
black maleC	0.356C	0.349C	0.343C	0.369C	0.340C	0.357C	0C	0.000036C	0.000026C

Table 22: SER Earningsfor year 1975

DemographicC GroupC	Mean Synthetic	Mean CompletedC	onfidence Interval Synthetic		Confidence IntervalC ompletedC		SyntheticC DF Not ExistC	Total VarianceC Synthetic	CompletedC
white femaleC	1,382C	1,457C	1,255C	1,509C	1,441C	1,472C	0C	2,410C	89C
black femaleC	1,085C	1,257C	1,000C	1,170C	1,222C	1,291C	0C	1,487C	441C
white maleC	3,533C	3,668C	3,486C	3,581C	3,631C	3,705C	1C	771C	446C
black maleC	2,170C	2,171C	2,021C	2,319C	2,089C	2,254C	0C	3,172C	2,215C

Table 23: SER work indicator for year 1985

DemographicC GroupC	Mean Synthetic	CompletedC	onfidence Interval Synthetic		Confidence IntervalC ompletedC		SyntheticC DF Not ExistC	Total VarianceC Synthetic	CompletedC
white femaleC	0.448C	0.449C	0.445C	0.451C	0.447C	0.452C	0C	0.000003C	0.000002C
black femaleC	0.400C	0.397C	0.391C	0.409C	0.391C	0.404C	0C	0.000025C	0.000015C
white maleC	0.549C	0.550C	0.546C	0.552C	0.548C	0.553C	0C	0.000003C	0.000003C
black maleC	0.453C	0.434C	0.445C	0.460C	0.426C	0.441C	0C	0.000020C	0.000021C

Table 24: SER Earnings for year 1985

DemographicC GroupC	Mean Synthetic	CompletedC	onfidence Interval Synthetic		Confidence IntervalC ompletedC		SyntheticC DF Not ExistC	Total VarianceC Synthetic	CompletedC
white femaleC	4,804C	4,723C	4,680C	4,929C	4,684C	4,761C	1C	5,388C	552C
black femaleC	3,733C	3,764C	3,657C	3,810C	3,674C	3,853C	1C	2,044C	2,968C
white maleC	9,892C	10,091C	9,763C	10,021C	10,023C	10,160C	1C	5,746C	1,728C
black maleC	6,139C	5,667C	5,849C	6,428C	5,518C	5,816C	1C	29,044C	8,131C

Table 25: SER work indicator for year 1995

DemographicC	Mean		onfidence Interval		Confidence IntervalC		SyntheticC	Total VarianceC	
GroupC	Synthetic	CompletedC	Synthetic		ompletedC		DF Not ExistC	Synthetic	CompletedC
white femaleC	0.567C	0.564C	0.564C	0.569C	0.560C	0.569C	0C	0.000002C	0.000006C
black femaleC	0.570C	0.545C	0.555C	0.586C	0.536C	0.554C	0C	0.000059C	0.000025C
white maleC	0.671C	0.675C	0.667C	0.675C	0.671C	0.679C	0C	0.000005C	0.000005C
black maleC	0.606C	0.581C	0.598C	0.615C	0.573C	0.590C	0C	0.000023C	0.000024C

Table 26: SER Earningsfor year 1995

DemographicC GroupC	Mean Synthetic	CompletedC	onfidence Interval Synthetic		Confidence IntervalC ompletedC		SyntheticC DF Not ExistC	Total VarianceC Synthetic	CompletedC
white femaleC	9,561C	9,443C	9,441C	9,680C	9,365C	9,522C	1C	4,964C	2,193C
black femaleC	8,153C	7,589C	7,991C	8,315C	7,412C	7,766C	1C	9,040C	11,134C
white maleC	17,047C	17,241C	16,837C	17,257C	17,114C	17,367C	1C	15,300C	5,512C
black maleC	11,023C	10,087C	10,658C	11,388C	9,870C	10,303C	1C	45,990C	17,301C

Table 27: Total SER Earnings 1951-2003

DemographicC GroupC	Mean		onfidence Interval Synthetic		Confidence IntervalC ompletedC		SyntheticC DF Not ExistC	Total VarianceC	
	Synthetic	CompletedC						Synthetic	CompletedC
white femaleC	211,817C	212,751C	210,419C	213,214C	211,508C	213,995C	0C	489,904C	564,662C
black femaleC	178,709C	174,331C	174,332C	183,085C	171,509C	177,153C	0C	5,112,046C	2,929,414C
white maleC	400,702C	409,869C	398,904C	402,500C	407,744C	411,994C	0C	795,526C	1,607,808C
black maleC	257,525C	240,933C	246,154C	268,896C	236,879C	244,987C	0C	24,866,833C	5,073,858C

Table 28: Total Years worked in SER (i.e. positive FICA earnings)

DemographicC	Mean		onfidence Interval		Confidence IntervalC		SyntheticC	Total VarianceC	
GroupC	Synthetic	CompletedC	Synthetic		ompletedC		DF Not ExistC	Synthetic	CompletedC
white femaleC	17.759C	17.992C	17.657C	17.862C	17.937C	18.048C	0C	0.002876C	0.001130C
black femaleC	16.310C	16.569C	16.128C	16.491C	16.421C	16.717C	0C	0.010407C	0.008026C
white maleC	22.698C	22.957C	22.638C	22.758C	22.889C	23.026C	0C	0.001320C	0.001710C
black maleC	18.992C	18.412C	18.607C	19.376C	18.184C	18.640C	0C	0.023317C	0.017944C

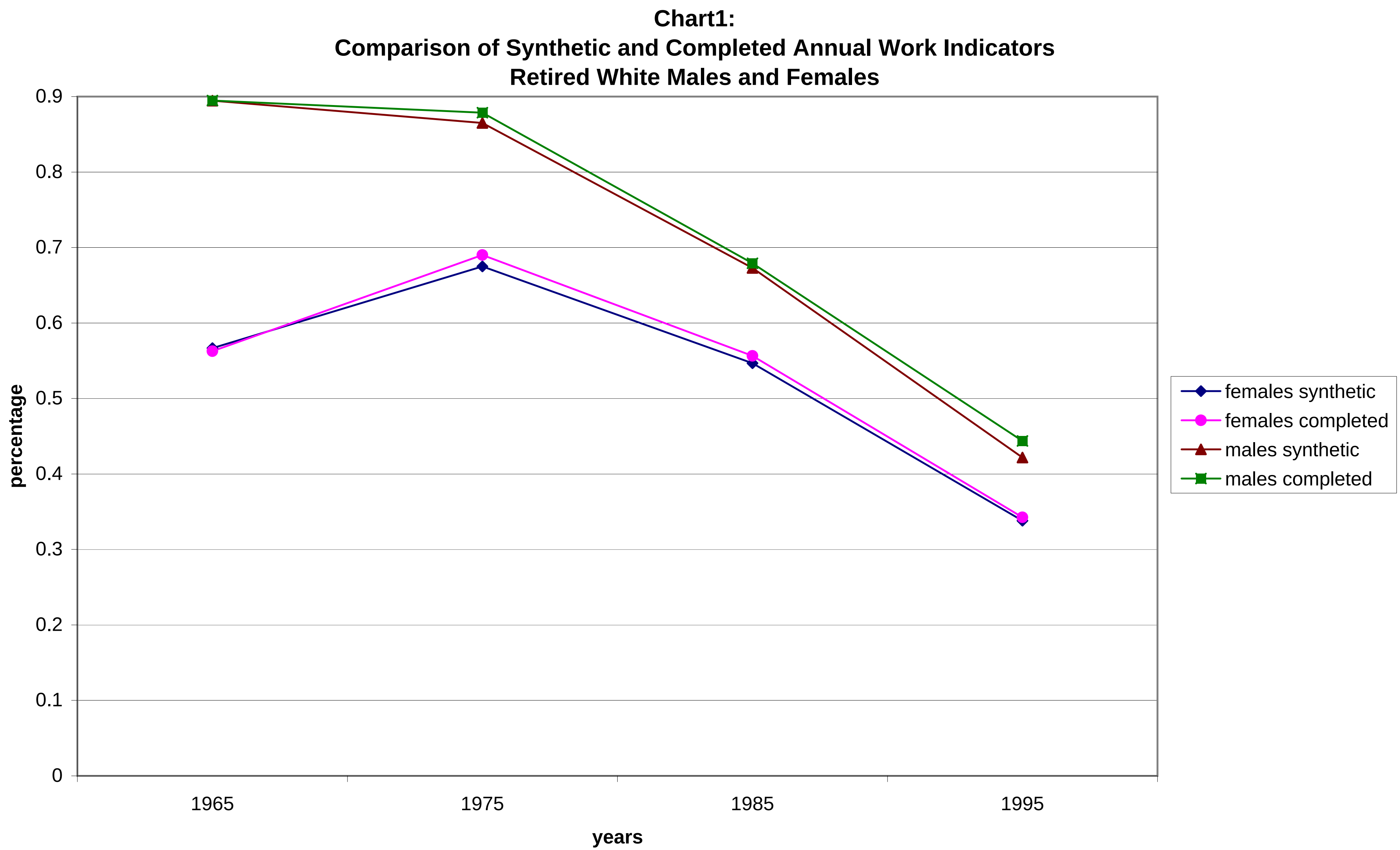


Chart1:
Comparison of Synthetic and Completed Annual Work Indicators
Retired White Males and Females
percentage
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1965
1975
1985
1995
years
females synthetic
females completed
males synthetic
males completed

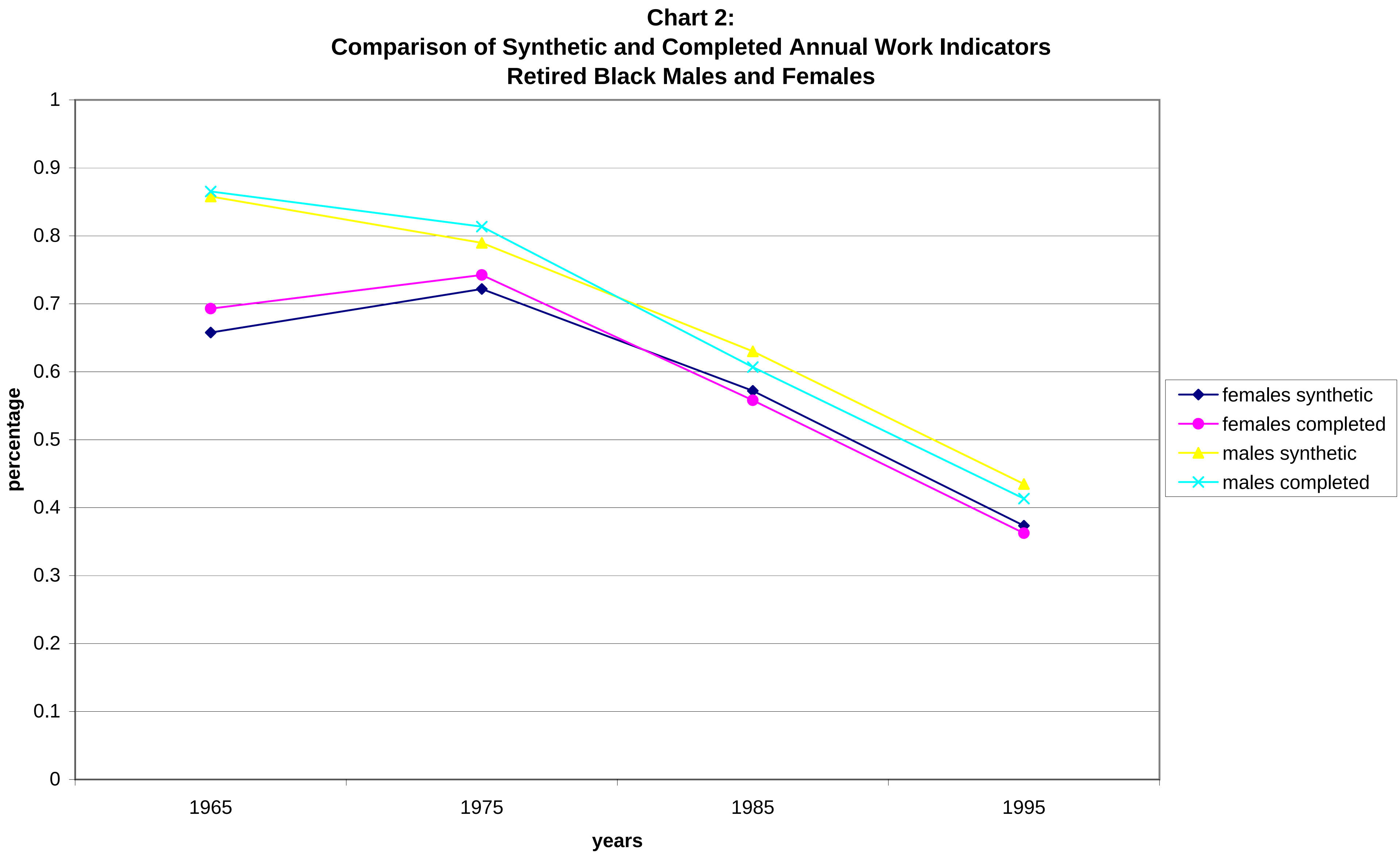
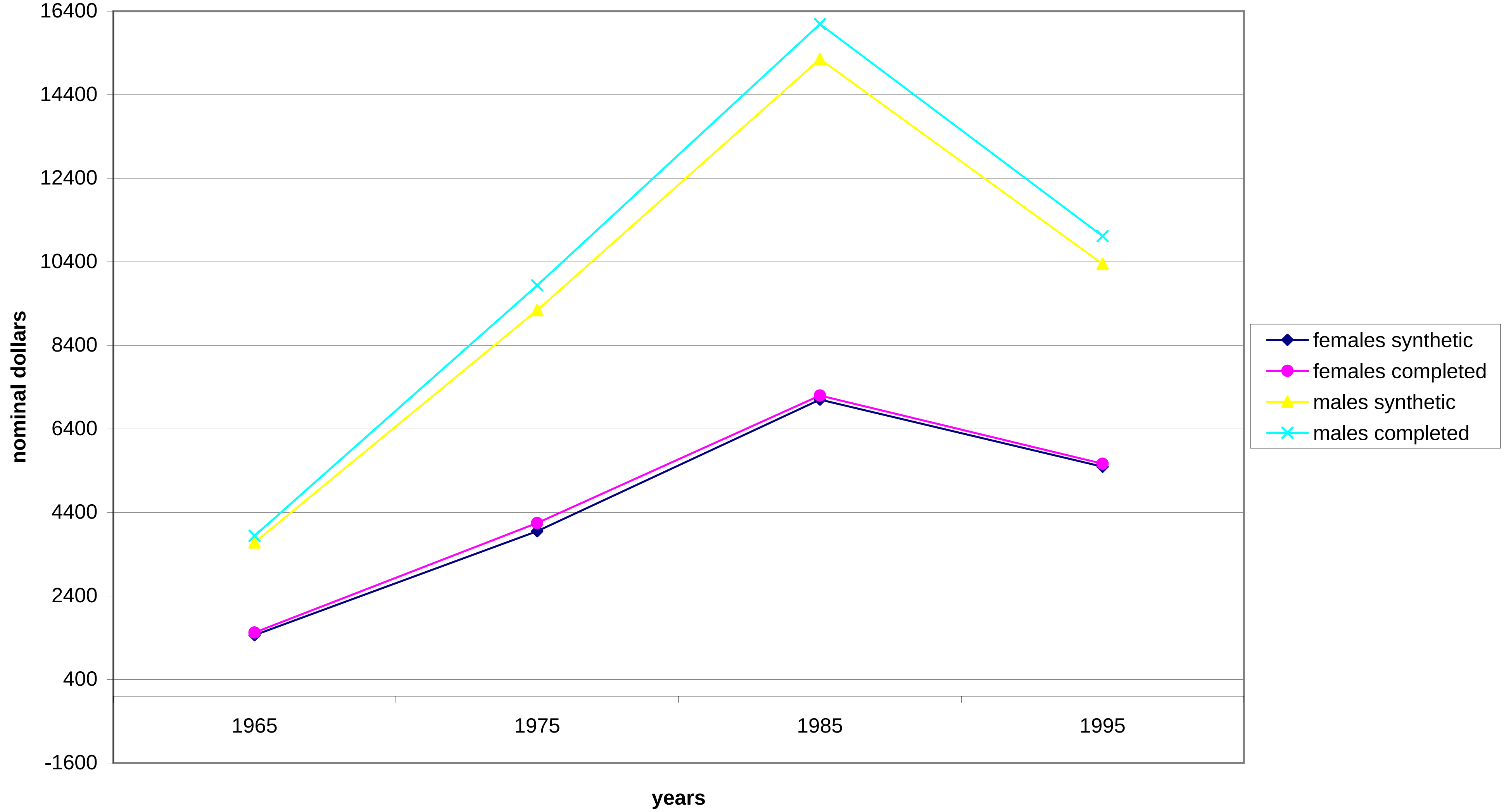


Chart 2:
Comparison of Synthetic and Completed Annual Work Indicators
Retired Black Males and Females
percentage
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1965
1975
1985
1995
years
females synthetic
females completed
males synthetic
males completed

**Chart 3:
Comparison of Synthetic and Completed Earnings
Retired White Males and Females**



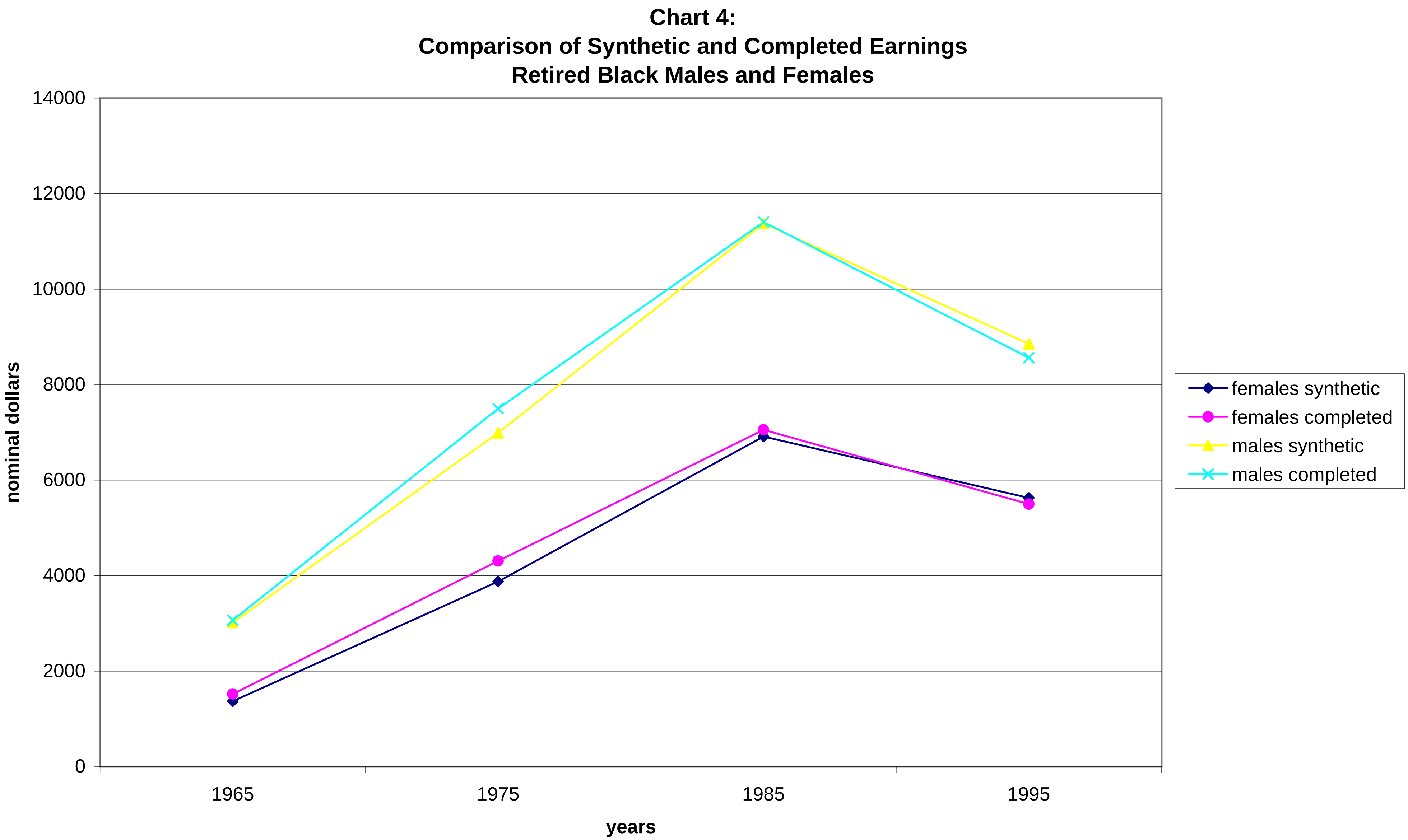


Chart 4:
Comparison of Synthetic and Completed Earnings
Retired Black Males and Females
14000
12000
10000
8000
6000
4000
2000
0
nominal dollars
1965
1975
1985
1995
years
females synthetic
females completed
males synthetic
males completed

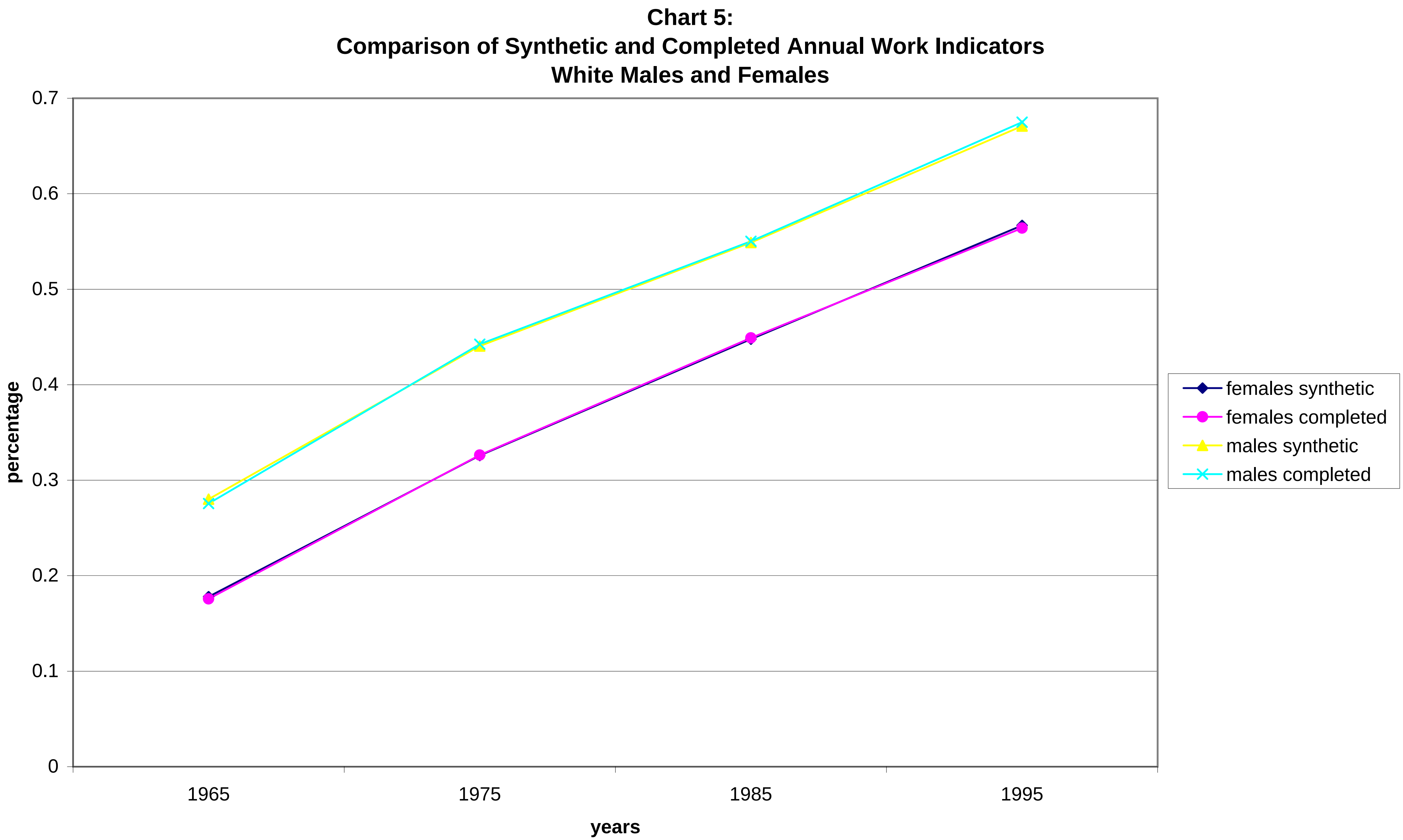


Chart 5:
Comparison of Synthetic and Completed Annual Work Indicators
White Males and Females
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
percentage
1965
1975
1985
1995
years
females synthetic
females completed
males synthetic
males completed

Chart 6:
Comparison of Synthetic and Completed Annual Work Indicators
Black Males and Females

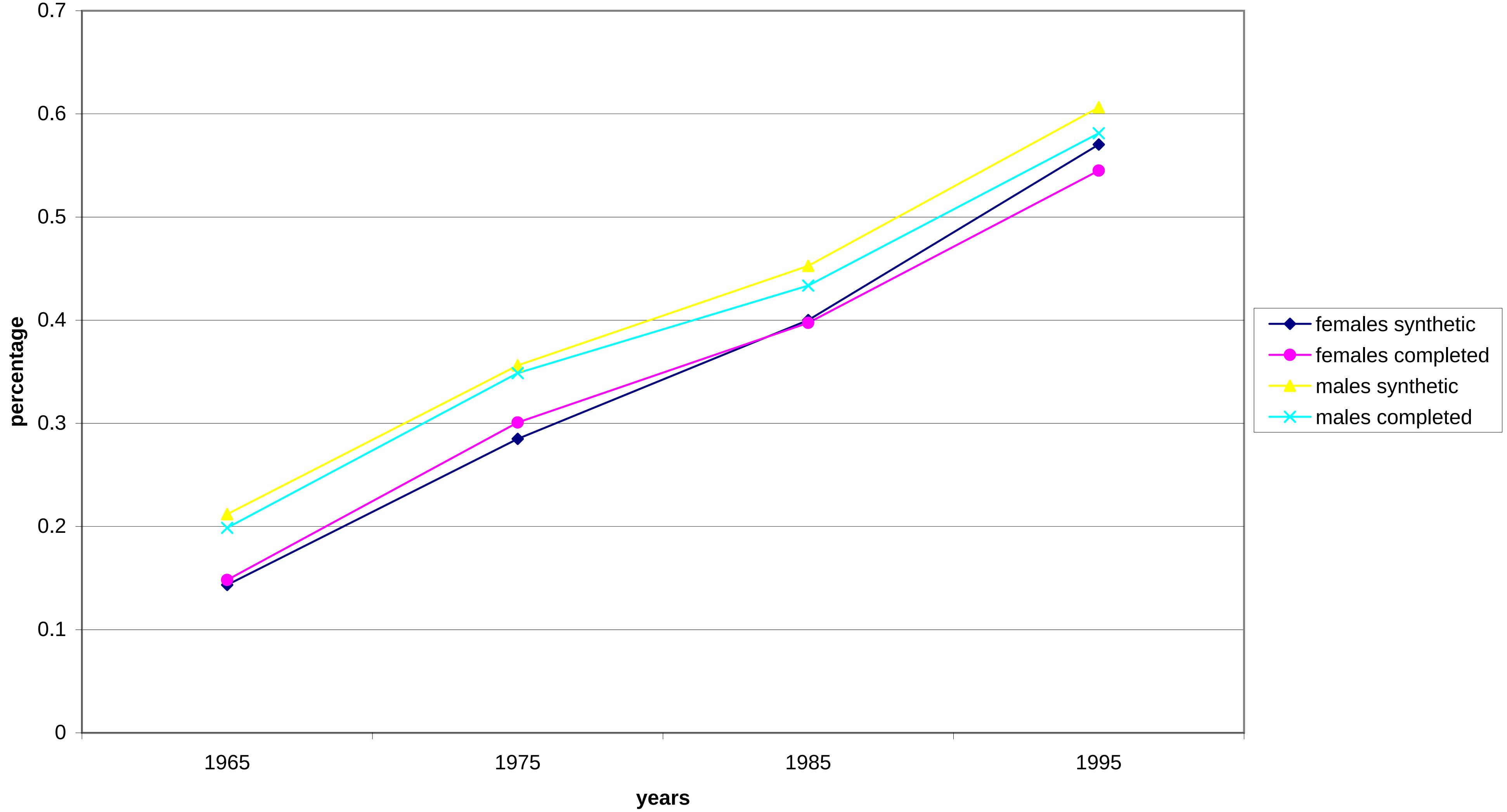


Chart 7:
Comparison of Synthetic and Completed Earnings
White Males and Females

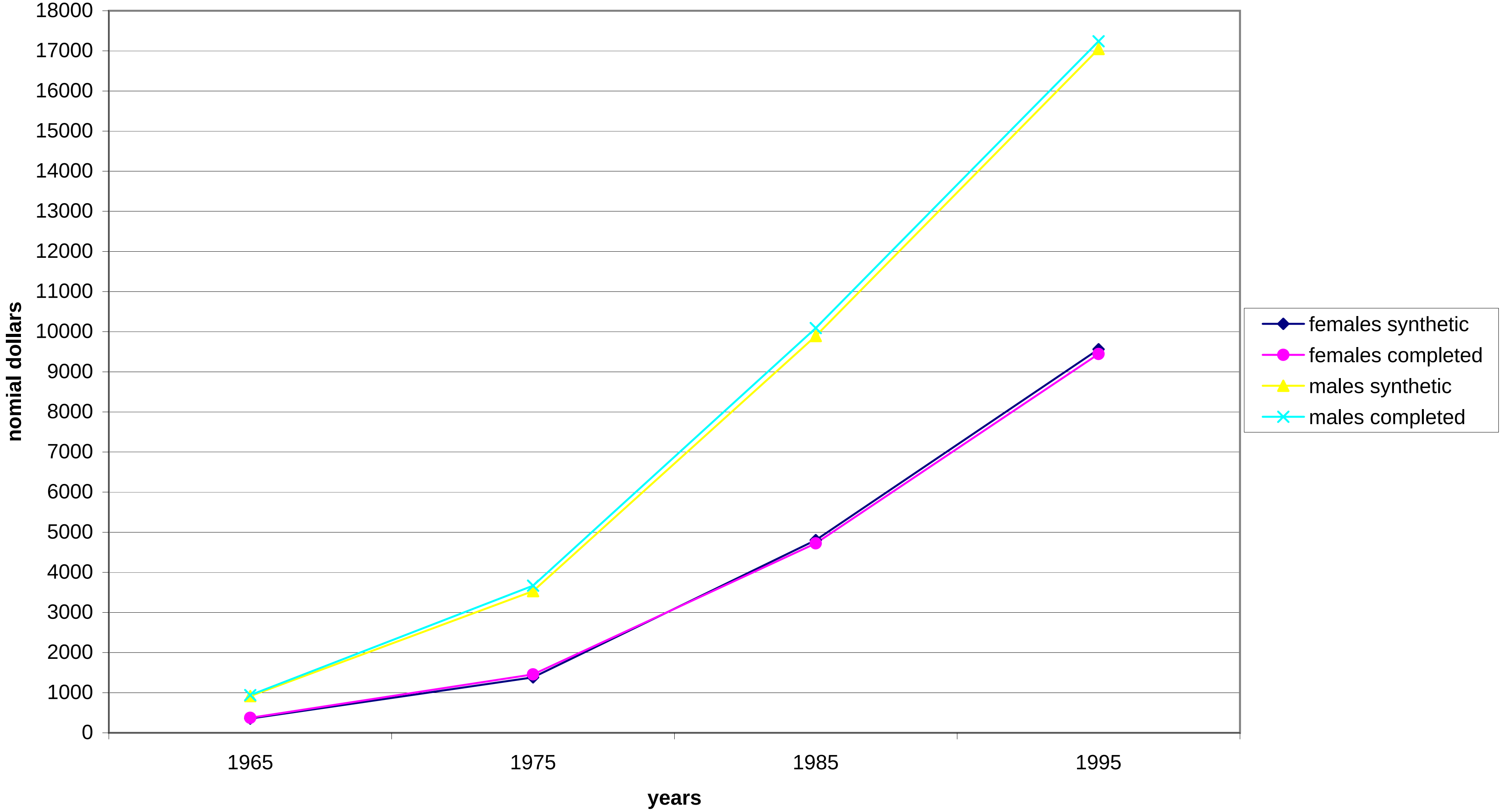


Chart 8:
Comparison of Synthetic and Completed Earnings
Black Males and Females

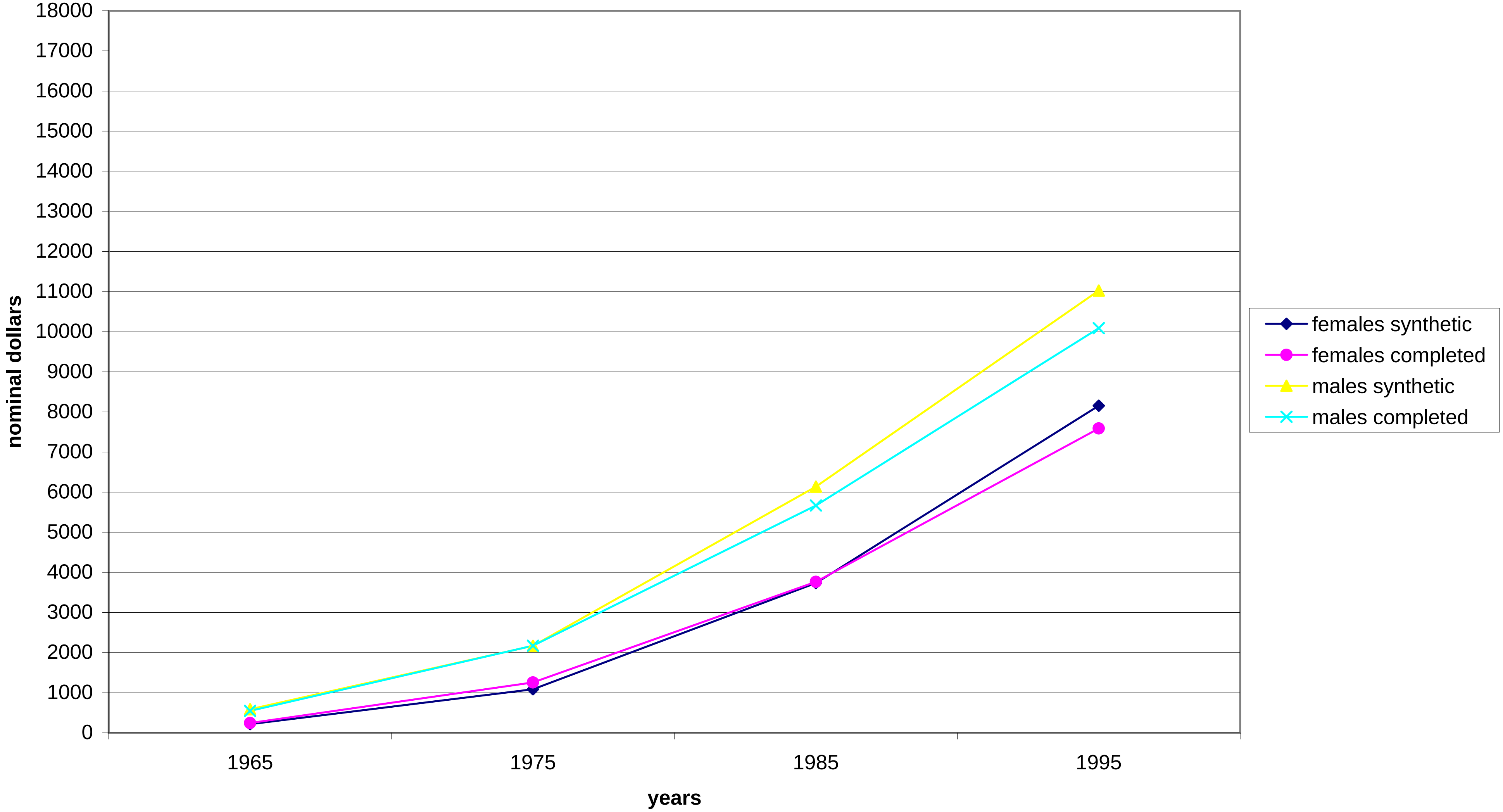


Table 29: MBA 2000 by demographic group and education

Demographic Groupv	Educationv Categoryv	Meanv Syntheticv	Completedv	Confidence Intervalv Syntheticv		Confidence Intervalv Completedv		Syntheticv DF Not Existv	Total Variancev Syntheticv	Completedv
white femalesv	no HSv	583v	563v	573v	594v	558v	569v	0v	27v	12v
	HSv	581v	598v	564v	598v	591v	605v	0v	59v	16v
	Some Collv	560v	598v	551v	569v	590v	606v	0v	24v	22v
	Collegev	542v	584v	530v	554v	567v	602v	0v	48v	98v
	Graduatev	594v	643v	581v	607v	626v	660v	0v	61v	109v
black femalesv	no HSv	485v	469v	472v	498v	457v	481v	0v	54v	53v
	HSv	442v	445v	429v	456v	431v	459v	0v	55v	69v
	Some Collv	430v	448v	400v	460v	418v	477v	0v	229v	254v
	Collegev	444v	450v	414v	474v	408v	492v	0v	305v	634v
	Graduatev	507v	620v	415v	598v	558v	682v	0v	1,891v	1,313v
white malesv	no HSv	715v	709v	682v	749v	700v	717v	0v	214v	25v
	HSv	719v	739v	685v	753v	731v	747v	0v	205v	23v
	Some Collv	708v	745v	682v	735v	734v	755v	0v	138v	40v
	Collegev	786v	812v	758v	814v	796v	828v	0v	200v	91v
	Graduatev	844v	886v	796v	893v	869v	904v	0v	421v	106v
black malesv	no HSv	598v	581v	563v	633v	562v	599v	0v	229v	118v
	HSv	538v	522v	489v	587v	501v	544v	0v	472v	167v
	Some Collv	514v	490v	487v	540v	462v	518v	0v	228v	276v
	Collegev	567v	584v	519v	616v	490v	678v	0v	837v	2,682v
	Graduatev	701v	650v	574v	828v	583v	717v	0v	3,801v	1,652v

Table 30: Non-deferred DER earnings at FICA covered jobs in year 2000

Demographicv Groupv	Educationv Categoryv	Meanv Syntheticv	Completedv	Confidence Intervalv Syntheticv		Confidence Intervalv Completedv		Syntheticv DF Not Existv	Total Variancev Syntheticv	Completedv
white femalesv	no HSv	14,347v	12,127v	13,370v	15,323v	11,900v	12,354v	1v	329,876v	18,339v
	HSv	19,365v	18,087v	18,346v	20,385v	17,889v	18,286v	0v	158,922v	14,163v
	Some Collv	23,788v	22,750v	22,951v	24,626v	22,478v	23,021v	1v	242,767v	26,524v
	Collegev	37,113v	40,244v	35,104v	39,122v	35,498v	44,990v	1v	1,396,219v	8,326,165v
	Graduatev	41,993v	45,673v	39,748v	44,238v	44,110v	47,237v	1v	1,743,860v	903,424v
black femalesv	no HSv	12,228v	10,116v	11,594v	12,862v	9,716v	10,515v	1v	139,051v	58,343v
	HSv	17,665v	15,890v	15,755v	19,574v	15,474v	16,306v	1v	1,261,484v	63,757v
	Some Collv	22,279v	20,658v	20,540v	24,018v	20,180v	21,136v	1v	1,046,677v	84,419v
	Collegev	34,678v	34,163v	28,613v	40,743v	32,760v	35,566v	1v	12,729,526v	726,972v
	Graduatev	42,786v	46,089v	24,999v	60,572v	39,969v	52,209v	0v	54,819,308v	13,842,902v
white malesv	no HSv	22,488v	19,493v	20,356v	24,620v	19,107v	19,878v	0v	889,127v	50,091v
	HSv	33,480v	28,981v	31,743v	35,217v	28,618v	29,344v	0v	457,599v	48,371v
	Some Collv	41,935v	37,778v	36,181v	47,689v	35,693v	39,863v	1v	11,455,543v	1,606,545v
	Collegev	73,805v	71,557v	60,892v	86,718v	67,724v	75,390v	0v	39,656,678v	5,415,726v
	Graduatev	87,676v	97,780v	52,174v	123,177v	87,226v	108,334v	0v	244,178,702v	41,171,337v
black malesv	no HSv	17,240v	13,829v	15,180v	19,299v	13,111v	14,547v	0v	795,733v	185,727v
	HSv	24,834v	23,625v	22,925v	26,744v	21,257v	25,993v	1v	1,261,829v	2,071,917v
	Some Collv	30,040v	26,578v	25,958v	34,123v	25,562v	27,593v	1v	5,767,275v	380,971v
	Collegev	51,602v	46,514v	33,670v	69,533v	41,896v	51,133v	0v	88,347,426v	7,792,256v
	Graduatev	79,714v	64,460v	-35,258v	194,687v	48,561v	80,360v	0v	4,023,002,613	93,392,302v

Table 31: Total SER Earnings in year 2000

Demographicv Groupv	Educationv Categoryv	Meanv Syntheticv	Completedv	Confidence Intervalv Syntheticv		Confidence Intervalv Completedv		Syntheticv DF Not Existv	Total Variancev Syntheticv	Completedv
white femalesv	no HSv	6,475v	5,421v	5,379v	7,572v	5,299v	5,543v	0v	197,635v	5,258v
	HSv	10,623v	10,048v	10,383v	10,862v	9,911v	10,185v	1v	19,877v	6,592v
	Some Collv	15,443v	15,187v	15,255v	15,630v	15,016v	15,358v	1v	12,137v	10,790v
	Collegev	21,319v	21,206v	20,733v	21,904v	20,834v	21,579v	1v	118,553v	50,338v
	Graduatev	23,213v	25,799v	22,519v	23,907v	25,311v	26,286v	1v	166,698v	87,334v
black femalesv	no HSv	6,039v	5,047v	5,536v	6,541v	4,839v	5,256v	1v	87,461v	15,884v
	HSv	10,918v	9,722v	10,459v	11,376v	9,414v	10,031v	1v	72,827v	34,296v
	Some Collv	15,324v	14,827v	14,667v	15,980v	14,427v	15,227v	1v	149,168v	59,123v
	Collegev	21,304v	21,630v	17,955v	24,654v	20,476v	22,785v	0v	1,674,893v	491,591v
	Graduatev	23,047v	24,907v	19,996v	26,097v	22,977v	26,836v	0v	1,835,723v	1,304,147v
white malesv	no HSv	12,534v	11,598v	10,828v	14,241v	11,373v	11,824v	0v	488,189v	17,105v
	HSv	20,307v	18,559v	19,893v	20,722v	18,343v	18,776v	1v	59,358v	16,462v
	Some Collv	25,595v	25,041v	24,795v	26,396v	24,781v	25,302v	1v	221,947v	24,726v
	Collegev	34,103v	35,830v	31,265v	36,942v	35,385v	36,275v	0v	1,325,716v	72,845v
	Graduatev	33,841v	38,602v	32,121v	35,560v	38,094v	39,110v	0v	561,532v	95,213v
black malesv	no HSv	8,890v	7,496v	7,216v	10,565v	7,085v	7,907v	0v	516,139v	57,999v
	HSv	15,135v	13,730v	14,669v	15,601v	13,268v	14,192v	1v	75,131v	77,575v
	Some Collv	18,978v	18,796v	18,181v	19,774v	18,118v	19,475v	1v	219,702v	167,113v
	Collegev	27,267v	27,542v	22,751v	31,783v	25,375v	29,710v	0v	4,519,161v	1,576,041v
	Graduatev	26,890v	29,630v	23,752v	30,029v	27,396v	31,864v	0v	2,186,459v	1,834,061v

Table 32: Foreign Born Indicator

Demographicv	Educationv	Meanv		Confidence Intervalv		Confidence Intervalv		Syntheticv	Total Variancev	
Groupv	Categoryv	Syntheticv	Completedv	Syntheticv		Completedv		DF Not Existv	Syntheticv	Completedv
white femalesv	no HSv	0.200v	0.235v	0.195v	0.205v	0.229v	0.241v	1v	0.000009v	0.000011v
	HSv	0.092v	0.096v	0.088v	0.096v	0.094v	0.099v	1v	0.000006v	0.000002v
	Some Collv	0.091v	0.089v	0.088v	0.095v	0.084v	0.095v	1v	0.000003v	0.000009v
	Collegev	0.121v	0.119v	0.114v	0.127v	0.114v	0.124v	0v	0.000007v	0.000009v
	Graduatev	0.105v	0.102v	0.100v	0.110v	0.096v	0.108v	1v	0.000008v	0.000013v
black femalesv	no HSv	0.058v	0.063v	0.048v	0.067v	0.055v	0.071v	0v	0.000011v	0.000022v
	HSv	0.059v	0.068v	0.055v	0.064v	0.062v	0.074v	0v	0.000004v	0.000013v
	Some Collv	0.063v	0.060v	0.060v	0.066v	0.053v	0.067v	1v	0.000003v	0.000018v
	Collegev	0.108v	0.099v	0.082v	0.135v	0.076v	0.123v	1v	0.000245v	0.000171v
	Graduatev	0.100v	0.088v	0.046v	0.154v	0.071v	0.105v	0v	0.000433v	0.000111v
white malesv	no HSv	0.204v	0.232v	0.196v	0.212v	0.227v	0.237v	0v	0.000011v	0.000009v
	HSv	0.109v	0.109v	0.052v	0.165v	0.088v	0.130v	0v	0.000568v	0.000085v
	Some Collv	0.093v	0.094v	0.084v	0.102v	0.090v	0.098v	0v	0.000012v	0.000006v
	Collegev	0.111v	0.104v	0.107v	0.115v	0.097v	0.112v	1v	0.000005v	0.000017v
	Graduatev	0.125v	0.131v	0.116v	0.134v	0.125v	0.138v	0v	0.000014v	0.000016v
black malesv	no HSv	0.069v	0.067v	0.059v	0.079v	0.059v	0.076v	0v	0.000024v	0.000026v
	HSv	0.075v	0.075v	0.057v	0.092v	0.062v	0.088v	0v	0.000059v	0.000045v
	Some Collv	0.078v	0.079v	0.070v	0.086v	0.069v	0.089v	1v	0.000022v	0.000034v
	Collegev	0.120v	0.119v	0.102v	0.138v	0.092v	0.146v	0v	0.000095v	0.000240v
	Graduatev	0.116v	0.145v	0.074v	0.159v	0.121v	0.170v	0v	0.000320v	0.000223v

Table 33: Hispanic Indicator

Demographicv Groupv	Educationv Categoryv	Meanv		Confidence Intervalv Syntheticv		Confidence Intervalv Completedv		Syntheticv DF Not Existv	Total Variancev	
		Syntheticv	Completedv						Syntheticv	Completedv
white femalesv	no HSv	0.239v	0.270v	0.227v	0.250v	0.265v	0.275v	0v	0.000016v	0.000009v
	HSv	0.094v	0.091v	0.089v	0.099v	0.089v	0.093v	0v	0.000003v	0.000002v
	Some Collv	0.088v	0.082v	0.083v	0.092v	0.079v	0.084v	0v	0.000003v	0.000003v
	Collegev	0.050v	0.040v	0.047v	0.053v	0.037v	0.043v	1v	0.000004v	0.000003v
	Graduatev	0.037v	0.032v	0.033v	0.040v	0.029v	0.035v	1v	0.000004v	0.000003v
black femalesv	no HSv	0.043v	0.048v	0.035v	0.051v	0.043v	0.053v	0v	0.000016v	0.000010v
	HSv	0.033v	0.034v	0.028v	0.038v	0.030v	0.038v	0v	0.000008v	0.000006v
	Some Collv	0.041v	0.036v	0.030v	0.053v	0.031v	0.040v	0v	0.000028v	0.000008v
	Collegev	0.034v	0.031v	0.023v	0.044v	0.021v	0.040v	0v	0.000035v	0.000032v
	Graduatev	0.033v	0.029v	0.021v	0.044v	0.019v	0.039v	0v	0.000046v	0.000036v
white malesv	no HSv	0.264v	0.288v	0.246v	0.282v	0.283v	0.293v	0v	0.000064v	0.000009v
	HSv	0.116v	0.114v	0.109v	0.122v	0.111v	0.117v	0v	0.000006v	0.000003v
	Some Collv	0.090v	0.090v	0.087v	0.093v	0.087v	0.093v	0v	0.000003v	0.000003v
	Collegev	0.047v	0.040v	0.042v	0.052v	0.038v	0.043v	0v	0.000006v	0.000003v
	Graduatev	0.040v	0.036v	0.035v	0.045v	0.033v	0.039v	0v	0.000006v	0.000003v
black malesv	no HSv	0.046v	0.053v	0.042v	0.050v	0.048v	0.059v	0v	0.000006v	0.000012v
	HSv	0.038v	0.037v	0.034v	0.043v	0.032v	0.041v	0v	0.000008v	0.000008v
	Some Collv	0.037v	0.036v	0.031v	0.043v	0.030v	0.042v	0v	0.000011v	0.000013v
	Collegev	0.046v	0.029v	0.035v	0.058v	0.019v	0.039v	0v	0.000046v	0.000037v
	Graduatev	0.041v	0.030v	0.030v	0.053v	0.018v	0.042v	0v	0.000045v	0.000051v

Table 34: SIPP Disability Indicator

Demographicv Groupv	Educationv Categoryv	Meanv Syntheticv	Completedv	Confidence Intervalv Syntheticv		Confidence Intervalv Completedv		Syntheticv DF Not Existv	Total Variancev Syntheticv	Completedv
white femalesv	no HSv	0.179v	0.186v	0.173v	0.185v	0.181v	0.192v	0v	0.000008v	0.000011v
	HSv	0.121v	0.108v	0.117v	0.125v	0.104v	0.112v	1v	0.000006v	0.000006v
	Some Collv	0.088v	0.081v	0.080v	0.095v	0.077v	0.085v	1v	0.000020v	0.000005v
	Collegev	0.062v	0.050v	0.032v	0.091v	0.046v	0.053v	0v	0.000157v	0.000004v
	Graduatev	0.064v	0.053v	0.061v	0.067v	0.049v	0.057v	1v	0.000003v	0.000006v
black femalesv	no HSv	0.199v	0.223v	0.183v	0.216v	0.203v	0.242v	0v	0.000083v	0.000115v
	HSv	0.125v	0.133v	0.114v	0.136v	0.122v	0.143v	0v	0.000026v	0.000037v
	Some Collv	0.087v	0.086v	0.078v	0.097v	0.079v	0.094v	0v	0.000026v	0.000021v
	Collegev	0.052v	0.032v	0.030v	0.073v	0.021v	0.042v	0v	0.000115v	0.000039v
	Graduatev	0.099v	0.057v	0.065v	0.133v	0.041v	0.073v	0v	0.000276v	0.000096v
white malesv	no HSv	0.164v	0.164v	0.162v	0.167v	0.154v	0.174v	1v	0.000003v	0.000026v
	HSv	0.112v	0.112v	0.111v	0.114v	0.107v	0.116v	1v	0.000001v	0.000006v
	Some Collv	0.089v	0.085v	0.081v	0.097v	0.081v	0.089v	0v	0.000008v	0.000006v
	Collegev	0.051v	0.043v	0.048v	0.055v	0.039v	0.047v	0v	0.000004v	0.000005v
	Graduatev	0.061v	0.048v	0.047v	0.075v	0.044v	0.052v	0v	0.000033v	0.000006v
black malesv	no HSv	0.189v	0.200v	0.175v	0.204v	0.186v	0.213v	0v	0.000032v	0.000065v
	HSv	0.134v	0.146v	0.116v	0.153v	0.135v	0.157v	0v	0.000051v	0.000041v
	Some Collv	0.105v	0.095v	0.090v	0.120v	0.083v	0.107v	0v	0.000057v	0.000052v
	Collegev	0.077v	0.066v	0.044v	0.110v	0.046v	0.087v	0v	0.000242v	0.000139v
	Graduatev	0.148v	0.097v	0.054v	0.241v	0.075v	0.119v	0v	0.001727v	0.000180v

Table 35: Non-deferred DER Earnings at FICA covered jobs in year   000

Demographic Group2	Foreign Born2 Category2	Mean2 Synthetic2	Mean2 Completed2	Confidence Interval2 Synthetic2		Confidence Interval2 Completed2		Synthetic2 DF Not Exist2	Total Variance2 Synthetic2	Total Variance2 Completed2
white female2	yes2	11,8232	10,8282	9,7092	13,9372	10,5132	11,142	02	960,8502	36,442
	no2	13,9992	14,1412	13,7362	14,262	13,7142	14,5682	1	3,9152	67,5162
black female2	yes2	14,232	12,5252	12,7262	15,7382	11,4062	13,6442	12	784,4652	429,8172
	no2	12,3602	11,1282	11,2832	13,4362	10,8362	11,4192	12	401,1142	31,0162
white male2	yes	25,102	4,8082	19,252	30,952	,599	27,0162	02	7,862,8442	1,798,7162
	no2	30,453	29,986	8,12	32,784	8,9142	31,0592	12	1,880,3232	424,8242
black male2	yes2	19,7862	17,812	15,493	24,0792	13,952	1,6732	02	4,399,3062	5,016,7752
	no2	17,5532	15,2752	11,840	23,2662	14,4462	16,1042	02	9,117,175	252,1762

Table 36: Total SER Earnings in year   000

Demographic Group2	Foreign Born2 Category2	Mean2 Synthetic2	Completed2	Confidence Interval2 Synthetic2		Confidence Interval2 Completed2		Synthetic2 DF Not Exist2	Total Variance2 Synthetic2	Completed2
white female2	yes2	10,8832	10,5312	10,5202	11,2452	10,2772	10,7862	12	45,494	23,8932
	no2	13,442	13,3112	13,2482	13,6372	13,192	13,4302	12	13,0812	4,8762
black female2	yes2	13,372	12,6882	13,0952	13,6502	11,7012	13,6762	1	6,6242	350,092
	no2	11,8062	11,0212	11,3802	12,2332	10,7842	11,2582	12	62,905	20,0492
white male2	yes2	19,7742	19,5682	17,461	2,0862	19,133	20,0032	02	931,942	65,4112
	no	23,352	3,494	23,219	23,486	23,335	23,6542	12	6,1692	9,1052
black male2	yes2	16,9302	15,7752	15,8482	18,0112	14,0762	17,4752	02	415,5472	903,5692
	no2	15,1272	14,192	14,652	15,6012	13,8682	14,5162	12	77,832	38,2382

Table 37: Monthly Benefit Amount April 000

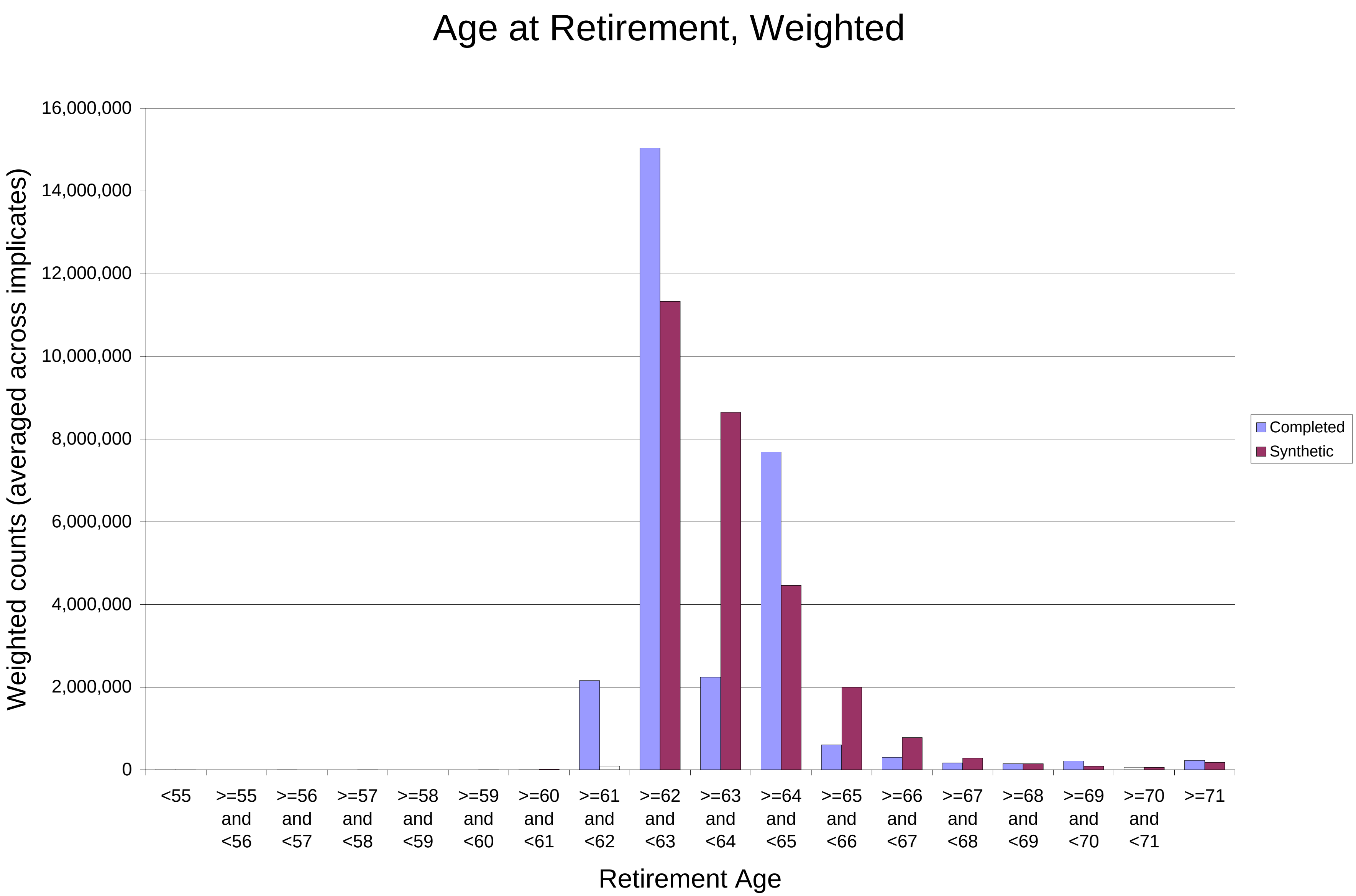
Demographic Group2	Foreign Born2 Category2	Mean2 Synthetic2	Completed2	Confidence Interval2 Synthetic2		Confidence Interval2 Completed2		Synthetic2 DF Not Exist2	Total Variance2 Synthetic2	Completed2
white female2	yes2	5282	5362	5172	5392	5152	5572	02	332	1172
	no2	5802	5952	5742	5872	5912	5992	02	12	52
black female2	yes2	4342	4942	4092	4602	4582	5292	0	32	472
	no2	4602	4582	4512	4682	4492	4682	0	262	302
white male2	yes2	6282	6872	592	6642	6472	7282	0	2952	3892
	no2	7472	7602	742	752	7552	7652	02	92	102
black male2	yes2	4902	5632	4482	532	4992	6282	12	6232	1,3782
	no2	5732	5452	562	5832	5332	5572	02	342	542

Table 38: Marital History Variables

Variable	Mean		Confidence Interval Synthetic		Confidence Interval Completed		Synthetic	Total Variance	
	Synthetic	Completed					DF Not Exist	Synthetic	Completed
Number of marriages	0.88	0.87			0.87	0.87	1	-5.38E-08	9.68E-07
Percent ever divorced	0.24	0.24			0.24	0.24	1	1.15E-07	3.89E-07
Percent ever widowed	0.07	0.07	0.07	0.07	0.07	0.07	0	2.53E-07	1.03E-07
Duration of 1st marriage	23.17	26.18	22.50	23.83	26.07	26.29	0	9.28E-03	2.21E-03
Duration of 2nd marriage	12.96	14.41	11.31	14.61	14.14	14.68	0	0.17	0.01
Age at first marriage	23.05	23.28			23.24	23.32	1	-3.94E-03	3.13E-04

Table 39: Retirement Age – Weighted and Unweighted Counts

	WEIGHTED		UNWEIGHTED	
Age	**Completed**	**Synthetic**	**Completed**	**Synthetic**
<55	25,115	24,089	35	33
>=55 and <56	0	0	0	0
>=56 and <57	1,724	141	1	1
>=57 and <58	0	2,396	1	2
>=58 and <59	268	0	1	1
>=59 and <60	733	3,467	1	5
>=60 and <61	4,686	13,601	11	21
>=61 and <62	2,160,505	92,222	4,919	139
>=62 and <63	15,042,278	11,330,115	21,771	18,046
>=63 and <64	2,243,418	8,648,017	3,820	14,081
>=64 and <65	7,688,675	4,461,798	12,248	7,274
>=65 and <66	605,260	1,998,145	988	3,335
>=66 and <67	302,360	781,017	474	1,388
>=67 and <68	165,783	284,813	305	472
>=68 and <69	148,499	149,679	218	229
>=69 and <70	216,884	89,701	298	137
>=70 and <71	62,259	58,498	106	91
>=71	228,409	182,367	348	289


Age at Retirement, Weighted
Weighted counts (averaged across implicates)
16,000,000
14,000,000
12,000,000
10,000,000
8,000,000
6,000,000
4,000,000
2,000,000
0
<55
>=55 and <56
>=56 and <57
>=57 and <58
>=58 and <59
>=59 and <60
>=60 and <61
>=61 and <62
>=62 and <63
>=63 and <64
>=64 and <65
>=65 and <66
>=66 and <67
>=67 and <68
>=68 and <69
>=69 and <70
>=70 and <71
>=71
Retirement Age
Completed
Synthetic

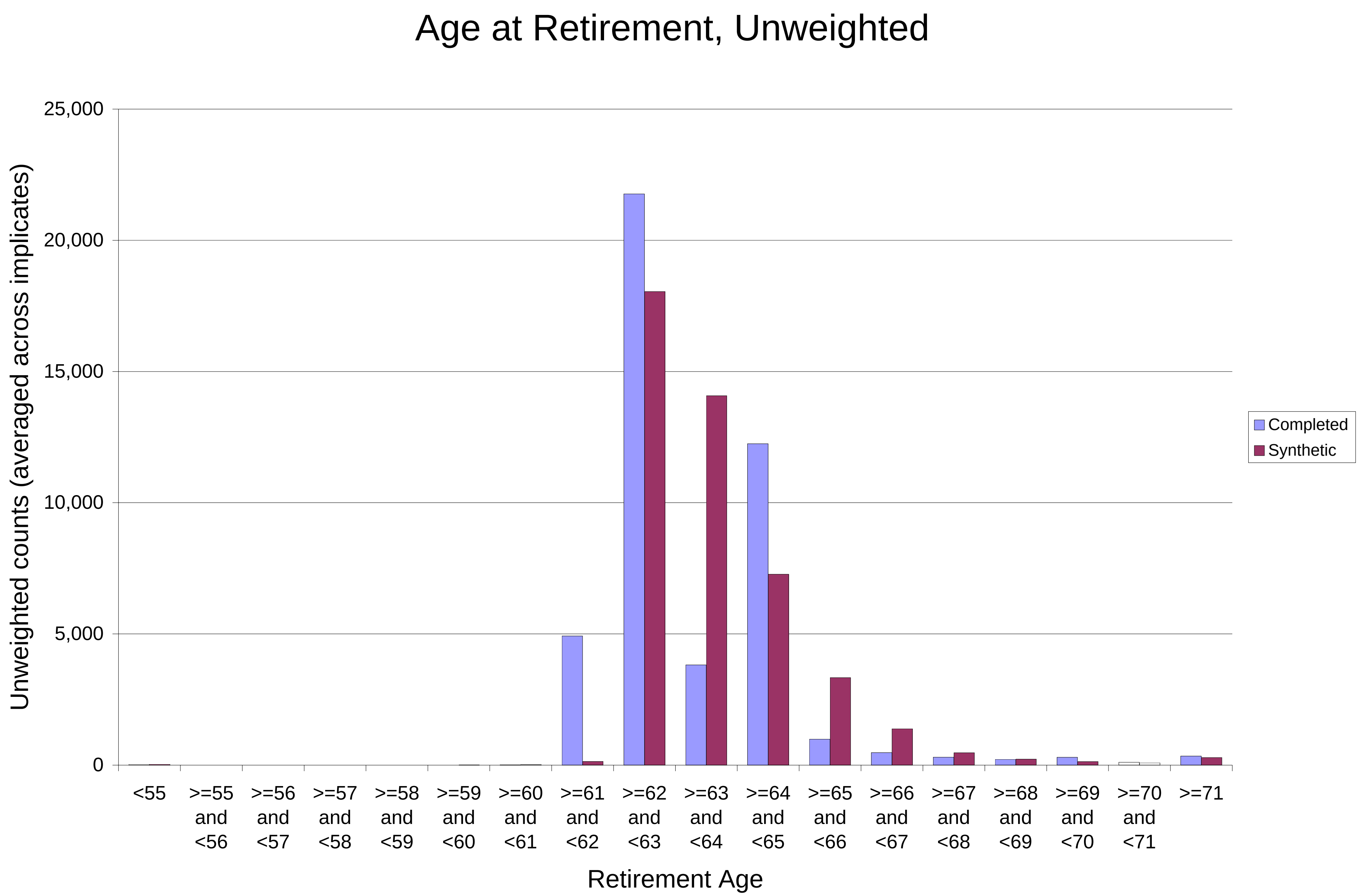

Age at Retirement, Unweighted
Unweighted counts (averaged across implicates)
25,000
20,000
15,000
10,000
5,000
0
<55
>=55 and <56
>=56 and <57
>=57 and <58
>=58 and <59
>=59 and <60
>=60 and <61
>=61 and <62
>=62 and <63
>=63 and <64
>=64 and <65
>=65 and <66
>=66 and <67
>=67 and <68
>=68 and <69
>=69 and <70
>=70 and <71
>=71
Retirement Age
Completed
Synthetic

Table 40: Log of Total DER Earnings in year 2000 for white males

	CoefficientS		Confidence Interval				Standard Error		SyntheticS
Explanatory Variables	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.377S	7.855S	8.266S	8.487S	7.793S	7.917S	0.065S	0.037S	1S
highschool_onlyS	0.214S	0.230S	0.133S	0.294S	0.205S	0.255S	0.036S	0.015S	0S
somecollegeS	0.400S	0.431S	0.263S	0.537S	0.404S	0.457S	0.059S	0.016S	0S
college_onlyS	0.738S	0.880S	0.530S	0.947S	0.851S	0.909S	0.086S	0.017S	0S
graduateS	0.830S	1.110S	0.632S	1.028S	1.080S	1.140S	0.085S	0.018S	0S
disabS	-0.354S	-0.610S	-0.380S	-0.328S	-0.657S	-0.562S	0.014S	0.026S	0S
foreign_bornS	0.064S	0.042S	-0.029S	0.157S	0.013S	0.070S	0.042S	0.017S	0S
hispanicS	-0.072S	-0.013S	-0.113S	-0.031S	-0.040S	0.013S	0.021S	0.016S	0S
ser_totyrs_2000S	0.179S	0.275S	0.142S	0.216S	0.259S	0.292S	0.014S	0.010S	0S
ser_totyrs_2000_2S	-0.073S	-0.140S	-0.085S	-0.062S	-0.153S	-0.128S	0.007S	0.007S	1S
ser_totyrs_2000_3S	0.016S	0.034S	0.013S	0.018S	0.030S	0.038S	0.001S	0.002S	1S
ser_totyrs_2000_4S	-0.001S	-0.003S	-0.002S	-0.001S	-0.004S	-0.003S	0.000S	0.000S	1S

Table 41: Log of Total DER Earnings in year 2000 for black males

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.080S	7.070S	7.929S	8.230S	6.885S	7.254S	0.089S	0.108S	1S
highschool_onlyS	0.163S	0.322S	-0.031S	0.357S	0.231S	0.413S	0.090S	0.053S	0S
somecollegeS	0.375S	0.551S	0.204S	0.546S	0.476S	0.627S	0.074S	0.046S	0S
college_onlyS	0.680S	0.860S	0.415S	0.945S	0.735S	0.985S	0.124S	0.075S	0S
graduateS	0.797S	1.169S	0.461S	1.133S	1.018S	1.320S	0.156S	0.091S	0S
disabS	-0.400S	-0.631S	-0.533S	-0.267S	-0.763S	-0.499S	0.062S	0.075S	0S
foreign_bornS	0.082S	0.046S	-0.098S	0.262S	-0.106S	0.197S	0.084S	0.084S	0S
hispanicS	-0.030S	0.156S	-0.128S	0.067S	0.017S	0.296S	0.051S	0.084S	0S
ser_totyrs_2000S	0.173S	0.388S	0.154S	0.191S	0.336S	0.440S	0.011S	0.030S	1S
ser_totyrs_2000_2S	-0.067S	-0.240S	-0.078S	-0.055S	-0.284S	-0.197S	0.007S	0.025S	1S
ser_totyrs_2000_3S	0.013S	0.067S	0.009S	0.018S	0.053S	0.080S	0.003S	0.008S	1S
ser_totyrs_2000_4S	-0.001S	-0.007S	-0.002S	-0.001S	-0.008S	-0.005S	0.000S	0.001S	1S

Table 42: Log of Total DER Earnings in year 2000 for white males

Explanatory VariableS	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.373S	7.717S	8.295S	8.451S	7.658S	7.777S	0.046S	0.036S	1S
highschool_onlyS	0.180S	0.248S	0.157S	0.203S	0.221S	0.274S	0.013S	0.016S	1S
somecollegeS	0.405S	0.491S	0.284S	0.527S	0.463S	0.520S	0.051S	0.017S	0S
college_onlyS	0.719S	0.834S	0.561S	0.876S	0.802S	0.865S	0.063S	0.019S	0S
graduateS	0.790S	1.043S	0.628S	0.952S	1.008S	1.078S	0.064S	0.021S	0S
disabS	-0.259S	-0.470S	-0.296S	-0.222S	-0.511S	-0.428S	0.022S	0.024S	1S
foreign_bornS	0.075S	0.097S	0.032S	0.117S	0.062S	0.132S	0.025S	0.020S	1S
hispanicS	-0.008S	0.043S	-0.049S	0.033S	0.007S	0.079S	0.022S	0.021S	0S
ser_totyrs_2000S	0.104S	0.216S	0.039S	0.169S	0.201S	0.230S	0.025S	0.009S	0S
ser_totyrs_2000_2S	-0.038S	-0.119S	-0.091S	0.014S	-0.131S	-0.107S	0.020S	0.007S	0S
ser_totyrs_2000_3S	0.009S	0.032S	-0.007S	0.025S	0.029S	0.036S	0.006S	0.002S	0S
ser_totyrs_2000_4S	-0.001S	-0.003S	-0.003S	0.001S	-0.004S	-0.003S	0.001S	0.000S	0S

Table 43: Log of Total DER Earnings in year 2000 for white males

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.236S	7.194S	8.101S	8.372S	7.059S	7.329S	0.080S	0.082S	1S
highschool_onlyS	0.263S	0.347S	0.197S	0.330S	0.282S	0.412S	0.032S	0.039S	0S
somecollegeS	0.494S	0.587S	0.376S	0.612S	0.518S	0.656S	0.046S	0.041S	0S
college_onlyS	0.842S	1.118S	0.770S	0.914S	1.030S	1.207S	0.042S	0.054S	1S
graduateS	0.953S	1.289S	0.748S	1.157S	1.179S	1.399S	0.077S	0.067S	0S
disabS	-0.300S	-0.453S	-0.408S	-0.192S	-0.543S	-0.363S	0.052S	0.054S	0S
foreign_bornS	0.114S	0.183S	-0.016S	0.245S	0.050S	0.317S	0.066S	0.073S	0S
hispanicS	-0.011S	-0.007S	-0.124S	0.101S	-0.113S	0.099S	0.064S	0.064S	0S
ser_totyrs_2000S	0.086S	0.297S	-0.010S	0.183S	0.260S	0.335S	0.037S	0.023S	0S
ser_totyrs_2000_2S	-0.019S	-0.175S	-0.092S	0.054S	-0.207S	-0.144S	0.031S	0.019S	0S
ser_totyrs_2000_3S	0.003S	0.048S	-0.019S	0.024S	0.038S	0.059S	0.010S	0.006S	0S
ser_totyrs_2000_4S	0.000S	-0.005S	-0.003S	0.002S	-0.006S	-0.004S	0.001S	0.001S	0S

Table 44: Log of ER Earnings in year 2000 for white males

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.277S	7.807S	8.135S	8.419S	7.746S	7.867S	0.084S	0.036S	1S
highschool_onlyS	0.190S	0.198S	0.102S	0.278S	0.175S	0.221S	0.038S	0.014S	0S
somecollegeS	0.357S	0.366S	0.246S	0.467S	0.342S	0.390S	0.048S	0.015S	0S
college_onlyS	0.596S	0.700S	0.430S	0.762S	0.672S	0.727S	0.071S	0.016S	0S
graduateS	0.638S	0.794S	0.458S	0.819S	0.765S	0.823S	0.077S	0.018S	0S
disabS	-0.340S	-0.580S	-0.364S	-0.316S	-0.634S	-0.527S	0.014S	0.028S	0S
foreign_bornS	0.015S	0.042S	-0.086S	0.116S	0.017S	0.067S	0.045S	0.015S	0S
hispanicS	-0.041S	-0.028S	-0.090S	0.007S	-0.053S	-0.003S	0.024S	0.015S	0S
ser_totyrs_2000S	0.194S	0.282S	0.150S	0.238S	0.266S	0.298S	0.016S	0.009S	0S
ser_totyrs_2000_2S	-0.087S	-0.148S	-0.102S	-0.071S	-0.160S	-0.136S	0.009S	0.007S	1S
ser_totyrs_2000_3S	0.019S	0.036S	0.016S	0.023S	0.033S	0.040S	0.002S	0.002S	1S
ser_totyrs_2000_4S	-0.002S	-0.003S	-0.002S	-0.001S	-0.004S	-0.003S	0.000S	0.000S	1S
marriedS	0.111S	0.110S	0.078S	0.145S	0.088S	0.132S	0.016S	0.013S	0S
divorcedS	-0.072S	-0.042S	-0.094S	-0.050S	-0.073S	-0.010S	0.012S	0.019S	0S
widowedS	-0.357S	-0.423S	-0.599S	-0.115S	-0.545S	-0.302S	0.120S	0.074S	0S

Table 45: Log of ER earnings in year 2000 for black males

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.022S	7.106S	7.861S	8.183S	6.931S	7.281S	0.095S	0.105S	1S
highschool_onlyS	0.130S	0.292S	-0.049S	0.308S	0.194S	0.390S	0.084S	0.056S	0S
somecollegeS	0.317S	0.500S	0.186S	0.449S	0.421S	0.579S	0.058S	0.048S	0S
college_onlyS	0.557S	0.658S	0.329S	0.785S	0.525S	0.792S	0.109S	0.080S	0S
graduateS	0.598S	0.872S	0.333S	0.864S	0.717S	1.027S	0.129S	0.094S	0S
disabS	-0.367S	-0.571S	-0.515S	-0.218S	-0.702S	-0.439S	0.067S	0.076S	0S
foreign_bornS	0.032S	0.066S	-0.140S	0.204S	-0.096S	0.228S	0.079S	0.090S	0S
hispanicS	0.001S	0.187S	-0.107S	0.108S	0.050S	0.325S	0.056S	0.083S	0S
ser_totyrs_2000S	0.172S	0.359S	0.131S	0.213S	0.311S	0.407S	0.024S	0.029S	1S
ser_totyrs_2000_2S	-0.066S	-0.215S	-0.098S	-0.034S	-0.255S	-0.175S	0.019S	0.024S	1S
ser_totyrs_2000_3S	0.013S	0.059S	0.003S	0.023S	0.047S	0.072S	0.006S	0.008S	1S
ser_totyrs_2000_4S	-0.001S	-0.006S	-0.002S	0.000S	-0.007S	-0.005S	0.001S	0.001S	1S
marriedS	0.114S	0.097S	0.057S	0.171S	0.022S	0.172S	0.034S	0.045S	0S
divorcedS	-0.187S	-0.190S	-0.278S	-0.095S	-0.340S	-0.040S	0.053S	0.085S	0S
widowedS	-0.311S	-0.492S	-0.716S	0.093S	-0.999S	0.016S	0.223S	0.293S	0S

Table 46: Log of ER Earnings in year 2000 for white females

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.300S	7.641S	8.216S	8.385S	7.579S	7.702S	0.050S	0.037S	1S
highschool_onlyS	0.168S	0.242S	0.134S	0.202S	0.215S	0.269S	0.020S	0.016S	1S
somecollegeS	0.374S	0.431S	0.244S	0.503S	0.403S	0.460S	0.055S	0.017S	0S
college_onlyS	0.646S	0.710S	0.471S	0.820S	0.677S	0.742S	0.071S	0.020S	0S
graduateS	0.686S	0.833S	0.531S	0.840S	0.797S	0.869S	0.059S	0.022S	0S
disabS	-0.249S	-0.445S	-0.302S	-0.195S	-0.486S	-0.403S	0.019S	0.024S	0S
foreign_bornS	0.076S	0.194S	0.006S	0.147S	0.159S	0.229S	0.025S	0.020S	0S
hispanicS	0.019S	0.030S	-0.027S	0.066S	-0.010S	0.070S	0.024S	0.023S	0S
ser_totyrs_2000S	0.111S	0.229S	0.043S	0.179S	0.214S	0.244S	0.027S	0.009S	0S
ser_totyrs_2000_2S	-0.028S	-0.109S	-0.081S	0.025S	-0.121S	-0.096S	0.021S	0.008S	0S
ser_totyrs_2000_3S	0.004S	0.027S	-0.011S	0.020S	0.023S	0.031S	0.006S	0.002S	0S
ser_totyrs_2000_4S	-0.001S	-0.003S	-0.002S	0.001S	-0.003S	-0.002S	0.001S	0.000S	0S
marriedS	-0.241S	-0.342S	-0.275S	-0.208S	-0.370S	-0.314S	0.020S	0.016S	1S
divorcedS	-0.190S	-0.243S	-0.265S	-0.116S	-0.276S	-0.210S	0.029S	0.020S	0S
widowedS	-0.473S	-0.563S	-0.571S	-0.376S	-0.636S	-0.491S	0.049S	0.042S	0S

Table 47: Log of ER Earnings in year 2000 for black females

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	8.107S	7.150S	7.970S	8.244S	7.006S	7.294S	0.081S	0.087S	1S
highschool_onlyS	0.230S	0.307S	0.152S	0.307S	0.244S	0.370S	0.036S	0.038S	0S
somecollegeS	0.439S	0.493S	0.325S	0.553S	0.424S	0.562S	0.047S	0.041S	0S
college_onlyS	0.733S	0.913S	0.625S	0.842S	0.818S	1.007S	0.064S	0.057S	1S
graduateS	0.811S	1.013S	0.619S	1.002S	0.899S	1.126S	0.085S	0.069S	0S
disabS	-0.284S	-0.407S	-0.375S	-0.193S	-0.501S	-0.312S	0.047S	0.056S	0S
foreign_bornS	0.094S	0.229S	-0.053S	0.241S	0.102S	0.357S	0.072S	0.071S	0S
hispanicS	0.008S	0.033S	-0.109S	0.125S	-0.074S	0.139S	0.067S	0.065S	0S
ser_totyrs_2000S	0.102S	0.292S	0.041S	0.162S	0.252S	0.333S	0.036S	0.024S	1S
ser_totyrs_2000_2S	-0.021S	-0.159S	-0.098S	0.056S	-0.193S	-0.126S	0.029S	0.020S	0S
ser_totyrs_2000_3S	0.003S	0.043S	-0.019S	0.024S	0.032S	0.054S	0.009S	0.007S	0S
ser_totyrs_2000_4S	0.000S	-0.004S	-0.002S	0.002S	-0.006S	-0.003S	0.001S	0.001S	0S
marriedS	-0.073S	-0.091S	-0.126S	-0.021S	-0.157S	-0.025S	0.031S	0.039S	0S
divorcedS	-0.170S	-0.214S	-0.218S	-0.123S	-0.294S	-0.133S	0.028S	0.047S	0S
widowedS	-0.435S	-0.469S	-0.606S	-0.264S	-0.601S	-0.336S	0.094S	0.081S	0S

Table 48: Log of Average Indexed Monthly Earnings (AIME) or Average Monthly Wage (AMW) for all individuals

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	7.604S	7.252S	7.554S	7.654S	7.170S	7.335S	0.029S	0.048S	1S
age_2000S	0.0004S	0.0093S	-0.0067S	0.0075S	0.0055S	0.0131S	0.003S	0.002S	0S
age_2000_sqS	-0.0002S	-0.0003S	-0.0003	-0.0001S	0.0003	-0.0002S	0.000S	0.000S	0S
blackfemaleS	-0.928S	-0.995S	-0.949S	-0.906S	-1.019S	-0.972S	0.010S	0.014S	0S
blackmaleS	-0.403S	-0.457S	-0.444S	-0.362S	-0.499S	-0.415S	0.019S	0.022S	0S
whitefemaleS	-0.822S	-0.843S	-0.836S	-0.807S	-0.853S	-0.832S	0.007S	0.006S	0S
highschool_onlyS	0.337S	0.400S	0.235S	0.438S	0.382S	0.417S	0.043S	0.010S	0S
somecollegeS	0.570S	0.690S	0.441S	0.699S	0.673S	0.708S	0.055S	0.010S	0S
college_onlyS	0.717S	0.866S	0.571S	0.862S	0.840S	0.891S	0.062S	0.014S	0S
graduateS	0.748S	0.911S	0.641S	0.855S	0.879S	0.942S	0.046S	0.017S	0S
disabS	-0.365S	-0.559S	-0.488S	-0.241S	-0.580S	-0.538S	0.053S	0.012S	0S
hispanicS	-0.249S	-0.257S	-0.280S	-0.218S	-0.276S	-0.237S	0.014S	0.011S	0S
divorcedS	0.136S	0.159S	0.108S	0.164S	0.118S	0.200S	0.015S	0.021S	0S
marriedS	0.134S	0.132S	0.105S	0.162S	0.099S	0.165S	0.014S	0.017S	0S
widowedS	-0.106S	-0.024S	-0.145S	-0.067S	-0.062S	0.014S	0.022S	0.023S	0S

*AIME for individuals who turned 62 after 1979, AMW otherwise

Table 49: Log of initial MBA for retired individuals (TOB_initial=1)

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	-61.534S	-67.501S	-66.344	-56.725S	-71.471S	-63.531S	2.501S	2.192S	0S
age_initial_entitleS	0.033S	0.038S	0.028S	0.038S	0.033S	0.044S	0.003S	0.003S	0S
blackfemaleS	-0.360S	-0.329S	-0.435S	-0.285S	-0.386S	-0.272S	0.036S	0.030S	0S
blackmaleS	-0.110S	-0.120S	-0.150S	-0.071S	-0.150S	-0.089S	0.015S	0.018S	0S
whitefemaleS	-0.301S	-0.297S	-0.364S	-0.238S	-0.354S	-0.240S	0.027S	0.026S	0S
highschool_onlyS	0.070S	0.061S	0.042S	0.097S	0.026S	0.096S	0.014S	0.017S	0S
somecollegeS	0.121S	0.089S	0.078S	0.163S	0.054S	0.124S	0.020S	0.018S	0S
college_onlyS	0.164S	0.124S	0.143S	0.184S	0.080S	0.168S	0.011S	0.022S	0S
graduateS	0.191S	0.147S	0.150S	0.232S	0.119S	0.175S	0.020S	0.016S	0S
disabS	-0.048S	-0.039S	-0.076S	-0.021S	-0.067S	-0.011S	0.013S	0.015S	0S
hispanicS	-0.098S	-0.058S	-0.161S	-0.035S	-0.124S	0.009S	0.029S	0.032S	0S
divorcedS	0.114S	0.132S	0.069S	0.159S	0.098S	0.166S	0.023S	0.020S	0S
marriedS	0.078S	0.052S	0.052S	0.104S	0.019S	0.085S	0.015S	0.019S	0S
widowedS	0.179S	0.162S	0.146S	0.213S	0.126S	0.197S	0.020S	0.021S	0S
log_totnetworthS	0.015S	0.046S	0.005S	0.025S	0.037S	0.055S	0.005S	0.005S	0S
ser_pct_yrs_wrkedS	1.052S	1.044S	0.609S	1.496S	0.689S	1.398S	0.187S	0.151S	0S
year_initial_entitleS	0.033S	0.035S	0.030S	0.035S	0.033S	0.037S	0.001S	0.001S	0S

Table 50: Log of initial MBA for disabled individuals (TOB_initial=2)

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	-76.179S	-75.378S	-80.608S	-71.751S	-80.502S	-70.253S	2.283S	2.663S	0S
age_initial_entitleS	0.010S	0.010S	0.009S	0.011S	0.009S	0.012S	0.001S	0.001S	0S
blackfemaleS	-0.299S	-0.255S	-0.353S	-0.244S	-0.321S	-0.188S	0.029S	0.035S	0S
blackmaleS	-0.055S	-0.022S	-0.093S	-0.018S	-0.059S	0.015S	0.021S	0.022S	0S
whitefemaleS	-0.328S	-0.326S	-0.347S	-0.309S	-0.348S	-0.304S	0.011S	0.013S	0S
highschool_onlyS	0.070S	0.143S	0.028S	0.113S	0.106S	0.180S	0.021S	0.020S	0S
somecollegeS	0.139S	0.201S	0.102S	0.176S	0.168S	0.233S	0.020S	0.019S	0S
college_onlyS	0.213S	0.287S	0.164S	0.262S	0.225S	0.349S	0.023S	0.034S	0S
graduateS	0.233S	0.343S	0.192S	0.274S	0.287S	0.399S	0.023S	0.033S	0S
disabS	-0.045S	-0.004S	-0.068S	-0.022S	-0.025S	0.017S	0.012S	0.013S	0S
hispanicS	-0.058S	-0.047S	-0.091S	-0.025S	-0.096S	0.001S	0.019S	0.027S	0S
divorcedS	0.099S	0.101S	0.067S	0.131S	0.070S	0.133S	0.019S	0.019S	0S
marriedS	0.125S	0.133S	0.102S	0.149S	0.097S	0.170S	0.014S	0.021S	0S
widowedS	0.046S	0.067S	0.003S	0.088S	0.018S	0.116S	0.025S	0.030S	0S
log_totnetworthS	0.005S	0.016S	-0.002S	0.011S	0.010S	0.022S	0.003S	0.004S	0S
ser_pct_yrs_wrkedS	0.542S	0.536S	0.351S	0.734S	0.388S	0.683S	0.085S	0.069S	0S
year_initial_entitleS	0.041S	0.040S	0.039S	0.043S	0.038S	0.043S	0.001S	0.001S	0S

Table 51: Log of MBA 2000 for retired individuals (TOB_2000=1)

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	21.365S	16.985S	11.656S	31.074S	7.475S	26.496S	4.487S	4.459S	0S
age_2000S	0.009S	0.009S	0.005S	0.012S	0.007S	0.011S	0.002S	0.001S	0S
blackfemaleS	-0.272S	-0.250S	-0.339S	-0.206S	-0.301S	-0.199S	0.032S	0.026S	0S
blackmaleS	-0.095S	-0.092S	-0.122S	-0.068S	-0.127S	-0.056S	0.014S	0.020S	0S
whitefemaleS	-0.211S	-0.192S	-0.284S	-0.139S	-0.257S	-0.126S	0.032S	0.029S	0S
highschool_onlyS	0.054S	0.054S	0.039S	0.068S	0.024S	0.084S	0.007S	0.015S	0S
somecollegeS	0.098S	0.078S	0.075S	0.121S	0.052S	0.104S	0.012S	0.014S	0S
college_onlyS	0.136S	0.128S	0.103S	0.169S	0.084S	0.172S	0.016S	0.022S	0S
graduateS	0.157S	0.150S	0.132S	0.182S	0.125S	0.175S	0.013S	0.015S	0S
disabS	-0.019S	0.001S	-0.035S	-0.003S	-0.019S	0.021S	0.009S	0.011S	0S
hispanicS	-0.102S	-0.059S	-0.164S	-0.040S	-0.108S	-0.009S	0.028S	0.025S	0S
divorcedS	0.110S	0.126S	0.071S	0.148S	0.075S	0.177S	0.021S	0.027S	0S
marriedS	0.107S	0.081S	0.073S	0.142S	0.051S	0.110S	0.019S	0.017S	0S
widowedS	0.232S	0.217S	0.187S	0.278S	0.171S	0.264S	0.024S	0.026S	0S
famwelpart1999S	-0.048S	-0.022S	-0.103S	0.008S	-0.058S	0.013S	0.026S	0.019S	0S
hicovannual1999S	0.023S	-0.001S	0.013S	0.033S	-0.012S	0.010S	0.006S	0.007S	0S
log_totnetworthS	0.012S	0.040S	0.004S	0.021S	0.035S	0.046S	0.004S	0.003S	0S
ser_pct_yrs_wrkedS	0.948S	1.001S	0.473S	1.423S	0.593S	1.409S	0.202S	0.174S	0S
year_initial_entitleS	-0.008S	-0.006S	-0.013S	-0.003S	-0.011S	-0.001S	0.002S	0.002S	0S

Table 52: Log of MBA 2000 for disabled individuals (TOB_2000=2)

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	-15.944S	-18.763S	-19.131	-12.758S	-22.303S	-15.222S	1.849S	2.129S	0S
age_2000S	0.010S	0.010S	0.009S	0.011S	0.009S	0.012S	0.001S	0.001S	0S
blackfemaleS	-0.260S	-0.226S	-0.319S	-0.201S	-0.284S	-0.168S	0.033S	0.033S	0S
blackmaleS	-0.048S	-0.008S	-0.086S	-0.011S	-0.056S	0.041S	0.022S	0.028S	0S
whitefemaleS	-0.293S	-0.295S	-0.314S	-0.271S	-0.326S	-0.264S	0.013S	0.018S	0S
highschool_onlyS	0.065S	0.111S	0.024S	0.106S	0.073S	0.149S	0.019S	0.022S	0S
somecollegeS	0.126S	0.175S	0.093S	0.158S	0.141S	0.209S	0.018S	0.020S	0S
college_onlyS	0.189S	0.280S	0.152S	0.226S	0.204S	0.357S	0.021S	0.043S	0S
graduateS	0.207S	0.280S	0.161S	0.253S	0.215S	0.345S	0.026S	0.039S	0S
disabS	-0.040S	0.019S	-0.072S	-0.008S	-0.007S	0.045S	0.016S	0.015S	0S
hispanicS	-0.034S	-0.057S	-0.073S	0.005S	-0.097S	-0.017S	0.023S	0.024S	0S
divorcedS	0.037S	0.029S	-0.025S	0.099S	-0.020S	0.077S	0.031S	0.028S	0S
marriedS	0.067S	0.065S	0.029S	0.104S	0.014S	0.115S	0.021S	0.028S	0S
widowedS	-0.009S	0.028S	-0.076S	0.059S	-0.048S	0.103S	0.039S	0.045S	0S
famwelpart1999S	-0.032S	-0.017S	-0.072S	0.008S	-0.059S	0.025S	0.022S	0.024S	0S
hicovannual1999S	0.038S	0.017S	0.007S	0.069S	-0.013S	0.046S	0.017S	0.018S	0S
log_totnetworthS	0.002S	0.019S	-0.006S	0.010S	0.012S	0.026S	0.004S	0.004S	0S
ser_pct_yrs_wrkedS	0.413S	0.379S	0.318S	0.507S	0.288S	0.470S	0.048S	0.049S	0S
year_initial_entitleS	0.011S	0.012S	0.009S	0.013S	0.010S	0.014S	0.001S	0.001S	0S

Table 53: Log of Total family income in year 1999, all individuals

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	-0.491S	3.914S	-1.852S	0.870S	-0.420S	8.249S	0.801S	1.905S	1S
blackfemaleS	-0.354S	-0.450S	-0.418S	-0.291S	-0.545S	-0.356S	0.038S	0.041S	1S
blackmaleS	-0.239S	-0.293S	-0.349S	-0.129S	-0.620S	0.034S	0.065S	0.139S	1S
whitefemaleS	-0.156S	-0.171S	-0.254S	-0.058S	-0.504S	0.161S	0.058S	0.142S	1S
highschool_onlyS	0.118S	0.211S	-0.040S	0.275S	0.059S	0.364S	0.055S	0.065S	0S
somecollegeS	0.252S	0.402S	0.164S	0.340S	0.282S	0.522S	0.052S	0.051S	1S
college_onlyS	0.354S	0.715S	0.321S	0.387S	0.446S	0.983S	0.020S	0.114S	1S
graduateS	0.412S	0.895S	0.375S	0.449S	0.490S	1.300S	0.022S	0.172S	1S
disabS	-0.121S	-0.252S	-0.151S	-0.091S	-0.288S	-0.215S	0.013S	0.017S	0S
foreign_bornS	-0.051S	-0.060S	-0.092S	-0.011S	-0.098S	-0.022S	0.018S	0.017S	0S
hispanicS	-0.117S	-0.150S	-0.186S	-0.048S	-0.220S	-0.081S	0.030S	0.031S	0S
divorcedS	-0.190S	-0.176S	-0.248S	-0.132S	-0.800S	0.447S	0.034S	0.265S	1S
marriedS	0.485S	0.420S	0.407S	0.564S	0.206S	0.634S	0.046S	0.091S	1S
widowedS	-0.031S	-0.035S	-0.091S	0.030S	-0.443S	0.374S	0.036S	0.174S	1S
retiredS	-0.236S	-0.141S	-0.316S	-0.156S	-0.171S	-0.111S	0.029S	0.015S	0S
disabledssaS	-0.353S	-0.202S	-0.412S	-0.295S	-0.245S	-0.158S	0.025S	0.022S	0S
agedspouseS	-0.291S	-0.248S	-0.351S	-0.230S	-0.434S	-0.061S	0.036S	0.083S	1S
widowspouseS	-0.305S	-0.244S	-0.376S	-0.234S	-0.349S	-0.139S	0.028S	0.050S	0S
otherbenefitS	-0.111S	-0.092S	-0.145S	-0.077S	-0.131S	-0.053S	0.016S	0.018S	0S
totfam_kidsS	-0.054S	0.027S	-0.057S	-0.051S	0.009S	0.045S	0.002S	0.008S	1S
year_birthS	0.006S	0.003S	0.005S	0.006S	0.001S	0.005S	0.000S	0.001S	1S

Table 54: Log of total personal income in year 1999 for all individuals

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	40.724S	39.397S	36.384S	45.064S	31.231S	47.564S	2.553S	3.529S	1S
blackfemaleS	-0.663S	-0.574S	-0.873S	-0.454S	-0.716S	-0.433S	0.073S	0.061S	0S
blackmaleS	-0.381S	-0.326S	-0.411S	-0.350S	-0.354S	-0.299S	0.018S	0.016S	1S
whitefemaleS	-0.705S	-0.693S	-0.913S	-0.497S	-1.021S	-0.365S	0.086S	0.140S	0S
highschool_onlyS	0.247S	0.299S	0.140S	0.354S	0.260S	0.338S	0.042S	0.018S	0S
somecollegeS	0.423S	0.452S	0.302S	0.543S	0.387S	0.518S	0.048S	0.029S	0S
college_onlyS	0.722S	0.922S	0.502S	0.941S	0.684S	1.160S	0.087S	0.102S	0S
graduateS	0.787S	1.165S	0.537S	1.036S	0.835S	1.495S	0.097S	0.141S	0S
disabS	-0.155S	-0.313S	-0.188S	-0.121S	-0.356S	-0.270S	0.015S	0.020S	0S
foreign_bornS	-0.063S	-0.101S	-0.116S	-0.010S	-0.125S	-0.077S	0.021S	0.013S	0S
hispanicS	-0.149S	-0.058S	-0.229S	-0.069S	-0.087S	-0.029S	0.033S	0.015S	0S
divorcedS	0.254S	0.223S	-0.018S	0.527S	0.039S	0.407S	0.096S	0.079S	0S
marriedS	0.246S	0.211S	0.214S	0.278S	0.095S	0.326S	0.019S	0.050S	1S
widowedS	0.233S	0.335S	0.187S	0.278S	0.142S	0.529S	0.027S	0.085S	1S
retiredS	-0.622S	-0.527S	-0.742S	-0.502S	-0.639S	-0.415S	0.051S	0.049S	0S
disabledssaS	-0.405S	-0.273S	-0.470S	-0.340S	-0.351S	-0.194S	0.029S	0.037S	0S
agedspouseS	-1.011S	-1.017S	-1.187S	-0.835S	-1.173S	-0.862S	0.074S	0.073S	0S
widowspouseS	-0.658S	-0.634S	-0.830S	-0.487S	-0.882S	-0.386S	0.076S	0.110S	0S
otherbenefitS	-0.054S	-0.005S	-0.095S	-0.013S	-0.052S	0.043S	0.019S	0.023S	0S
totfam_kidsS	0.004S	-0.037S	-0.020S	0.028S	-0.053S	-0.021S	0.009S	0.007S	0S
year_birthS	-0.016S	-0.015S	-0.018S	-0.014S	-0.019S	-0.011S	0.001S	0.002S	1S

Table 55: Indicator for whether individual has either a DB or DC pension, all individuals age/employment eligible f

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	-4.573S	-4.999S	-4.633S	-4.514S	-5.227S	-4.771S	0.035S	0.097S	1S
age_2000S	0.049S	0.079S	0.043S	0.055S	0.065S	0.093S	0.004S	0.006S	1S
age_2000_sqS	-0.0003S	-0.0006S	-0.0004	-0.0002S	-0.0008S	-0.0005S	0.000S	0.000S	1S
blackfemaleS	-0.112S	-0.125S	-0.281S	0.057S	-0.319S	0.070S	0.064S	0.083S	0S
blackmaleS	0.077S	0.046S	0.071S	0.083S	-0.068S	0.161S	0.003S	0.049S	1S
whitefemaleS	-0.213S	-0.246S	-0.284S	-0.142S	-0.314S	-0.178S	0.028S	0.029S	0S
highschool_onlyS	0.304S	0.387S	0.217S	0.391S	0.376S	0.398S	0.031S	0.005S	0S
somecollegeS	0.492S	0.579S	0.368S	0.617S	0.528S	0.629S	0.048S	0.022S	0S
college_onlyS	0.744S	0.804S	0.612S	0.877S	0.742S	0.865S	0.050S	0.026S	0S
graduateS	0.656S	0.678S	0.525S	0.788S	0.666S	0.690S	0.047S	0.005S	0S
disabS	-0.063S	-0.205S	-0.123S	-0.003S	-0.308S	-0.102S	0.035S	0.044S	1S
hispanicS	-0.187S	-0.261S	-0.291S	-0.083S	-0.324S	-0.199S	0.041S	0.027S	0S
divorcedS	0.084S	0.121S	0.047S	0.120S	0.063S	0.178S	0.021S	0.024S	1S
marriedS	0.198S	0.257S	0.178S	0.218S	0.209S	0.305S	0.012S	0.020S	1S
widowedS	-0.063S	0.016S	-0.135S	0.009S	-0.203S	0.234S	0.042S	0.093S	1S
ltotearn_ser_2000S	0.254S	0.229S	0.244S	0.264S	0.208S	0.251S	0.006S	0.009S	1S
managerialS	0.177S	0.315S	0.087S	0.268S	0.295S	0.335S	0.032S	0.009S	0S
tech_supportS	0.085S	0.208S	0.039S	0.132S	0.191S	0.226S	0.027S	0.007S	1S
manufacturingS	0.144S	0.292S	0.135S	0.153S	0.231S	0.352S	0.005S	0.026S	1S
retailS	0.011S	-0.349S	-0.025S	0.047S	-0.407S	-0.291S	0.021S	0.025S	1S
servicesS	0.041S	-0.063S	0.016S	0.066S	-0.123S	-0.003S	0.015S	0.026S	1S

Table 56: Indicator for positive weeks with pay in year 1999, all individuals

	CoefficientS		Confidence Interval				Standard Error		SyntheticS
Explanatory Variables	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	0.227S	0.061S	0.186S	0.268S	-0.255S	0.378S	0.024S	0.134S	1S
blackfemaleS	-0.404S	-0.393S	-0.547S	-0.262S	-0.504S	-0.281S	0.052S	0.047S	0S
blackmaleS	-0.389S	-0.486S	-0.548S	-0.230S	-0.591S	-0.382S	0.059S	0.045S	0S
whitefemaleS	-0.346S	-0.387S	-0.360S	-0.332S	-0.426S	-0.348S	0.008S	0.017S	1S
highschool_onlyS	0.571S	0.795S	0.547S	0.595S	0.630S	0.960S	0.014S	0.070S	1S
somecollegeS	0.775S	0.921S	0.749S	0.800S	0.788S	1.054S	0.015S	0.057S	1S
college_onlyS	0.961S	1.245S	0.902S	1.019S	1.150S	1.340S	0.035S	0.040S	1S
graduateS	0.919S	1.214S	0.860S	0.978S	1.123S	1.305S	0.035S	0.039S	1S
totfam_kidsS	0.161S	0.155S	0.123S	0.199S	0.127S	0.182S	0.015S	0.012S	0S
foreign_bornS	0.008S	0.095S	-0.127S	0.143S	-0.017S	0.206S	0.056S	0.047S	0S
hispanicS	0.146S	0.176S	0.115S	0.176S	0.090S	0.262S	0.018S	0.037S	1S
pos_mba2000S	-1.148S	-1.525S	-1.167S	-1.128S	-1.601S	-1.448S	0.011S	0.033S	1S
der_avg_log_real_ear	0.024S	-0.011S	0.020S	0.029S	-0.020S	-0.003S	0.003S	0.004S	1S
qder2S	0.199S	0.292S	0.162S	0.236S	0.194S	0.391S	0.022S	0.042S	1S
qder3S	0.323S	0.439S	0.201S	0.445S	0.341S	0.536S	0.046S	0.041S	0S
qder4S	0.257S	0.278S	0.087S	0.427S	0.151S	0.405S	0.069S	0.054S	0S
ser_pct_yrs_wrkedS	0.307S	1.179S	-0.273S	0.888S	0.188S	2.170S	0.238S	0.421S	0S
qser2S	0.182S	0.110S	0.012S	0.352S	-0.209S	0.429S	0.070S	0.135S	0S
qser3S	0.704S	0.567S	0.419S	0.988S	0.060S	1.074S	0.117S	0.215S	0S
qser4S	1.102S	1.105S	0.739S	1.465S	0.521S	1.689S	0.150S	0.248S	0S
divorcedS	-0.264S	-0.289S	-0.295S	-0.232S	-0.395S	-0.184S	0.018S	0.045S	1S
divorced1S	0.086S	0.128S	0.075S	0.098S	0.100S	0.157S	0.007S	0.012S	1S
marriedS	-0.498S	-0.522S	-0.574S	-0.423S	-0.594S	-0.451S	0.029S	0.030S	0S
widowedS	-0.984S	-0.955S	-1.028S	-0.940S	-1.059S	-0.852S	0.026S	0.044S	1S
widowed1S	-0.623S	-0.472S	-0.667S	-0.579S	-0.530S	-0.413S	0.026S	0.025S	1S

Table 57: Log of total networth in 2000 dollars, one observation per married couple

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	67.110S	75.883S	60.560S	73.660S	68.672S	83.094S	2.928S	3.352S	0S
blackS	-0.654S	-0.315S	-0.827S	-0.481S	-0.474S	-0.155S	0.070S	0.090S	0S
black_spouseS	-0.026S	-0.294S	-0.164S	0.111S	-0.446S	-0.142S	0.064S	0.087S	0S
highschool_onlyS	0.482S	0.315S	0.464S	0.500S	0.283S	0.346S	0.011S	0.019S	1S
somecollegeS	0.634S	0.443S	0.565S	0.703S	0.399S	0.486S	0.040S	0.025S	1S
college_onlyS	1.057S	0.769S	0.959S	1.155S	0.718S	0.819S	0.058S	0.029S	1S
graduateS	1.152S	0.841S	1.054S	1.250S	0.788S	0.894S	0.058S	0.030S	1S
highschool_only_spouseS	0.031S	0.290	0.002S	0.060S	0.258S	0.323S	0.018S	0.019S	0S
somecollege_spouseS	0.042S	0.398S	-0.003S	0.087S	0.366S	0.431S	0.025S	0.020S	0S
college_only_spouseS	0.074S	0.677S	0.040S	0.107S	0.641S	0.713S	0.019S	0.022S	0S
graduate_spouseS	0.083S	0.696S	0.051S	0.115S	0.652S	0.740S	0.019S	0.026S	0S
totfam_kidsS	-0.039S	-0.042S	-0.062S	-0.016S	-0.067S	-0.017S	0.010S	0.012S	0S
foreign_bornS	0.036S	-0.001S	0.007S	0.066S	-0.051S	0.048S	0.018S	0.028S	0S
foreign_born_spouseS	0.012S	0.061S	-0.018S	0.042S	-0.010S	0.133S	0.018S	0.037S	0S
hispanicS	-0.068S	-0.168S	-0.119S	-0.017S	-0.263S	-0.074S	0.029S	0.051S	0S
hispanic_spouseS	-0.056S	-0.126S	-0.092S	-0.020S	-0.217S	-0.035S	0.018S	0.049S	0S
own_homeS	1.349S	1.765S	1.326S	1.371S	1.712S	1.818S	0.013S	0.027S	1S
retiredS	0.024S	-0.089S	-0.033S	0.081S	-0.138S	-0.039S	0.024S	0.028S	0S
disabledssaS	-0.163S	-0.253S	-0.348S	0.022S	-0.317S	-0.188S	0.085S	0.038S	0S
agedspouseS	0.023S	-0.089S	-0.061S	0.106S	-0.174S	-0.003S	0.049S	0.050S	1S
widowspouseS	-0.034S	-0.282S	-0.220S	0.151S	-0.389S	-0.174S	0.073S	0.060S	0S
otherbenefitS	-0.011S	-0.058S	-0.052S	0.030S	-0.100S	-0.015S	0.024S	0.025S	0S
retired_spouseS	0.032S	0.079S	-0.023S	0.087S	0.038S	0.120S	0.028S	0.024S	0S
disabledssa_spouseS	-0.026S	-0.079S	-0.122S	0.070S	-0.135S	-0.023S	0.047S	0.034S	0S
agedspouse_spouseS	0.045S	0.104S	-0.042S	0.132S	0.033S	0.175S	0.047S	0.043S	0S
widowspouse_spouseS	-0.061S	-0.024S	-0.206S	0.083S	-0.200S	0.151S	0.070S	0.095S	0S
otherbenefit_spouseS	-0.019S	-0.039S	-0.064S	0.026S	-0.087S	0.010S	0.025S	0.028S	0S
year_birthS	-0.026S	-0.021S	-0.028S	-0.024S	-0.026S	-0.016S	0.001S	0.002S	1S
year_birth_spouseS	-0.003S	-0.014S	-0.007S	0.000S	-0.016S	-0.011S	0.002S	0.001S	1S

Table 58: Log of total net worth in 2000 dollars, single individuals

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	16.892S	15.110S	14.374S	19.410S	11.606S	18.613S	1.481S	1.773S	1S
maleS	0.073S	0.067S	0.015S	0.132S	0.000S	0.133S	0.034S	0.030S	1S
blackS	-0.687S	-0.646S	-0.775S	-0.599S	-0.798S	-0.493S	0.052S	0.067S	1S
highschool_onlyS	0.057S	0.148S	-0.021S	0.136S	0.074S	0.222S	0.046S	0.034S	1S
somecollegeS	0.198S	0.392S	0.110S	0.285S	0.320S	0.465S	0.051S	0.034S	1S
college_onlyS	0.376S	0.645S	0.316S	0.436S	0.586S	0.704S	0.035S	0.031S	1S
graduateS	0.460S	0.778S	0.393S	0.526S	0.690S	0.866S	0.039S	0.044S	1S
widowedS	-0.070S	-0.079S	-0.107S	-0.032S	-0.138S	-0.019S	0.022S	0.033S	1S
divorcedS	-0.301S	-0.346S	-0.353S	-0.250S	-0.429S	-0.263S	0.030S	0.038S	1S
foreign_bornS	0.032S	0.036S	-0.013S	0.077S	-0.050S	0.121S	0.026S	0.041S	1S
totfam_kidsS	0.009S	-0.017S	-0.004S	0.021S	-0.027S	-0.007S	0.007S	0.006S	1S
hispanicS	-0.132S	-0.367S	-0.187S	-0.076S	-0.429S	-0.305S	0.024S	0.031S	0S
own_homeS	1.774S	2.126S	1.745S	1.803S	2.090S	2.163S	0.017S	0.018S	1S
retiredS	0.232S	0.312S	0.184S	0.279S	0.266S	0.357S	0.028S	0.027S	1S
disabledssaS	-0.069S	-0.159S	-0.133S	-0.005S	-0.221S	-0.097S	0.035S	0.036S	0S
agedspouseS	0.137S	0.309S	-0.155S	0.428S	0.033S	0.586S	0.151S	0.150S	0S
otherbenefitS	-0.043S	-0.110S	-0.093S	0.006S	-0.155S	-0.065S	0.025S	0.024S	0S
widowspouseS	0.157S	0.289S	0.066S	0.247S	0.235S	0.343S	0.035S	0.033S	0S
year_birthS	-0.004S	-0.003S	-0.005S	-0.003S	-0.005S	-0.001S	0.001S	0.001S	1S

Table 59: Log of home equity in 2000 dollars, one observation per married couple

Explanatory Variables	CoefficientS		Confidence Interval				Standard Error		SyntheticS
	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	57.204S	68.769S	52.663S	61.746S	62.126S	75.411S	2.671S	3.172S	1S
blackS	-0.271S	-0.127S	-0.673S	0.130S	-0.257S	0.002S	0.148S	0.078S	0S
black_spouseS	-0.028S	-0.269S	-0.147S	0.090S	-0.395S	-0.144S	0.055S	0.076S	0S
highschool_onlyS	0.218S	0.134S	0.064S	0.373S	0.067S	0.202S	0.091S	0.034S	1S
somecollegeS	0.314S	0.187S	0.156S	0.471S	0.123S	0.251S	0.093S	0.034S	1S
college_onlyS	0.552S	0.335S	0.449S	0.654S	0.215S	0.455S	0.061S	0.057S	1S
graduateS	0.577S	0.383S	0.460S	0.695S	0.313S	0.454S	0.069S	0.038S	1S
highschool_only_spouseS	0.012S	0.140S	-0.016S	0.041S	0.104S	0.177S	0.016S	0.021S	0S
somecollege_spouseS	0.010S	0.188S	-0.032S	0.053S	0.150S	0.226S	0.020S	0.023S	0S
college_only_spouseS	0.021S	0.309S	-0.034S	0.076S	0.262S	0.355S	0.028S	0.027S	0S
graduate_spouseS	0.032S	0.313S	-0.021S	0.086S	0.255S	0.371S	0.027S	0.033S	0S
foreign_bornS	0.126S	0.090S	0.071S	0.182S	0.019S	0.161S	0.029S	0.037S	0S
foreign_born_spouseS	0.000S	0.069S	-0.028S	0.028S	0.014S	0.125S	0.015S	0.031S	0S
totfam_kidsS	-0.015S	-0.001S	-0.029S	-0.002S	-0.025S	0.023S	0.007S	0.012S	0S
hispanicS	-0.087S	-0.066S	-0.128S	-0.047S	-0.133S	0.001S	0.023S	0.040S	0S
hispanic_spouseS	-0.001S	-0.121S	-0.043S	0.040S	-0.184S	-0.058S	0.024S	0.038S	0S
qder2S	0.016S	-0.004S	-0.016S	0.049S	-0.043S	0.036S	0.019S	0.023S	0S
qder3S	-0.008S	0.016S	-0.054S	0.038S	-0.022S	0.053S	0.024S	0.022S	0S
qder4S	0.060S	0.178S	-0.005S	0.125S	0.133S	0.223S	0.039S	0.025S	1S
qder2_spouseS	0.013S	0.016S	-0.029S	0.055S	-0.013S	0.046S	0.023S	0.018S	0S
qder3_spouseS	0.031S	0.023S	-0.016S	0.078S	-0.007S	0.053S	0.023S	0.018S	0S
qder4_spouseS	0.091S	0.219S	0.040S	0.143S	0.188S	0.250S	0.030S	0.019S	1S
qser2S	-0.002S	-0.036S	-0.048S	0.043S	-0.076S	0.004S	0.024S	0.022S	0S
qser3S	-0.027S	-0.077S	-0.077S	0.024S	-0.133S	-0.021S	0.026S	0.029S	0S
qser4S	-0.062S	-0.103S	-0.153S	0.029S	-0.180S	-0.026S	0.039S	0.038S	0S
qser2_spouseS	-0.008S	-0.028S	-0.054S	0.037S	-0.063S	0.007S	0.016S	0.020S	0S
qser3_spouseS	0.004S	-0.088S	-0.037S	0.045S	-0.126S	-0.050S	0.021S	0.022S	0S
qser4_spouseS	-0.001S	-0.120S	-0.094S	0.093S	-0.175S	-0.065S	0.034S	0.029S	0S
retiredS	0.012S	-0.083S	-0.029S	0.054S	-0.120S	-0.047S	0.017S	0.022S	0S
disabledssaS	-0.043S	-0.063S	-0.277S	0.190S	-0.135S	0.010S	0.104S	0.042S	0S
agedspouseS	0.008S	-0.052S	-0.066S	0.082S	-0.128S	0.023S	0.043S	0.045S	0S
widowspouseS	-0.031S	-0.237S	-0.116S	0.055S	-0.338S	-0.136S	0.045S	0.057S	0S
otherbenefitS	-0.002S	-0.028S	-0.044S	0.039S	-0.078S	0.021S	0.023S	0.029S	0S
retired_spouseS	0.034S	0.046S	0.015S	0.054S	0.005S	0.088S	0.012S	0.024S	1S
disabledssa_spouseS	0.030S	0.008S	-0.024S	0.084S	-0.052S	0.068S	0.032S	0.036S	0S
agedspouse_spouseS	0.044S	0.125S	-0.045S	0.133S	0.044S	0.205S	0.048S	0.047S	0S
widowspouse_spouseS	0.012S	-0.026S	-0.073S	0.097S	-0.157S	0.104S	0.035S	0.075S	0S
otherbenefit_spouseS	-0.033S	-0.020S	-0.079S	0.014S	-0.081S	0.041S	0.028S	0.034S	0S
year_birthS	-0.021S	-0.019S	-0.024S	-0.019S	-0.021S	-0.018S	0.001S	0.001S	1S
year_birth_spouseS	-0.003S	-0.011S	-0.006S	0.001S	-0.014S	-0.007S	0.002S	0.002S	1S

Table 60: Log of home equity in 2000 dollars, single individuals

	CoefficientS		Confidence Interval				Standard Error		SyntheticS
Explanatory Variables	SyntheticS	Completed	SyntheticS		Completed		Synthetic	Completed	DF Not ExistS
InterceptS	23.169S	14.041S	16.960S	29.378S	9.200S	18.881S	3.652S	2.360S	1S
maleS	0.042S	0.019S	-0.112S	0.196S	-0.152S	0.190S	0.091S	0.074S	1S
blackS	-0.296S	-0.437S	-0.507S	-0.085S	-0.600S	-0.274S	0.124S	0.072S	1S
highschool_onlyS	0.047S	0.103S	-0.004S	0.099S	0.065S	0.142S	0.031S	0.021S	1S
somecollegeS	0.201S	0.244S	0.054S	0.348S	0.219S	0.269S	0.086S	0.015S	1S
college_onlyS	0.301S	0.354S	0.220S	0.382S	0.310S	0.397S	0.048S	0.025S	1S
graduateS	0.328S	0.407S	0.243S	0.412S	0.351S	0.462S	0.050S	0.032S	1S
foreign_bornS	0.069S	0.066S	-0.019S	0.157S	0.009S	0.124S	0.039S	0.031S	0S
hispanicS	-0.111S	-0.222S	-0.166S	-0.055S	-0.316S	-0.129S	0.026S	0.045S	0S
totfam_kidsS	-0.023S	-0.012S	-0.038S	-0.007S	-0.024S	0.000S	0.009S	0.007S	1S
qder2S	0.020S	-0.002S	-0.039S	0.079S	-0.037S	0.034S	0.023S	0.020S	0S
qder3S	0.014S	-0.060S	-0.009S	0.037S	-0.100S	-0.020S	0.014S	0.022S	1S
qder4S	0.023S	0.045S	-0.008S	0.055S	-0.024S	0.113S	0.015S	0.035S	0S
qser2S	-0.011S	-0.023S	-0.058S	0.036S	-0.054S	0.008S	0.018S	0.018S	0S
qser3S	-0.022S	-0.072S	-0.089S	0.046S	-0.116S	-0.029S	0.031S	0.023S	0S
qser4S	-0.024S	-0.099S	-0.102S	0.054S	-0.148S	-0.049S	0.033S	0.026S	0S
retiredS	0.119S	0.271S	0.101S	0.138S	0.212S	0.330S	0.011S	0.033S	1S
disabledssaS	-0.005S	-0.016S	-0.064S	0.053S	-0.076S	0.045S	0.028S	0.036S	0S
agedspouseS	0.120S	0.218S	-0.099S	0.340S	0.007S	0.429S	0.113S	0.123S	0S
widowspouseS	0.130S	0.274S	0.109S	0.152S	0.200S	0.348S	0.012S	0.043S	1S
otherbenefitS	-0.017S	-0.047S	-0.073S	0.039S	-0.093S	-0.001S	0.020S	0.025S	0S
divorcedS	-0.326S	-0.324S	-0.349S	-0.304S	-0.487S	-0.161S	0.013S	0.072S	1S
widowedS	-0.053S	-0.024S	-0.160S	0.054S	-0.198S	0.149S	0.063S	0.079S	1S
year_birthS	-0.006S	-0.002S	-0.009S	-0.003S	-0.004S	0.001S	0.002S	0.001S	1S

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 1

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
					Date variables9						
birthdate9	completed9	1/22/19559	1/12/19139	4/28/1922	9/6/19289	4/21/19439	6/13/19579	4/1/1969	2/1/1977	9/10/1979	4/20/19819
birthdate9	synthesized9	2/17/19559	4/24/19139	8/22/19229	3/23/1929	10/1/19439	7/2/19579	1/27/1969	8/25/19769	6/10/1979	3/7/19819
date_initial_entitle9	completed9	3/9/19889	12/9/19639	1/31/19709	10/9/19739	12/24/19809	10/24/1989	5/24/19 69	6/1/2000	9/9/2001	9/1/20029
date_initial_entitle9	synthesized9	5/17/19889	3/3/19649	4/5/19709	11/21/19739	3/7/19819	12/21/1989	7/30/19 69	6/20/20009	8/31/2001	9/29/20029
deathdate9	completed9	7/5/20019	4/12/20009	5/17/20009	7/16/20009	12/2/20009	7/3/20019	2/17/20029	6/24/20029	8/5/2002	9/14/20029
deathdate9	synthesized9	10/19/20009	2/4/19 39	4/22/19 69	8/6/19 89	7/13/20009	3/18/20019	11/26/20019	6/2/20029	8/28/20029	12/7/20029
					MBA Variables9						
mba_20009	completed9	6439	28	919	1569	3539	611	919	11369	12609	1549
mba_20009	synthesized9	6429	37	949	1559	3509	609	219	11379	1259	15379
mba_initial_real9	completed9	609	349	1059	1659	3379	549	8739	11169	12369	14339
mba_initial_real9	synthesized9	6129	459	1109	1709	339	5519	8809	11209	12379	14319
					Marital History Variables9						
age_mar19	completed9	23.49	15.89	17.29	18.19	19.89	22.39	25.69	309	33.49	42.49
age_mar19	synthesized9	23.19	15.79	179	17.9	19.69	22.19	25.29	29.29	32.49	40.9
duration_end19	completed	9289	09	19	19	39	139	19619	19739	19779	19819
duration_end19	synthesized	9639	09	19	29	49	2429	19489	19759	20059	20709
duration_end29	completed9	10749	09	19	29	49	19339	19609	19689	19729	19789
duration_end29	synthesized9	1059	09	29	49	219	18609	19519	19889	20159	20629
duration_end39	completed9	18149	19	69	19289	19429	19539	19619	19679	1969	19749
duration_end39	synthesized9	18429	19	7239	1929	19449	19549	19619	19669	1969	19739
duration_mar19	completed9	14.59	0.39	1.29	2.29	4.7	9.69	20.19	369	44.79	55.59
duration_mar19	synthesized9	13.49	0.39	1.39	2.29	4.49	8.89	18.29	33.19	42.39	53.69
duration_mar29	completed9	1169	09	29	49		19559	19709	19759	19789	19819
duration_mar29	synthesized9	12009	19	29	59	179	19419	1969	19769	19839	20659
duration_mar39	completed9	12989	09	29	39	109	19539	19649	19709	19739	1979
duration_mar39	synthesized9	12189	29	119	279	1419	18039	19659	19879	20269	21139
duration_mar49	completed9	19559	19259	19369	19419	1949	19569	19629	19679	19709	19729
duration_mar49	synthesized9	19569	19249	19349	19429	19509	19579	19639	19689	19709	19739
					Wealth Variables9						
homeequity9	completed9	723149	-90009	40009	80009	220009	500009	1000009	1636259	2157509	3200009
homeequity9	synthesized9	744919	-262729	13279	52849	18539	489429	1010629	1788729	2493269	3809429
nonhouswealth9	completed9	749259	-70009	10009	20009	60009	170009	610009	1815009	3172509	7650009
nonhouswealth9	synthesized9	729219	-752359	-5139	10569	46959	151819	560429	1776019	3240719	8776949
totnetworth9	completed9	1196329	-330009	-60009	1000	90009	515009	1410009	2945009	449500	9250009
totnetworth9	synthesized9	1131459	-428449	-76319	-9 59	75259	497619	1374189	2870709	4362229	8796409
					DER Earnings Arrays9						
nondefer_der_fica_19789	completed9	132009	49	3119	7389	25409	71889	140269	210729	27729	669669
nondefer_der_fica_19789	synthesized9	141119	1119	516	969	26919	71069	13879	21139	283279	831519
nondefer_der_fica_1979	completed9	132689	569	3459	8139	27459	77309	149339	229009	288969	586539
nondefer_der_fica_1979	synthesized9	143529	69	394	9439	28989	77269	150209	229039	289589	631229
nondefer_der_fica_19809	completed9	119309	589	349	8259	28729	82729	160149	250689	303369	532909
nondefer_der_fica_19809	synthesized9	120419	1719	729	13389	3369	83069	159729	246439	301859	565059
nondefer_der_fica_19819	completed9	126019	619	397	969	329	3159	177119	274059	333749	56929
nondefer_der_fica_19819	synthesized9	125659	109	539	11579	3481	92189	175769	270849	328339	562989
nondefer_der_fica_19829	completed9	13439	709	439	10209	35869	100719	187529	289419	354019	606739
nondefer_der_fica_19829	synthesized9	134269	1709	7319	14159	37679	100079	184009	283269	347829	609439

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel92

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
nondefer_der_fica_19839	completed9	141849	689	4459	10469	36879	104349	196869	305309	375329	652439
nondefer_der_fica_19839	synthesized9	140919	1219	6229	13489	37749	102019	190329	297629	369879	643349
nondefer_der_fica_19849	completed9	150539	779	4719	11049	39759	111059	208629	327309	39 059	69 219
nondefer_der_fica_19849	synthesized9	153939	1809	8189	15719	41429	112629	206229	322669	396659	754069
nondefer_der_fica_19859	completed9	158159	709	479	11419	40449	115149	219209	343769	418009	744779
nondefer_der_fica_19859	synthesized9	159 89	1379	6589	14319	41909	115049	216609	339079	42649	780729
nondefer_der_fica_19869	completed9	166679	789	4929	11549	42219	12019	229309	361129	444789	814209
nondefer_der_fica_19869	synthesized9	172919	1409	6859	14679	43029	118649	228249	35629	442179	876179
nondefer_der_fica_19879	completed9	173929	79	5009	11819	43859	125389	237209	373229	452329	849 49
nondefer_der_fica_19879	synthesized9	176819	1579	7859	16089	46039	122559	232849	37089	460979	891749
nondefer_der_fica_19889	completed9	18259	859	5369	12539	45709	130669	24759	392119	477479	896559
nondefer_der_fica_19889	synthesized9	182549	889	5479	14089	4589	127059	24179	386149	47761	900159
nondefer_der_fica_1989	completed9	18877	949	5729	13409	48629	136419	257449	406019	49613	938859
nondefer_der_fica_1989	synthesized9	187719	1469	7279	15639	47479	133869	253929	401089	49767	933579
nondefer_der_fica_19 09	completed9	19588	909	5969	1429	51229	141809	265059	411439	51300	965439
nondefer_der_fica_19 09	synthesized9	195559	1569	8149	17849	5139	139689	262319	407109	52098	981979
nondefer_der_fica_19 19	completed9	20495	949	59	14439	52959	145489	272119	426159	551369	1095679
nondefer_der_fica_19 19	synthesized9	207549	1309	6819	16149	52449	143549	268649	421419	556389	1140559
nondefer_der_fica_19 29	completed9	21543	919	5819	14109	53569	150339	284439	447409	581309	1204509
nondefer_der_fica_19 29	synthesized9	21869	1439	7359	17249	52779	148079	282189	44329	587749	1214419
nondefer_der_fica_19 39	completed9	22267	919	5789	1459	55959	15449	29289	46379	610939	1277539
nondefer_der_fica_19 39	synthesized9	226319	214	959	2079	58929	153319	289379	460819	617969	1277789
nondefer_der_fica_19 49	completed9	226349	889	5729	14559	55979	156339	295619	467319	614179	1258679
nondefer_der_fica_19 49	synthesized9	22989	215	9809	20679	59319	154289	293259	46729	62059	1305959
nondefer_der_fica_19 59	completed9	23562	909	6009	14889	57719	160829	303229	482589	637019	1319039
nondefer_der_fica_19 59	synthesized9	239189	2769	12209	24209	65809	162159	301609	480159	641709	1350039
nondefer_der_fica_19 69	completed9	25237	979	6709	15879	60009	166559	313019	49 539	664179	1383019
nondefer_der_fica_19 69	synthesized9	251069	201	9859	2209	66589	171609	315879	504119	67309	1461849
nondefer_der_fica_19 79	completed9	258719	1069	7849	18569	65139	176459	329669	528369	706879	1528649
nondefer_der_fica_19 79	synthesized9	263339	180	9229	21969	68289	17719	332129	536919	727859	1539709
nondefer_der_fica_19 89	completed9	277329	133	9249	21709	73659	189289	347849	555109	746459	1597829
nondefer_der_fica_19 89	synthesized9	285549	2729	12609	27049	78409	192859	35359	559749	752839	1636109
nondefer_der_fica_19	completed9	296479	1409	10589	25539	82739	201989	364069	580829	787239	170779
nondefer_der_fica_19	synthesized9	315949	3409	16489	3409	0179	208069	369789	587459	798909	1813589
nondefer_der_fica_20009	completed9	323209	1519	1249	2970	94849	217679	383959	613409	839319	1828949
nondefer_der_fica_20009	synthesized9	338289	3909	19309	39579	102979	225759	393769	63049	87879	1893909
nondefer_der_fica_20019	completed9	330959	1569	12639	32049	103689	23159	401859	638829	869 69	1845849
nondefer_der_fica_20019	synthesized9	347109	3909	19619	4209	109229	2379	408569	646379	882579	1974139
nondefer_der_fica_20029	completed9	336979	1239	12049	32459	106459	240049	414089	657719	88989	1882329
nondefer_der_fica_20029	synthesized9	357029	3439	17539	40479	110189	243959	422379	67194	914139	1956409
nondefer_der_fica_20039	completed9	347579	1329	13239	33649	112089	249 79	428359	67701	912129	193449
nondefer_der_fica_20039	synthesized9	371309	4519	21949	46339	115319	249889	43179	68410	945249	2245619
nondefer_der_nonfica_19789	completed9	85359	19	759	1859	8179	58459	136049	202119	257429	381789
nondefer_der_nonfica_19789	synthesized9	86509	469	2079	3949	1149	57009	135339	198939	249 29	378189
nondefer_der_nonfica_1979	completed9	88459	179	819	1989	869	64119	145459	211449	253209	361519
nondefer_der_nonfica_1979	synthesized	91009	1019	4229	7589	18309	63039	144729	208519	25309	362369
nondefer_der_nonfica_19809	completed	9019	219	1009	2489	1119	73849	16079	230749	28059	401439

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel93

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
nondefer_der_nonfica_19809	synthesized9	101609	1229	4709	8179	1919	6769	163219	235219	285429	41489
nondefer_der_nonfica_19819	completed9	10949	229	869	2119	8719	68449	179129	265569	328489	501079
nondefer_der_nonfica_19819	synthesized9	107739	1069	4629	8189	19109	63269	171689	257179	316939	480109
nondefer_der_nonfica_19829	completed9	112729	169	719	1659	6449	45639	19 839	288269	352139	553069
nondefer_der_nonfica_19829	synthesized9	11236	939	3869	679	15949	47439	190159	285309	351139	532249
nondefer_der_nonfica_19839	completed9	1229	19	739	1889	7659	69459	21869	301979	362849	537159
nondefer_der_nonfica_19839	synthesized9	12079	1079	459	8259	2029	66539	201319	288869	355349	516069
nondefer_der_nonfica_19849	completed9	138539	189	759	169	759	1379	242139	321169	386609	575679
nondefer_der_nonfica_19849	synthesized9	130619	1049	4159	7429	17759	5959	227839	31979	389049	568109
nondefer_der_nonfica_19859	completed9	135029	19	729	1639	7179	68529	246289	334719	397359	57649
nondefer_der_nonfica_19859	synthesized9	154869	89	3719	679	18909	106689	264539	359249	433169	606749
nondefer_der_nonfica_19869	completed9	14329	209	759	1889	7979	72259	255139	345449	419749	665039
nondefer_der_nonfica_19869	synthesized9	14020	909	3749	6869	16779	58489	247039	342019	418759	656839
nondefer_der_nonfica_19879	completed9	14429	239	839	1959	8139	72379	259529	354889	428139	624779
nondefer_der_nonfica_19879	synthesized9	140679	1129	4609	8309	20479	63909	241559	350129	425929	609 19
nondefer_der_nonfica_19889	completed9	15509	219	879	2049	8429	85179	276189	376079	450959	656089
nondefer_der_nonfica_19889	synthesized9	15077	929	3369	5619	12319	66909	267139	373649	455059	681379
nondefer_der_nonfica_1989	completed9	166609	209	839	209	8389	82829	286409	393369	479149	688579
nondefer_der_nonfica_1989	synthesized9	153469	829	3529	6359	14649	60779	273239	392449	482249	706559
nondefer_der_nonfica_19 09	completed9	1619	25	929	218	9029	78929	292569	409219	495279	706949
nondefer_der_nonfica_19 09	synthesized9	154219	1049	4009	6929	15149	54609	278159	406219	492309	703489
nondefer_der_nonfica_19 19	completed9	181969	269		2459	10809	112009	319439	44379	531159	747839
nondefer_der_nonfica_19 19	synthesized9	1749	1049	4609	8379	22689	81529	306539	437009	526769	719589
nondefer_der_nonfica_19 29	completed9	190969	249	1049	2579	11239	118629	332219	463879	555339	807379
nondefer_der_nonfica_19 29	synthesized9	176449	1219	4919	8849	21879	76979	306079	449 79	542749	765859
nondefer_der_nonfica_19 39	completed9	201729	309	1239	3149	12939	115539	34179	487409	59780	924249
nondefer_der_nonfica_19 39	synthesized9	190309	1259	506	9009	21409	8289	322779	47649	59339	27639
nondefer_der_nonfica_19 49	completed9	203489	259	1119	3009	12679	117559	342539	488829	589459	876049
nondefer_der_nonfica_19 49	synthesized9	189709	1339	5739	10879	2817	91039	320239	474279	569739	801009
nondefer_der_nonfica_19 59	completed9	209549	289	1139	2859	13069	117449	347889	504289	623539	1006549
nondefer_der_nonfica_19 59	synthesized9	196039	1309	5659	10139	23049	84609	328549	482379	59353	984819
nondefer_der_nonfica_19 69	completed9	211649	339	1129	2669	12059	112689	353929	510589	62305	951219
nondefer_der_nonfica_19 69	synthesized9	195909	1429	5929	10769	26209	82269	325619	496269	610149	869 39
nondefer_der_nonfica_19 79	completed9	223089	359	1449	3419	15089	122059	365789	530459	660639	1074579
nondefer_der_nonfica_19 79	synthesized9	223949	1819	7539	13989	34449	102219	340139	537229	671439	1076309
nondefer_der_nonfica_19 89	completed9	222589	479	1739	379	14859	117009	368879	53559	662719	1078119
nondefer_der_nonfica_19 89	synthesized9	21579	1969	7979	1419	33179	101019	345849	533569	663389	1035519
nondefer_der_nonfica_19	completed9	235079	409	1779	4359	17619	127879	376589	552879	677889	1072959
nondefer_der_nonfica_19	synthesized9	238659	2509	10319	18559	44189	122179	366069	559089	703169	1100349
nondefer_der_nonfica_20009	completed9	24279	439	1779	4419	1779	129229	390629	57009	69 29	1148639
nondefer_der_nonfica_20009	synthesized9	238869	2739	10309	18139	40889	119 69	37009	572509	716009	1126479
nondefer_der_nonfica_20019	completed9	250259	359	1689	4549	19069	137459	401609	586449	722379	111589
nondefer_der_nonfica_20019	synthesized9	251019	2909	11649	20889	4939	13479	383519	592509	732659	111949
nondefer_der_nonfica_20029	completed9	282479	469	2279	6009	28589	194769	438759	63539	780239	1232959
nondefer_der_nonfica_20029	synthesized9	27569	1649	7649	15009	45789	180929	425609	623509	774219	1149009
nondefer_der_nonfica_20039	completed9	291249	459	2289	649	33719	210969	456659	652839	788069	1249839
nondefer_der_nonfica_20039	synthesized9	332949	4819	20729	3793	9309	281979	486829	671109	82069	117449

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel94

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
defer_der_fica_19879	completed9	11059	4109	4109	4109	8709	11019	12269	16379	20489	20489
defer_der_fica_19879	synthesized	9239	6159	6159	6539	775	9219	10709	11329	11329	12439
defer_der_fica_19889	completed9	77719	09	1229	1419	5909	17619	36729	58749	76279	76279
defer_der_fica_19889	synthesized9	135459	89	609	1559	5859	19429	42179	62569	774539	2896579
defer_der_fica_1989	completed9	32639	109	89	1579	5309	12009	24549	47519	61789	264889
defer_der_fica_1989	synthesized9	25779	359	1359	239	5559	12989	27209	49 39	7479	205509
defer_der_fica_19 09	completed9	21139	359	1489	2619	6009	13279	28009	49 79	66659	84009
defer_der_fica_19 09	synthesized9	19659	309	1319	2479	5609	11889	25489	48449	65149	88039
defer_der_fica_19 19	completed9	23219	449	1549	2669	6109	13709	28219	51279	70249	87509
defer_der_fica_19 19	synthesized9	2119	449	1809	3079	6089	12839	26579	50659	6914	96889
defer_der_fica_19 29	completed9	22809	469	1739	2909	6449	14439	29	5429	73139	87289
defer_der_fica_19 29	synthesized9	22759	69	229	3619	6749	13889	29149	54589	7535	9529
defer_der_fica_19 39	completed9	24489	459	1729	3009	6789	15039	31909	57639	7774	94919
defer_der_fica_19 39	synthesized9	23659	619	219	3619	6969	14589	30919	57539	7916	96939
defer_der_fica_19 49	completed9	27049	449	1819	3129	7019	15849	33349	60039	8051	92409
defer_der_fica_19 49	synthesized9	25359	519	2189	3709	7239	15749	32249	58539	7938	95849
defer_der_fica_19 59	completed9	27759	469	1719	3129	7089	1619	34549	6279	8341	92409
defer_der_fica_19 59	synthesized9	25329	559	2459	4249	8489	15619	33349	61139	8312	95429
defer_der_fica_19 69	completed9	26419	509	1949	3369	7659	17149	36349	66729	8719	5009
defer_der_fica_19 69	synthesized9	2639	59	2439	4249	839	17259	35749	65749	8652	97119
defer_der_fica_19 79	completed9	27969	49	1919	3389	7869	18129	38929	7185	9369	5009
defer_der_fica_19 79	synthesized9	27909	609	2519	4369	8679	18189	38569	7075	9157	96279
defer_der_fica_19 89	completed9	2979	519	2059	3689	849	19279	419	7722	9369	100009
defer_der_fica_19 89	synthesized9	29869	689	2849	483	9419	19419	41129	7548	95579	105919
defer_der_fica_19	completed9	31409	489	2039	3789	89	20249	44539	80969	100009	100009
defer_der_fica_19	synthesized9	31789	59	2439	439	259	20179	43789	7975	97969	126469
defer_der_fica_20009	completed9	3429	459	2059	378	9059	20859	46469	85749	105009	105009
defer_der_fica_20009	synthesized9	34439	709	2979	5079	10249	21619	46889	85429	103269	117419
defer_der_fica_20019	completed9	33749	459	2019	387	939	21549	4824	90339	105009	105009
defer_der_fica_20019	synthesized9	3429	29	3389	5479	10739	22259	48759	89259	103519	111789
defer_der_fica_20029	completed9	3489	459	2009	390	9539	21979	4920	96449	110009	120009
defer_der_fica_20029	synthesized9	35139	839	3109	5159	10379	21949	4856	9489	109739	125229
defer_der_fica_20039	completed9	37209	559	2259	4089	10009	22719	5119	104529	120009	140009
defer_der_fica_20039	synthesized9	40269	1029	4469	6609	12529	25329	54829	107159	120289	142069
defer_der_nonfica_19879	completed9	61949	3649	3649	5489	1369	2286	97719	169759	169759	169759
defer_der_nonfica_19879	synthesized9	34029	709	709	489	20439	34359	4719	58949	58949	68709
defer_der_nonfica_19889	completed9	17619	2209	2209	2209	3319	4079	17419	66069	102129	102129
defer_der_nonfica_19889	synthesized9	17989	509	509	509	8759	12359	25829	42039	42039	43719
defer_der_nonfica_1989	completed9	18729	419	729	7839	11509	16009	19509	36009	38029	82009
defer_der_nonfica_1989	synthesized9	1929	829	4769	6869	11589	1759	22979	35849	44939	52179
defer_der_nonfica_19 09	completed9	38849	769	2939	480	9649	18859	34019	6779	79 59	125679
defer_der_nonfica_19 09	synthesized9	29729	409	2289	430	9169	17529	30669	55939	76389	116959
defer_der_nonfica_19 19	completed9	25869	259	1919	3549	8369	1779	3179	6309	75009	122549
defer_der_nonfica_19 19	synthesized9	25239	729	2669	436	919	17259	30919	59039	78639	119769
defer_der_nonfica_19 29	completed9	24729	79	679	2249	7379	17019	3209	60009	75009	116739
defer_der_nonfica_19 29	synthesized9	23429	449	1729	3069	7249	1539	29809	56409	77459	116649
defer_der_nonfica_19 39	completed9	25319	59	609	2069	6869	16489	33229	60529	75009	125009

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel95

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
defer_der_nonfica_19 39	synthesized9	2319	409	1689	3029	6949	14989	28869	55529	74959	122669
defer_der_nonfica_19 49	completed9	24489	69	819	19	6509	16229	31589	60009	75009	117579
defer_der_nonfica_19 49	synthesized9	22109	39	1539	2649	6209	14159	28239	54319	73879	110859
defer_der_nonfica_19 59	completed9	24859	129	819	1949	6509	16849	32039	61519	75009	122319
defer_der_nonfica_19 59	synthesized9	22219	419	1519	2629	5979	14189	28189	54659	74319	114319
defer_der_nonfica_19 69	completed9	2479	79	609	2089	6509	18009	33409	60319	75009	111149
defer_der_nonfica_19 69	synthesized9	23309	49	1659	2869	6429	15219	30479	57919	75809	112719
defer_der_nonfica_19 79	completed9	26719	79	769	2259	7749	18479	35069	66679	78369	120009
defer_der_nonfica_19 79	synthesized9	25379	459	1819	3189	7319	17059	32689	62589	80129	116509
defer_der_nonfica_19 89	completed9	279	11	979	2589	8449	1939	37629	70089	80009	122409
defer_der_nonfica_19 89	synthesized9	24459	439	1719	3089	709	16409	31479	61089	78419	112989
defer_der_nonfica_19	completed9	28759	119	1019	2369	8179	19719	38749	71259	80359	122409
defer_der_nonfica_19	synthesized9	25079	449	1739	3079	719	16729	32539	62039	80259	113379
defer_der_nonfica_20009	completed9	29279	129	109	2679	8519	20939	39229	73449	80049	120009
defer_der_nonfica_20009	synthesized9	25739	519	1939	3389	7809	17759	33459	62839	80939	110369
defer_der_nonfica_20019	completed9	31429	219	1209	269	8259	21669	43059	78959	85859	125429
defer_der_nonfica_20019	synthesized9	27319	509	2159	3769	8219	18169	36069	68929	85209	111189
defer_der_nonfica_20029	completed9	35529	179	1449	327	9689	24009	49289	85009	110009	139 59
defer_der_nonfica_20029	synthesized9	31589	659	2359	3989	8749	19539	43919	80459	101759	126419
defer_der_nonfica_20039	completed9	37779	219	139	3009	8979	24009	5335	93619	120009	146329
defer_der_nonfica_20039	synthesized9	35239	739	2649	452	9869	21559	47729	85979	112149	144709
				SER Earnings Arrays9							
earn1937_to_19519	completed9	55559		589	1529	6849	28639	80729	148129	195209	304249
earn1937_to_19519	synthesized9	60819	39	119	229	1509	16629	85089	185539	266589	397609
totearn_ser_19519	completed9	1659		489	1129	4439	14639	28219	36009	36009	36009
totearn_ser_19519	synthesized9	15289	159	69	1349	3709	11939	26139	36009	36009	36009
totearn_ser_19529	completed9	17549	109	529	1239	4609	16079	30859	36009	36009	36009
totearn_ser_19529	synthesized9	16239	19	819	1539	3979	13429	28329	36009	36009	36009
totearn_ser_19539	completed9	18569	109	559	1389	5219	17719	33609	36009	36009	36009
totearn_ser_19539	synthesized9	17309	21	909	1669	439	14909	31379	36009	36009	36009
totearn_ser_19549	completed9	18959	109	549	1339	5379	18519	34789	36009	36009	36009
totearn_ser_19549	synthesized9	17649	22	969	1789	4769	15529	31729	36009	36009	36009
totearn_ser_19559	completed9	20609	149	669	1529	5509	18789	36649	42009	42009	42009
totearn_ser_19559	synthesized9	19339	259	1059	1959	509	16459	33589	42009	42009	42009
totearn_ser_19569	completed9	21769	149	749	1719	6459	20809	39169	42009	42009	42009
totearn_ser_19569	synthesized9	20789	279	1189	2189	5959	18839	36679	42009	42009	42009
totearn_ser_19579	completed9	22529	169	879	2149	7749	21909	40659	42009	42009	42009
totearn_ser_19579	synthesized9	2159	329	1349	2519	7039	19 89	37909	42009	42009	42009
totearn_ser_19589	completed9	22829	149	839	2039	7759	2239	41969	42009	42009	42009
totearn_ser_19589	synthesized9	2159	29	1259	2359	6959	20779	37369	4129	41889	42689
totearn_ser_1959	completed9	25059	17	929	219	8139	24039	44639	48009	48009	48009
totearn_ser_1959	synthesized9	24009	309	129	2439	7149	22179	41819	48009	48009	48009
totearn_ser_19609	completed9	2559	189	1039	2389	8579	24739	4629	48009	48009	48009
totearn_ser_19609	synthesized9	24389	359	1429	2619	7289	22539	43469	48009	48009	48009
totearn_ser_19619	completed9	26029	209	1079	2479	8949	25449	47829	48009	48009	48009
totearn_ser_19619	synthesized9	24389	319	1319	2489	7149	22339	44039	48009	48009	48009
totearn_ser_19629	completed9	26759	189	1079	274	9559	26679	48009	48009	48009	48009

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel96

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
totearn_ser_19629	synthesized9	25329	389	1549	2849	8059	24109	45579	48009	48009	48009
totearn_ser_19639	completed9	27189	19	1129	270	969	27669	48009	48009	48009	48009
totearn_ser_19639	synthesized9	25849	359	1489	2789	8269	25259	46239	48009	48009	48009
totearn_ser_19649	completed9	27949	209	1179	2869	10279	29379	48009	48009	48009	48009
totearn_ser_19649	synthesized9	26239	329	1389	2649	8189	26249	47389	48009	48009	48009
totearn_ser_19659	completed9	28339	239	1289	2919	10489	30269	48009	48009	48009	48009
totearn_ser_19659	synthesized9	26919	389	159	2969	8689	2739	48009	48009	48009	48009
totearn_ser_19669	completed9	33839	239	1329	3129	10889	31469	60109	66009	66009	66009
totearn_ser_19669	synthesized9	31909	379	1629	316	9389	28579	55229	66009	66009	66009
totearn_ser_19679	completed9	3479	249	1379	3359	11719	33079	62689	66009	66009	66009
totearn_ser_19679	synthesized9	33039	469	1889	349	79	29 69	58769	66009	66009	66009
totearn_ser_19689	completed9	38979	289	1569	3619	12669	35829	67369	78009	78009	78009
totearn_ser_19689	synthesized9	36829	49	2029	3769	10819	32479	62749	78009	78009	78009
totearn_ser_1969	completed9	40579	29	1769	3989	13509	38189	72489	78009	78009	78009
totearn_ser_1969	synthesized9	38809	49	209	4029	11739	35259	68179	78009	78009	78009
totearn_ser_19709	completed9	42139	289	1779	4019	14369	41069	76609	78009	78009	78009
totearn_ser_19709	synthesized9	40149	539	2229	4229	12439	37659	70989	78009	78009	78009
totearn_ser_19719	completed9	4339	289	1789	4119	15029	43689	78009	78009	78009	78009
totearn_ser_19719	synthesized9	4129	539	2219	4249	12679	39829	739	78009	78009	78009
totearn_ser_19729	completed9	47949	309	1929	4449	15959	46189	8683	9000	9000	90009
totearn_ser_19729	synthesized9	45539	559	2329	4489	13489	42449	8055	9000	9000	90009
totearn_ser_19739	completed9	53989	389	2109	4789	169	4968	93279	108009	108009	108009
totearn_ser_19739	synthesized9	51759	609	259	4969	14879	46609	88749	108009	108009	108009
totearn_ser_19749	completed9	60879	369	2269	5339	19109	5346	9689	132009	132009	132009
totearn_ser_19749	synthesized9	58239	639	289	5659	16769	5026	95769	132009	132009	132009
totearn_ser_19759	completed9	64589	439	2439	5519	19659	56839	106269	141009	141009	141009
totearn_ser_19759	synthesized9	61949	759	3159	6019	17489	53339	101909	141009	141009	141009
totearn_ser_19769	completed9	70019	459	2729	6149	21649	61229	114839	153009	153009	153009
totearn_ser_19769	synthesized9	67209	829	3569	6779	1929	57249	110679	153009	153009	153009
totearn_ser_19779	completed9	75159	489	2879	6579	23169	65409	123529	165009	165009	165009
totearn_ser_19779	synthesized9	70349	859	3669	69	19709	5809	115339	165009	165009	165009
				SIPP Arrays9							
famwelamt19 09	completed9	21779	259	1139	2109	5319	11949	28479	57229	77639	114549
famwelamt19 09	synthesized9	23129	429	1729	2989	6249	12879	2979	59319	80209	121839
famwelamt19 19	completed9	26649	369	1549	2879	6689	15409	36259	68129	89049	131589
famwelamt19 19	synthesized9	22389	279	1249	2289	519	11429	27259	60479	83129	130909
famwelamt19 29	completed9	23269	219	1039	2019	5159	12749	31569	60939	80749	123019
famwelamt19 29	synthesized9	22089	289	1259	229	5149	1139	2769	58879	80759	127609
famwelamt19 39	completed9	2469	269	1169	2269	5709	13759	34129	64919	84719	123939
famwelamt19 39	synthesized9	2489	379	1549	2759	6019	13179	3219	6679	0159	132069
famwelamt19 49	completed9	22649	209	1069	2079	5259	1249	30079	59609	77939	118119
famwelamt19 49	synthesized9	20969	279	1209	2179	4779	10349	24909	57159	79329	129069
famwelamt19 59	completed9	19209	21	989	1939	4879	10829	22769	50909	71839	108159
famwelamt19 59	synthesized9	21029	289	119	2149	4789	10479	24069	58819	81669	131209
famwelamt19 69	completed9	28889	339	1279	2449	6219	15259	40159	7476	98469	150979
famwelamt19 69	synthesized9	30269	469	1859	319	6839	1489	39129	79739	10789	174219
famwelamt19 79	completed9	25949	289	1219	2249	5619	1339	33329	7025	94249	134589

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel97

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
famwelamt19 79	synthesized9	29	409	1659	2869	6169	13779	36819	84439	110659	193879
famwelamt19 89	completed9	22439	239	1039	1939	489	1169	27979	60759	82569	128879
famwelamt19 89	synthesized9	25469	369	139	239	5109	11389	29279	72739	104189	164059
famwelamt19	completed9	19409	279	1169	2039	4749	10339	22949	49489	72949	120559
famwelamt19	synthesized9	23449	39	1519	2569	5249	11169	27249	6466	90249	146439
fpov19 09	completed9	1296789	739579	797469	84031	96089	1190339	1572439	1877779	2083409	2722329
fpov19 09	synthesized9	1318629	746779	805819	860479	1017309	1213919	1582449	1889489	2085979	2676949
fpov19 19	completed9	1359129	766759	821739	863529	1038249	1250069	1656339	1973049	2189639	279869
fpov19 19	synthesized9	140209	771839	832879	890939	1065649	128769	168589	2013339	2233989	2831609
fpov19 29	completed9	1391759	792849	840939	883089	1063439	1276029	1702889	202009	2247669	2902619
fpov19 29	synthesized9	1438189	798059	85760	917319	109769	1323069	1731479	2061969	2279659	291539
fpov19 39	completed9	1436049	817829	87016	907329	1091869	1315489	1757189	209 279	2322739	3004789
fpov19 39	synthesized9	1468049	822619	87906	930579	1114119	13479	1777979	2124759	2343279	2964319
fpov19 49	completed9	1466609	814779	87869	25059	1108209	1347509	1800449	2135369	2358509	29 5789
fpov19 49	synthesized9	1487629	823759	88859	43879	1127349	1370949	181409	2150459	2358489	2956359
fpov19 59	completed9	1494389	792589	88736	952079	1117519	1357179	1833849	2185909	2416369	3145519
fpov19 59	synthesized9	1523469	80973	90013	967129	1138289	1391919	1857559	2210239	2447029	3159439
fpov19 69	completed9	1552139	87867	92638	978839	1142869	1398289	1910289	2300749	255309	3330049
fpov19 69	synthesized9	1570469	88733	94111	94479	1169659	1426029	1922279	2300079	2534859	3230779
fpov19 79	completed9	154949	88870	93505	96839	1163229	1395159	1949039	2301449	2555719	3216509
fpov19 79	synthesized9	160714	90227	953359	1008219	1189779	1461719	1977079	2368129	2617609	3297789
fpov19 89	completed9	1589 69	89096	948249	1013829	1182649	1453769	198289	2347559	2593649	3238419
fpov19 89	synthesized9	162224	90015	96439	1021509	1205329	1475059	19 289	2393169	2641339	334169
fpov19	completed9	162718	92542	97229	1037319	1207719	1482339	2025769	2405409	2653069	3343639
fpov19	synthesized9	168355	92148	98489	1044939	1236259	152739	2059 59	2480629	2762569	3582439
ftotinc19 09	completed9	369659	10449	5735	92369	176929	313289	495289	712619	885429	1283979
ftotinc19 09	synthesized9	352719	6589	55489	89789	171569	301379	472069	674339	83269	1212169
ftotinc19 19	completed9	387139	17649	6461	9889	189679	333259	522889	74401	903309	1235239
ftotinc19 19	synthesized9	35629	-689	47439	80009	159 79	295209	483089	711159	882219	1254459
ftotinc19 29	completed9	389139	13019	6146	9669	186369	330949	528819	75954	921749	1259759
ftotinc19 29	synthesized9	368159	229	51229	85329	168639	30669	49 669	73493	90209	1263039
ftotinc19 39	completed9	40269	20809	69219	105679	195559	344379	543759	77512	945709	1296989
ftotinc19 39	synthesized9	37969	859	52839	88109	173459	317819	515139	75370	929619	1298979
ftotinc19 49	completed9	417839	15609	67009	105089	197679	353349	565659	820649	1002879	138129
ftotinc19 49	synthesized9	387889	-1949	51429	87519	173379	32109	529 19	78275	968709	1363579
ftotinc19 59	completed9	430209	12639	66159	107939	204329	370559	586709	834059	1002309	1423819
ftotinc19 59	synthesized9	40829	-9 79	40679	79149	170849	334329	565769	836889	103719	1478419
ftotinc19 69	completed9	467039	10949	65119	104219	202939	369169	605359	898819	1127249	2085889
ftotinc19 69	synthesized9	438979	-879	5679	5809	189349	349089	574159	854909	1071859	1881859
ftotinc19 79	completed9	463589	1219	6322	9549	194649	36169	60471	902289	1143009	2133619
ftotinc19 79	synthesized9	451809	1129	5671	94949	188589	350789	58859	884909	1125909	2082029
ftotinc19 89	completed9	489 29	12359	66239	104829	205479	382839	64043	956039	1208979	2310819
ftotinc19 89	synthesized9	482769	849	5885	98979	197929	37109	62751	951969	1219279	2391909
ftotinc19	completed9	525619	13269	67349	108589	21539	40509	672859	1018179	133239	2745239
ftotinc19	synthesized9	513429	-389	5574	9789	201939	387189	657319	1012919	1337609	2823129
helamt19 09	completed9	22139	389	1549	2859	5929	11489	239	50489	79839	172389
helamt19 09	synthesized9	20519	329	1449	2569	5389	10309	20949	45269	74809	173439

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel98

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
helamt19 19	completed9	28679	519	2219	3969	8149	15279	32119	6527	98409	206339
helamt19 19	synthesized9	28869	579	2289	3949	789	14879	30489	66119	103279	221019
helamt19 29	completed9	3189	559	2119	3889	8349	16889	35079	75079	112879	210159
helamt19 29	synthesized9	29749	519	219	3929	7959	15309	3109	65919	10669	225589
helamt19 39	completed9	29179	479	1949	3609	7949	15989	33279	6650	97769	208239
helamt19 39	synthesized9	30219	559	2269	3959	8049	15549	32019	67689	107739	228259
helamt19 49	completed9	3269	569	2279	4089	8979	17749	37919	75149	111749	225549
helamt19 49	synthesized9	33159	569	2359	4169	869	16979	36319	76069	120019	250779
helamt19 59	completed9	39659	1169	4279	7069	13469	25189	46949	86559	131879	2129
helamt19 59	synthesized9	35949	849	29	5149	10669	20859	40359	77689	129439	232009
helamt19 69	completed9	32879	549	2439	434	9039	17979	35779	67409	10479	277489
helamt19 69	synthesized9	38449	839	3089	517	9859	18429	37359	78489	136959	374469
helamt19 79	completed9	41029	849	2969	5129	10489	21119	45639	89259	136609	242569
helamt19 79	synthesized9	62159	1309	4369	7029	12959	25109	5459	114719	189879	332029
helamt19 89	completed9	47849	1139	3609	6379	1189	23479	4757	97459	143109	268379
helamt19 89	synthesized9	65009	1139	4209	6739	12669	24039	47689	108329	186669	893019
helamt19	completed9	49459	119	4469	7279	1329	25879	5098	95239	142109	265109
helamt19	synthesized9	73809	1209	4679	7639	14539	27439	56049	116659	200549	857269
totearn19 09	completed9	176439	1079	6979	15879	53559	138869	248819	379679	484119	7739
totearn19 09	synthesized9	164659	859	5629	11409	38819	12559	236739	365769	468809	75989
totearn19 19	completed9	185069	1489	8329	17689	5549	143369	259489	40109	513679	836169
totearn19 19	synthesized9	17662	959	579	12429	43009	132369	251639	394419	508939	841929
totearn19 29	completed9	193539	158	9189	19859	60409	149179	269829	418319	541709	870169
totearn19 29	synthesized9	178819	859	6149	13789	4579	133439	253249	398159	517719	853779
totearn19 39	completed9	19519	151	9159	19669	60869	15059	27189	421439	549789	87849
totearn19 39	synthesized9	190169		6319	13979	48129	142189	269509	424159	55377	901149
totearn19 49	completed9	198719	1979	10849	22319	63489	152149	275349	429169	55686	90189
totearn19 49	synthesized9	190769	1129	7029	15539	50619	142609	26869	424689	55240	906229
totearn19 59	completed9	198579	2859	15259	29439	70149	151329	270009	421979	54303	911429
totearn19 59	synthesized9	193859	3469	15689	27869	63649	141039	264389	422509	55008	923259
totearn19 69	completed9	243239	1769	11979	25749	74839	174189	315219	494569	66129	1383329
totearn19 69	synthesized9	243529	2319	11989	24579	71369	171659	315179	495589	665089	1457249
totearn19 79	completed9	245719	2339	13139	28119	78709	178559	319119	497239	653229	1301389
totearn19 79	synthesized9	255929	219	11729	24659	74989	180279	328749	519529	69279	1514869
totearn19 89	completed9	253519	2679	15449	31969	8459	185609	327589	514839	673959	1331949
totearn19 89	synthesized9	270549	2339	12859	27329	81369	189719	343179	551869	737319	169 179
totearn19	completed9	278349	3549	20069	3947	96619	200439	350509	552609	737039	1719 89
totearn19	synthesized9	291309	159	12059	28119	85749	198919	36079	584529	805809	2139529
tothoursannual19 09	completed9	16719	449	1819	346	9729	19459	21979	25959	29159	35379
tothoursannual19 09	synthesized9	15219	19	1039	2209	729	18019	21419	24359	26919	32579
tothoursannual19 19	completed9	16039	479	1789	3279	8529	18719	21459	25279	28439	36089
tothoursannual19 19	synthesized9	15119	19	1159	2359	7249	17629	2129	24359	27179	33179
tothoursannual19 29	completed9	16839	559	2169	419	10729	19549	2169	25179	28129	34509
tothoursannual19 29	synthesized9	15619	409	169	3089	8229	18029	21409	24709	27629	34039
tothoursannual19 39	completed9	17069	539	2059	3869	10109	19709	22089	26739	30039	38059
tothoursannual19 39	synthesized9	15849	439	1679	3039	8159	18369	21669	25429	28429	34789
tothoursannual19 49	completed9	16519	649	2269	405	9769	19239	21539	25589	28559	34789

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
tothoursannual19 49	synthesized9	15579	689	219	3689	8389	17629	21289	24609	2749	33329
tothoursannual19 59	completed9	16209	1119	3209	506	9729	18369	21359	24529	27269	32439
tothoursannual19 59	synthesized9	15509	1479	3669	545	9629	16779	2089	23189	25429	30449
tothoursannual19 69	completed9	17439	619	2359	439	11179	19659	22429	27029	30159	36489
tothoursannual19 69	synthesized9	17119	679	239	4319	10789	1929	22269	26459	29439	35469
tothoursannual19 79	completed9	17309	579	2349	4459	11169	19859	22009	26319	29469	36449
tothoursannual19 79	synthesized9	17009	489	1949	3709	10279	19519	22189	26459	29579	36659
tothoursannual19 89	completed9	17269	619	2389	4649	11179	19869	21829	26089	29239	35879
tothoursannual19 89	synthesized9	17119	569	2169	4139	1069	19549	22139	26239	29329	36329
tothoursannual19	completed9	17479	69	2779	5269	11829	19 59	21829	25979	28889	35539
tothoursannual19	synthesized9	16909	459	2019	3939		19249	22029	26559	30029	37789
totinc19 09	completed9	166159	-829	1529	10249	46289	126769	237709	367939	471609	740689
totinc19 09	synthesized9	162829	-7649		8279	38729	117369	234809	373459	481589	768329
totinc19 19	completed9	172849	-6759	2009	10219	47629	129419	245879	385809	495739	806369
totinc19 19	synthesized9	168159	-569	2689	8969	39419	117569	240419	388639	504919	83379
totinc19 29	completed9	177629	-7209	1929	11009	50509	130879	250429	397769	516639	843909
totinc19 29	synthesized9	174689	-5119	3249	10169	44089	121329	246639	402849	52919	876989
totinc19 39	completed9	183009	-6679	3519	14909	54939	136379	256089	404209	524139	857079
totinc19 39	synthesized9	180659	-4949	4579	13309	49739	128309	253409	409169	535329	885609
totinc19 49	completed9	186679	-6839	7349	20409	59669	139429	258669	411169	53059	863489
totinc19 49	synthesized9	187149	-4879	769	18259	53589	132129	25979	424859	55541	910049
totinc19 59	completed9	189 59	-7589	10819	26279	64009	140609	261319	41229	528849	875779
totinc19 59	synthesized9	189 69	-6369	10769	23649	569	130369	259629	428519	56489	4769
totinc19 69	completed9	218419	-7649	8939	25259	68929	155239	287689	457329	593709	1120629
totinc19 69	synthesized9	220119	-322	969	24119	6579	151589	28879	466449	61269	1168309
totinc19 79	completed9	231039	-2909	1089	28739	74449	165529	302859	47749	622969	117509
totinc19 79	synthesized9	227519	-2589	10919	24969	6809	156759	296939	478559	630149	1229
totinc19 89	completed9	238049	-39	12979	32339	78959	172929	312889	49679	647019	1159109
totinc19 89	synthesized9	2379	-2829	1359	29019	72689	163609	307139	502019	666329	1305209
totinc19	completed9	260429	-3739	16789	3669	85209	180879	32969	527389	707209	1634349
totinc19	synthesized9	259889	-2079	16639	33639	78019	172059	327889	537869	726329	1710449
wkspt19 09	completed9	15.89	0.39	1.19	2.29	5.29	12.39	22.59	36.79	44.59	529
wkspt19 09	synthesized9	15.79	0.29	1.29	2.29	5.19	129	22.49	36.79	44.29	51.19
wkspt19 19	completed9	15.89	0.39	1.19	2.19	59	12.19	22.59	37.19	45.79	529
wkspt19 19	synthesized9	15.79	0.39	1.29	2.29	5.19	11.9	22.29	379	45.29	529
wkspt19 29	completed9	16.9	0.39	1.19	2.29	5.69	13.49	24.59	38.29	45.89	539
wkspt19 29	synthesized9	16.39	0.29	1.29	2.29	5.49	12.89	23.69	37.29	44.39	529
wkspt19 39	completed9	16.49	0.39	1.19	2.29	5.39	12.89	23.69	37.89	46.79	529
wkspt19 39	synthesized9	16.29	0.39	1.29	2.39	5.39	12.49	23.19	389	46.59	529
wkspt19 49	completed9	15.79	0.39	1.29	2.19	5.19	12.49	21.49	36.59	45.79	529
wkspt19 49	synthesized9	15.9	0.39	1.49	2.59	5.59	12.79	21.69	36.29	44.79	529
wkspt19 59	completed9	16.59	0.69	2.29	3.89	8.29	13.9	23.59	32.49	409	49.49
wkspt19 59	synthesized9	18.19	0.79	2.39	3.9	8.59	15.79	25.19	36.69	42.29	49.59
wkspt19 69	completed9	22.39	0.79	2.89	4.9	11.29	19.19	33.29	44.69	48.89	529
wkspt19 69	synthesized9	229	0.59	2.39	4.29	10.49	18.9	33.49	44.89	48.69	51.9
wkspt19 79	completed9	19.89	0.49	1.79	3.39	8.19	16.59	29.79	43.29	49.49	539
wkspt19 79	synthesized9	20.69	0.49	1.9	3.59	8.39	17.29	31.49	44.69	49.89	529

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel910

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
wkspt19 89	completed9	209	0.69	2.29	3.9	8.39	16.59	29.79	439	49.49	529
wkspt19 89	synthesized9	21.29	0.69	2.59	4.3	9.29	17.9	329	459	49.9	529
wkspt19	completed9	209	0.59	2.29	3.89	8.69	16.89	29.79	42.39	48.59	529
wkspt19	synthesized9	21.9	0.69	2.79	4.79	109	18.89	32.9	459	49.39	529
wkswp19 09	completed9	419	3.29	109	17.39	35.39	47.59	51.29	51.9	529	529
wkswp19 09	synthesized9	40.79	29	7.59	13.89	34.79	48.59	51.29	51.9	529	529
wkswp19 19	completed9	41.59	3.69	10.49	17.29	35.79	48.89	51.69	529	529	529
wkswp19 19	synthesized9	41.29	2.6	9.19	15.79	34.59	49	51.59	529	529	529
wkswp19 29	completed9	41.9	3.89	11.59	19.49	36.9	47.69	51.59	539	539	539
wkswp19 29	synthesized9	41.59	2.5	9.59	17.19	36.79	48.29	51.39	529	529	529
wkswp19 39	completed9	42.49	3.89	11.19	18.59	38.19	49.39	51.79	529	529	529
wkswp19 39	synthesized9	42.19	2.6	9.69	16.9	37.19	49.79	51.69	529	529	529
wkswp19 49	completed9	429	4.59	12.89	19.39	36.39	48.59	51.69	529	529	529
wkswp19 49	synthesized9	42.29	3.9	12.19	19.19	36.9	49.19	51.49	51.9	529	529
wkswp19 59	completed9	41.29	6.89	15.19	20.29	35.49	46.59	50.59	51.69	51.89	529
wkswp19 59	synthesized9	41.39	4.59	12.49	19.49	35.49	47.59	50.79	51.69	51.89	529
wkswp19 69	completed9	43.29	49	13.59	22.49	39.79	49.39	51.59	529	529	529
wkswp19 69	synthesized9	449	3.79	13.69	249	429	49.9	51.59	529	529	529
wkswp19 79	completed9	43.39	3.69	11.59	209	39	50.9	51.9	529	539	539
wkswp19 79	synthesized9	43.89	2.89	11.19	20.79	41.69	50.89	51.9	529	529	529
wkswp19 89	completed9	449	4.39	14.9	24.9	40.9	50.19	51.9	529	529	529
wkswp19 89	synthesized9	44.79	4.79	15.59	26.49	43.19	50.59	51.89	529	529	529
wkswp19	completed9	43.9	5.89	169	26.19	40.49	49.59	51.89	529	529	529
wkswp19	synthesized9	44.89	4.79	16.19	27.29	439	50.59	51.89	529	529	529

Table 61:9Percentiles of9Synthetic and9Completed9Variables9 Panel911

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
					Cardinal Categorical Variables9						
time_arrive_usa9	completed9	5.569	19	19	29	49	69	89	89	89	89
time_arrive_usa9	synthesized9	5.449	19	19	1.19	49	69	89	89	89	89
totfam_kids9	completed9	0.9	09	09	09	09	09	29	39	39	59
totfam_kids9	synthesized9	0.9	09	09	09	09	09	29	39	39	59

Table 62: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)

	WEIGHTED				UNWEIGHTED			
	Count		Percentage		Count		Percentage	
Variable	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed
Male								
0	106,949,968	106,470,823	52.41	52.18	138,357	138,357	52.45	52.45
1	97,094,759	97,573,904	47.59	47.82	125,436	125,436	47.55	47.55
Black								
0	181,725,962	180,358,529	89.06	88.39	232,401	233,326	88.10	88.45
1	22,318,765	23,686,198	10.94	11.61	31,392	30,467	11.90	11.55
Hispanic								
0	184,181,614	181,560,851	90.27	88.98	238,277	238,558	90.33	90.43
1	19,863,113	22,483,876	9.73	11.02	25,516	25,235	9.67	9.57
Maritalstat								
1	103,680,995	102,213,429	50.81	50.09	141,292	141,292	53.56	53.56
2	9,673,233	9,958,884	4.74	4.88	17,483	17,483	6.63	6.63
3	23,885,992	24,400,916	11.71	11.96	31,103	31,103	11.79	11.79
4	66,804,507	67,471,498	32.74	33.07	73,915	73,915	28.02	28.02
Tob_initial								
	144,158,303	142,951,882	70.65	70.06	173,251	173,251	65.68	65.68
1	28,120,067	28,896,858	13.78	14.16	45,543	45,543	17.26	17.26
2	9,771,612	9,934,114	4.79	4.87	15,401	15,401	5.84	5.84
3	2,562,145	2,603,705	1.26	1.28	4,028	4,028	1.53	1.53
5	3,881,472	4,000,409	1.90	1.96	6,920	6,920	2.62	2.62
100	15,551,128	15,657,759	7.62	7.67	18,652	18,652	7.07	7.07
Tob_2000								
	150,570,926	149,502,998	73.79	73.27	196,028	196,028	74.31	74.31
1	26,411,823	27,162,166	12.94	13.31	34,306	34,306	13.00	13.00
2	5,873,327	5,926,142	2.88	2.90	7,571	7,571	2.87	2.87
3	1,860,289	1,880,083	0.91	0.92	2,433	2,433	0.92	0.92
5	4,311,070	4,460,705	2.11	2.19	5,781	5,781	2.19	2.19
100	15,017,292	15,112,633	7.36	7.41	17,675	17,675	6.70	6.70
Own_home								
0	69,152,585	69,785,113	33.89	34.20	87,230	86,839	33.07	32.92
1	134,892,142	134,259,614	66.11	65.80	176,563	176,955	66.93	67.08
Foreign_born								
0	180,969,389	179,239,395	88.69	87.84	234,496	235,262	88.89	89.18
1	23,075,338	24,805,332	11.31	12.16	29,297	28,531	11.11	10.82

Table 62: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)

	WEIGHTED				UNWEIGHTED			
	Count		Percentage		Count		Percentage	
Variable	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed
Educ_5cat								
1	43,271,926	42,898,854	21.21	21.02	56,851	55,903	21.55	21.19
2	69,217,821	68,520,001	33.92	33.58	89,529	90,001	33.94	34.12
3	51,100,854	52,842,654	25.04	25.90	64,585	66,115	24.48	25.06
4	22,930,251	23,458,739	11.24	11.50	29,804	29,230	11.30	11.08
5	17,523,875	16,324,479	8.59	8.00	23,024	22,544	8.73	8.55
Age_cat12								
<=21	10,993,657	12,355,657	5.39	6.06	4,452	4,480	1.69	1.70
22-24	11,083,762	11,637,091	5.43	5.70	8,296	8,484	3.14	3.22
25-29	19,330,649	18,791,213	9.47	9.21	24,922	25,344	9.45	9.61
30-34	20,827,765	19,839,766	10.21	9.72	28,311	27,657	10.73	10.48
35-39	22,623,857	22,137,354	11.09	10.85	28,824	28,715	10.93	10.89
40-44	22,702,993	22,384,522	11.13	10.97	28,493	28,579	10.80	10.83
45-49	20,541,360	20,091,637	10.07	9.85	26,195	26,150	9.93	9.91
50-54	18,114,622	17,779,895	8.88	8.71	23,761	23,797	9.01	9.02
55-59	13,593,649	13,497,244	6.66	6.61	17,822	17,895	6.76	6.78
60-64	10,666,518	10,893,885	5.23	5.34	14,636	14,469	5.55	5.48
65-69	9,336,754	9,547,527	4.58	4.68	12,829	12,849	4.86	4.87
>=70	24,229,138	25,088,937	11.87	12.30	45,255	45,375	17.16	17.20
Agecat_initial_entitle (Tob_initial=1,2,3,5)								
<62	11,699,746	11,688,294	26.39	25.73	18,555	18,184	25.81	25.29
>=62 and <63	7,273,818	18,508,709	16.41	40.74	11,673	29,067	16.24	40.43
>=63 and <64	11,912,607	3,419,745	26.87	7.53	19,452	5,391	27.06	7.50
>=64 and <65	7,337,083	2,757,925	16.55	6.07	11,964	4,961	16.64	6.90
>=65 and <66	3,351,021	7,440,877	7.56	16.38	5,536	11,672	7.70	16.24
>=66 and <67	1,514,780	482,726	3.42	1.06	2,587	798	3.60	1.11
>=67	1,208,960	1,077,981	2.73	2.37	2,051	1,726	2.85	2.40
>=80	37,283	58,829	0.08	0.13	73	90	0.10	0.13
Agecat_retire (Tob_initial=1 only)								
<62	56,373	45,537	0.20	0.16	83	73	0.18	0.16
>=62 and <63	5,843,165	15,661,703	20.78	54.20	9,187	24,252	20.17	53.25
>=63 and <64	10,449,792	2,770,806	37.16	9.59	16,945	4,374	37.21	9.60
>=64 and <65	6,457,351	2,351,612	22.96	8.14	10,486	4,237	23.02	9.30
>=65 and <66	2,941,391	6,703,497	10.46	23.20	4,850	10,471	10.65	22.99
>=66 and <67	1,323,219	406,445	4.71	1.41	2,259	663	4.96	1.46
>=67	1,020,418	910,199	3.63	3.15	1,677	1,403	3.68	3.08
>=80	28,357	47,059	0.10	0.16	56	71	0.12	0.16

Table 63: Agreement Probabilities for Individuals with Spouses

Field|	Comparison| Type|	Pr(agree match):| m|	Pr(agree non-match):| u|	Agree weight:| ln(m/u)|	Disagree weight:| ln(1-m)/(1-u)|
Hispanic|	c|	0.954479|	0.835287|	0.133390|	-1.286023|
Educ_5cat|	c|	0.330004|	0.241200|	0.313478|	-0.124467|
Disab_in_scope|	c|	0.949006|	0.777256|	0.199645|	-1.474307|
Disab|	c|	0.843075|	0.810676|	0.039187|	-0.187691|
Disab_nowork|	c|	0.637131|	0.541970|	0.161765|	-0.232893|
Totfam_kids_wave2|	c|	0.469601|	0.329187|	0.355257|	-0.234861|
Ind_4cat|	c|	0.361122|	0.309276|	0.154980|	-0.078026|
Foreign_born|	c|	0.844434|	0.788724|	0.068250|	-0.306097|
Time_arrive_usa|	c|	0.236797|	0.162303|	0.377738|	-0.093133|
Ind_exist|	c|	0.762450|	0.568762|	0.293074|	-0.596280|
Occ_exist|	c|	0.775007|	0.572171|	0.303434|	-0.642654|
Occ_4cat|	c|	0.446905|	0.343057|	0.264449|	-0.172067|
Mh_category|	c|	0.591162|	0.574111|	0.029268|	-0.040861|
Flag_mar4t|	c|	0.987294|	0.987260|	0.000035|	-0.002695|
Own_home|	c|	0.719070|	0.668007|	0.073660|	-0.167008|
Pension_in_scope_age|	c|	0.976252|	0.949419|	0.027870|	-0.756061|
Pension_in_scope_empl|	c|	0.702327|	0.557740|	0.230506|	-0.395902|

Table 64: Agreement Probabilities for Single Individuals

Field|	Comparison| Type|	Pr(agree match):| m|	Pr(agree non-match):| u|	Agree weight:| ln(m/u)|	Disagree weight:| ln(1-m)/(1-u)|
Hispanic|	c|	0.888222|	0.817697|	0.082729|	-0.489153|
Educ_5cat|	c|	0.360123|	0.252198|	0.356231|	-0.155862|
Disab_in_scope|	c|	0.923310|	0.744927|	0.214679|	-1.201784|
Disab|	c|	0.824805|	0.113998|	1.978968|	-1.620817|
Disab_nowork|	c|	0.679595|	0.222995|	1.114350|	-0.885862|
Totfam_kids_wave2|	c|	0.568113|	0.130233|	1.472992|	-0.700061|
Ind_4cat|	c|	0.356281|	0.305685|	0.153165|	-0.075664|
Foreign_born|	c|	0.852712|	0.094033|	2.204775|	-1.816610|
Time_arrive_usa|	c|	0.289757|	0.091983|	1.147440|	-0.245656|
Ind_exist|	c|	0.784428|	0.603121|	0.262838|	-0.610339|
Occ_exist|	c|	0.784490|	0.602726|	0.263572|	-0.611621|
Occ_4cat|	c|	0.465897|	0.388607|	0.181394|	-0.135150|
Mh_category|	c|	0.763459|	0.067933|	2.419334|	-1.371281|
Flag_mar4t|	c|	0.990087|	0.004855|	5.317686|	-4.609064|
Own_home|	c|	0.547307|	0.242271|	0.814954|	-0.515111|
Pension_in_scope_age|	c|	0.887510|	0.585350|	0.416210|	-1.304568|
Pension_in_scope_empl|	c|	0.693329|	0.211577|	1.186915|	-0.944258|

Table 65: Match Rates for Married Individuals, Split into Data Blocks

Segment	*Match Status*	*COUNT*	*PERCENT*
1|	FALSE|	29939|	99.31|
	TRUE|	209|	0.69|
2|	FALSE|	19660|	99.57|
	TRUE|	84|	0.43|
3|	FALSE|	19517|	99.62|
	TRUE|	74|	0.38|
4|	FALSE|	20202|	99.71|
	TRUE|	58|	0.29|
5|	FALSE|	20017|	99.71|
	TRUE|	58|	0.29|
6|	FALSE|	19811|	99.62|
	TRUE|	76|	0.38|
7|	FALSE|	19658|	99.65|
	TRUE|	69|	0.35|
8|	FALSE|	19564|	99.7|
	TRUE|	58|	0.3|
9|	FALSE|	18305|	99.63|
	TRUE|	68|	0.37|
10|	FALSE|	19724|	99.73|
	TRUE|	54|	0.27|

Table 66: Match Rates for Single Individuals, Split into Data Blocks

Segment	*matchstatus*	*COUNT*	*PERCENT*
1|	FALSE|	21717|	99.2|
	TRUE|	175|	0.8|
2|	FALSE|	18005|	98.82|
	TRUE|	215|	1.18|
3|	FALSE|	18028|	99.1|
	TRUE|	164|	0.9|
4|	FALSE|	18936|	99.28|
	TRUE|	138|	0.72|
5|	FALSE|	19102|	99.29|
	TRUE|	136|	0.71|
6|	FALSE|	18503|	99.18|
	TRUE|	153|	0.82|
7|	FALSE|	18682|	99.22|
	TRUE|	146|	0.78|
8|	FALSE|	18798|	99.34|
	TRUE|	124|	0.66|
9|	FALSE|	19034|	99.3|
	TRUE|	134|	0.7|
10|	FALSE|	19014|	99.31|
	TRUE|	132|	0.69|
11|	FALSE|	17411|	98.83|
	TRUE|	207|	1.17|
12|	FALSE|	19018|	99.25|
	TRUE|	144|	0.75|
13|	FALSE|	29939|	99.31|
	TRUE|	209|	0.69|

Table 67: Mahalanobis Distance Matching Results

Male	Marital Status	N Synth	N N GS	Match Rate 1 Maha1	Match Rate 2 Maha1	Ratio 2 to 1	Match Rate 3 Maha1	Ratio 3 to 2	Ratio 3, 2 to 1
1	1	70,814	70,814	1.11	0.50	0.45	0.44	0.88	0.84
0	1	70,478	70,478	1.03	0.55	0.53	0.44	0.81	0.96
1	4	39,434	39,434	0.97	0.52	0.54	0.39	0.74	0.93
0	4	34,481	34,481	1.18	0.73	0.62	0.55	0.74	1.09
0	3	18,733	18,733	1.05	0.54	0.51	0.33	0.61	0.83
0	2	14,668	14,668	1.04	0.67	0.64	0.50	0.74	1.12
1	3	12,370	12,370	1.04	0.46	0.44	0.38	0.82	0.81
1	2	2,815	2,815	2.91	1.53	0.52	0.78	0.51	0.79
Totals		263,793	263,793	1.09	0.57	0.52	0.44	0.79	0.93
Male	**Marital Status**	**N Synth**	**N N GS**	**Match Rate 1 Maha2**	**Match Rate 2 Maha2**	**Ratio 2 to 1**	**Match Rate 3 Maha2**	**Ratio 3 to 2**	**Ratio 3, 2 to 1**
1	1	70,814	70,814	0.80	0.39	0.48	0.31	0.81	0.87
0	1	70,478	70,478	0.67	0.38	0.57	0.32	0.83	1.05
1	4	39,434	39,434	0.68	0.39	0.58	0.28	0.71	0.99
0	4	34,481	34,481	0.80	0.50	0.63	0.42	0.84	1.15
0	3	18,733	18,733	0.64	0.40	0.62	0.34	0.85	1.15
0	2	14,668	14,668	0.78	0.41	0.53	0.38	0.93	1.02
1	3	12,370	12,370	0.74	0.30	0.41	0.35	1.16	0.88
1	2	2,815	2,815	2.20	0.99	0.45	0.75	0.75	0.79
Totals		263,793	263,793	0.75	0.41	0.55	0.34	0.83	1.00

Table 68: Euclidean Distance Matching Results

Male	Marital Status	N Synth	N N GS	Match Rate 1 EUCL1	Match Rate 2 EUCL1	Ratio 2 to 1	Match Rate 3 EUCL1	Ratio 3 to 2	Ratio 3, 2 to 1
1	1	70,814	70,814	0.60	0.40	0.66	0.31	0.77	1.17
0	1	70,478	70,478	0.58	0.39	0.67	0.27	0.71	1.15
1	4	39,434	39,434	0.49	0.28	0.58	0.21	0.75	1.01
0	4	34,481	34,481	0.53	0.32	0.61	0.30	0.93	1.18
0	3	18,733	18,733	0.90	0.57	0.63	0.36	0.63	1.03
0	2	14,668	14,668	0.47	0.42	0.90	0.22	0.53	1.38
1	3	12,370	12,370	0.74	0.45	0.61	0.40	0.88	1.14
1	2	2,815	2,815	0.82	0.50	0.61	0.36	0.71	1.04
Totals		263,793	263,793	0.59	0.38	0.65	0.29	0.75	1.14
Male	**Marital Status**	**N Synth**	**N N GS**	**Match Rate 1 EUCL2**	**Match Rate 2 EUCL2**	**Ratio 2 to 1**	**Match Rate 3 EUCL2**	**Ratio 3 to 2**	**Ratio 3, 2 to 1**
1	1	70,814	70,814	1.26	0.74	0.58	0.55	0.75	1.02
0	1	70,478	70,478	1.43	0.81	0.57	0.66	0.81	1.03
1	4	39,434	39,434	0.94	0.59	0.62	0.51	0.87	1.16
0	4	34,481	34,481	1.16	0.67	0.58	0.51	0.76	1.02
0	3	18,733	18,733	0.91	0.56	0.61	0.42	0.76	1.07
0	2	14,668	14,668	1.03	0.53	0.52	0.52	0.99	1.03
1	3	12,370	12,370	0.91	0.53	0.58	0.44	0.85	1.06
1	2	2,815	2,815	2.31	1.17	0.51	1.03	0.88	0.95
Totals		263,793	263,793	1.20	0.70	0.58	0.56	0.81	1.05

This version: $Id: SSAfinal.tex 188 2006-11-06 01:10:54Z jma7 $