WP.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (ii): New and Emerging Methods

MULTIPLE IMPUTATION METHODS FOR IMPUTING EARNINGS IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

Prepared by María García, Chandra Erdman, and Ben Klemens, US Census Bureau, USA[1]

I. INTRODUCTION

1. The Survey of Income and Program Participation (SIPP) is a longitudinal survey with data collected in panels. The panels are samples of approximately 14,000 to 65,000 households with interviews occurring at set frequencies (called waves) over a period of two to four years. The SIPP collects information on a wide range of demographic characteristics. It also asks questions about assets and liabilities, health insurance coverage, educational attainment, program participation, labor force participation, and earnings. The main goal of the SIPP is to provide comprehensive information about the income and program participation of households and individuals in the United States, which allows the government to evaluate the effectiveness of federal, state, and local programs.

2. In 2006, the U.S. Census Bureau began a major redesign of the SIPP with the goal of reducing costs while improving data quality. The redesign includes less frequent interviewing, the introduction of the Event History Calendar, and an updated automated survey instrument. In addition to making changes to the way that SIPP data are collected, the Census Bureau is researching ways to improve SIPP data processing, and we take this opportunity to explore methods for missing data imputation. The SIPP uses a hot-deck method for most missing data imputation. The hot deck matches a record with missing data to that of a donor with similar characteristics and uses the donor's values to fill-in missing items. In this paper, we examine alternative procedures for imputing missing job-level SIPP earnings, and evaluate their performance using data collected in the panel that began in January, 2004. We describe two procedures for imputing monthly SIPP earnings: a model-based approach and a randomized hot deck. We present results of a simulation study designed for comparing the model-based approach to the randomized hot deck. Each of these procedures use a multiple imputation approach, which gives us a framework to estimate variances.

3. In the following section we describe current SIPP imputation procedures. In Section 3.1 we describe the randomized hot deck as it is implemented in the CSRM generalized edit and imputation system **TEA** and in Section 3.2 we describe a sequential regression multiple imputation approach. Section 4 presents the results of the simulation study for comparing the randomized hot deck and the model-based approach. Section 5 contains final remarks along with our recommendations.

[1]This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

II. BACKGROUND

4. Historically, missing SIPP data are handled in one of three ways, according to the type of missingness:

- Whole-household nonresponse, in which no household member responds to the survey, is handled in the weighting step of SIPP data processing. That is, a nonresponse adjustment is factored into the weights for the households that respond to the survey.
- Partial household nonresponse, occurs when some household members respond and others do not respond or provide insufficient information. This type of missingness is handled via hot deck. The hot deck finds a respondent that is similar to that of the nonrespondent on demographics that are available for both individuals and fills in the entire record of the nonrespondent with the record of the respondent.
- Item nonresponse occurs when respondents provide answers to some but not all questions. In some cases this type of missingness is handled via edit rules. That is, missing items are deduced and filled in based on other information in the record of the partial respondent. In other cases, item nonresponse is handled via hot deck.

5. To ensure that missing data are imputed from respondents living nearby, the hot-deck procedure sorts records by the geographic area of each household. Valid responses for each variable are then supplied from an imputation matrix, which typically has cells that are grouped by demographics such as age, race and ethnicity, gender, and so on. Missing values are imputed by selecting the most recent, or "hottest," value to enter the cells of the imputation matrix. If no valid responses are available, the cells of the imputation matrix are collapsed until a suitable donor is found.

6. Hot-deck imputation has been used by the U.S. Census Bureau since the 1960 Census, and works well in many cases. However, hot deck as it is currently employed for SIPP needs improvement. Citro and Scholz [2009] report that the variables that define the categories of SIPP imputation matrices are often not carefully tailored to the variables being imputed. Including more categories in the hot-deck imputation matrix may lead to a smaller donor pool, and the likelihood that a single donor is used to fill-in many missing values increases. Conversely, collapsing imputation cells leads to a larger donor pool but it decreases the number of characteristics that are observed in donors and recipients.

7. In the remaining sections we discuss a model based alternative method for imputing SIPP data. Ideally, we would compare the model-based procedure to the production hot deck; however, the production software is not available for our use. Instead, we compare the model-based approach with the hot deck as it is implemented in the **TEA** software for editing and imputation.

III. METHODS

A. TEA's Randomized Hot Deck

8. **TEA** is a generalized system designed to unify and streamline demographic survey processing, from raw data to editing to imputation to dissemination of output (Klemens [2012]). It has an R front-end, and has a back-end based on Apophenia, a library of statistics functions and models (Klemens [2008]), which is in turn based on the GNU Scientific Library (Gough [2003]). It is available as an unofficial R package, and has been used by the U.S. Census Bureau for the processing of group quarters data for the American Community Survey and the 2010 Census. **TEA** has an imputation component that allows users to choose from several models for imputation of missing fields. We use the randomized

hot-deck model as it is implemented in **TEA** to simulate production hot-deck imputation of missing SIPP monthly earnings. In the random hot deck, the donor is selected randomly from a set of potential donors, which we call the donor pool. We use the method to impute missing SIPP job-level earnings from observed SIPP earnings in the donor pool. The imputation procedure is set in a multiple-imputation framework, so each imputed value can be assigned a variance.

9. The randomized hot-deck model specification in **TEA** fills-in missing values by drawing from the non-missing data in the same category as the record with missing data. This requires designing a suitable list of categories to partition the population. For example, the variable *Age* is one of the variables we used when partitioning the population to create imputation cells. Missing job-level earnings in SIPP are in scope for individuals who are at least 15 years of age; from those we consider the following four levels of the variable *Age*:

$$15 \leq Age < 18,$$

$$18 \leq Age < 35,$$

$$35 \leq Age < 65, \text{and}$$

$$65 \leq Age.$$

10. Similarly, we consider four separate levels for the variable representing the number of jobs an individual had in the year in scope: Zero jobs, one job, two jobs, or more than two jobs.

11. **TEA**'s hot-deck procedure randomly selects a donor from the set of potential donors using the partitions for the levels of each of the variables. The model specification within **TEA** includes a list of categories; for example, this specification,

```
min group size: 20

categories {
  num_sipp_jobs_2004 = 0
  num_sipp_jobs_2004 = 1
  num_sipp_jobs_2004 = 2
  num_sipp_jobs_2004 > 2
  15<=agesipp20042<18
  18<=agesipp20042<35
  35<=agesipp20042<65
  65<=agesipp20042
}
```

creates 16 (Age × Number of Jobs) categories with at least 20 observations according to the defined minimum group size. For each category, **TEA** draws donors from the hot-deck matrix of non-missing observations to fill-in missing job-level earnings within the category. If one of these categories has under twenty non-missing observations, then the age grouping is dropped, and **TEA** draws donors from observations within the category for job count. For example, if the {*65 and over, more than two jobs*} category has too few observations, then **TEA**'s hot-deck imputes missing values for individuals that fall in the age over 65 category and who held more than two jobs by randomly taking draws from the category {*more than two jobs*}. For these data, we use only the number of jobs in the year in scope to generate categories. That is, we use four hot-deck models: One each for those who had zero jobs, one job, two jobs, or more than two jobs in the 2004 SIPP data.

B. Sequential Regression Multiple Imputation

12. As an alternative to hot-deck imputation of SIPP earnings, we examine a model-based procedure that makes use of sequential regression multiple imputation (SRMI) which imputes missing values sequentially, variable by variable, conditioning on all the observed and imputed variables (Raghunathan et al. [2001]). Note that although we are imputing job-level earnings, the SIPP data is collected in waves and thus there are twelve distinct variables representing reported earnings that might be missing: $January\ SIPP\ earnings, February\ SIPP\ earnings, ..., December\ SIPP\ earnings$.

13. Let the complete data matrix be represented by $Y = (Y_1, Y_2, ..., Y_n)$, where for each $j = 1, 2, ..., n$, $Y_j = (Y_{j,obs}, Y_{j,miss})$ denote the jth column or variable. On each iteration, SRMI generates imputations for $Y_{j,miss}$ as random draws from the predictive distribution specified by the regression model, $P(Y_j|Y_i, i \neq j, \theta_j)$, conditioning on the other variables and unknown parameters θ_j, for $j = 1, 2, ..., n$. The iterative process imputes sequentially for each variable as follows:

(1) Draw θ_j from $P(\theta_j|Y_{j,obs}; Y_i, i \neq j)$
(2) Draw $Y_{j,miss}$ from $P(Y_{j,miss}|Y_i, i \neq j; \theta_j)$

The iterative process is repeated until there is convergence, although it is theoretically possible the method may not converge (see Raghunathan et al. [2001] for a detailed description of the methodology).

14. This technique of imputing each variable using a regression model conditional on all the others variables and sequentially iterating through all the variables that contain missing values is also known as a "chained equations" approach. There are several computer software packages implementing this methodology. The program **IVEware** is a suite of macros callable from SAS (Raghunathan et al. [2001]) that performs single or multiple imputations of missing values using the SRMI method. The R package **MICE**, from Statistics Netherlands (Buuren et al. [2011]), implements SRMI using chained equations to generate multivariate imputations. The R package **mi** (Su et al. [2011]) which generates multiple imputations for incomplete data also implements a chained equations approach to iterative regression imputation.

15. At the Census Bureau, Benedetto and Stinson [2009] have written a set of SAS macros implementing SRMI with the purpose of multiply-imputing job-level earnings in the SIPP. They use the Bayesian Bootstrap [Rubin, 1981] to impute an indicator of positive earnings and sequential regression to estimate a posterior predictive distribution of earnings conditional on a set of observed characteristics including an independent source of administrative earnings data. They document testing their model using the SIPP 2004-2005 panel data with impressive results (see Benedetto and Stinson [2009]). Benedetto and Stinson's model and associated SAS program are an important first step into investigating alternative methods for imputing missing data in the SIPP and we thus follow their lead. Our approach to missing earnings imputation is similar, but we use the sequential regression multiple imputation model as it is implemented in the general Multiple Imputation R software package **mi** of Su et al. [2011].

16. The package **mi** is flexible and user-friendly. Using the software requires several steps, including some basic set-up steps and data pre-processing taking into account the missing data patterns. **mi** has a generic core function which executes one of three methods depending on the type of data and it chooses the conditional model based on the type of variable. **mi** has the capability to handle various types of variables including continuous, semi-continuous, log-continuous, binary, categorical and/or count variables, etc. **mi** includes several imputation models and determines which imputation model to use based on the variables types; the regression models for each variable can also be user-defined. The program executes the iterative imputation based on the specified conditional model using the chained

equations approach in which the imputation algorithm proceeds to sequentially iterate through each of the variables to impute missing values according to the specified model. The user specifies how many multiply-imputed completed datasets to generate. The program verifies the imputed missing values are reasonable and tests for convergence of the procedure. The software provides additional features like sensitivity analysis and cross-validation along with useful convergence diagnostics. We refer the reader to Su et al. [2011] for a more detailed description of the software functions and features.

17. One advantage of the model-based approach is that regression models can incorporate different types of information than the hot deck -hot deck cells preserves interactions while regression models preserve main effects- without the risk of depleting the donor pool. In addition, because the model-based approach does not use direct substitutions as does the hot deck, sensitive information such as administrative records may be used to guide the imputations as in the work of Benedetto and Stinson [2009]. The model-based approach is generally more flexible than the hot deck. However, with flexibility often comes complexity, and careful consideration must be given to the implementation of such methods. Using this available R package for SRMI makes it possible to easily modify the model and apply it to other SIPP variables with minimal effort from the user. In the next section we present results of a simulation study to determine the feasibility of using a model-based multiple imputation approach with current and future SIPP data by examining earnings imputed using a randomized hot deck with those imputed using the general SRMI software of Su et al. [2011].

IV. SIMULATION

18. In this section we describe a simulation study for comparing the imputed SIPP job-level earnings using the SRMI model with the hot-deck method as it is done in production. For the simulation study, we use completed SIPP panel data from January 2004 to December 2005, also used in Benedetto and Stinson [2009] to evaluate their procedure. The dataset has one observation per person, per job, including person characteristics, job related variables (e.g. industry, occupation, firm size, usual weekly hours, type of job, union participation) and earnings. We begin with the 2004 SIPP data, and drop observations for which the respondents were less than 15 years of age in January, 2004, and those whose jobs were not in scope throughout the year. We then proceed to select a set of explanatory variables for our regression model (for example sex, age, race, education level, occupation, job type, etc.) We found some large differences in earnings in consecutive months for some individuals. We are using the unedited version of the data, thus these large fluctuations may be due to simple errors like reporting in the wrong units or keying errors. These types of errors are normally handled during editing processing. Since dealing with changing data due to edit failures is part of production and beyond the scope of this project, we flagged these observations with the assumption that this information denotes a wages bonus or a significant increase in earnings. We decided to create a new variable to assign an indicator for having received a wages bonus or a significant increase in earnings in any of the observed months and include this bonus indicator in the model. We then randomly select observations for which the job-level earnings are to be set to missing, and proceed to impute using the two methods being compared. Package **mi** imputes for the earnings indicator using logistic regression; if the monthly earnings indicator for a given individual is imputed to positive, then we impute the missing monthly earnings using SRMI.

19. The simulation study allows an objective evaluation of each model's performance and estimation of uncertainty due to missing values. Since we begin with a set of complete data, we know the "true" values and thus we can compute deviations from the truth. Also, since we are multiply-imputing, we are able to make use of Rubin's formulae (Rubin [1987]) to properly account for the variance due to imputation. Rubin's work shows that imputation adds variability to user calculated statistics. Because

traditional methods impute only once, imputers are not able to properly account for the variability due to imputation. Multiple imputation, in which missing values are replaced by $m > 1$ plausible values allows imputers to calculate the part of the total variance that is due to imputation and analysis of each of the m complete datasets could be combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.

20. For this simulation study, with m multiply-imputed data sets, we compute m estimates of both the individual and average monthly earnings along with their respective variance estimates as follows.

21. Let $\hat{Q}_j$ denote the estimate of the quantity of interest computed from completed data set j, and S_j^2 denote its variance. Overall estimates are then calculated as the average of the estimates over m. Specifically, the overall point estimate for Q is given by

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_j.$$

Following on Rubin [1987] work, to draw valid inferences about $\hat{Q}_j$ we must properly account for its variance. This variance has two components: Within-imputation variance and between-imputation variance. The within-imputation variance is the average of the m complete-data estimates for the variance,

$$W = \frac{1}{m} \sum_{i=1}^{m} S_j^2.$$

The point estimates, $\hat{Q}_j$, and their average, $\bar{Q}$, are used to calculate the variance of the statistic across the multiply-imputed data sets, i.e. between-imputation variance,

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{Q}_j - \bar{Q})(\hat{Q}_j - \bar{Q})\prime,$$

We use the within-imputation and between-imputation variances to calculate the estimated total variance,

$$T = W + \left(1 + \frac{1}{m}\right) B.$$

22. It has been demonstrated [Rubin, 1987, p. 114] that only a few imputations are needed to obtain efficient estimates unless the rate of missing values is very high, and thus we set m to 4 to create four completed datasets.

A. Comparison of SRMI and TEA's Randomized Hot Deck

23. In this section we present results of a simulation study to evaluate the effectiveness of these two models by comparing completed data using the model-based imputation method SRMI and the hot-deck method for imputing SIPP job-level earnings. While we believe that model-based imputation methods have the potential to improve SIPP earnings estimates, it is not possible to compare the model estimates to the production hot-deck estimates. We can however use **TEA**'s randomized hot-deck procedure to impute individual missing earnings with reported data from other individuals with similar characteristics and compare the completed data using **TEA**'s hot deck to the completed data using our SRMI imputation model. We take multiple draws from the hot-deck matrix so we can estimate the additional variance due to the hot-deck imputation.

24. We begin with the "truth" dataset containing the year of SIPP panel data we described previously and randomly select 100 sets of 10% of observations to be set to missing. For each of the 100 simulated datasets, we multiply-impute using the SRMI model as implemented in **mi** and the randomized hot deck as implemeted in **TEA**.

25. In order to evaluate each model's performance, we compute the average difference in root mean squared error (RMSE) over the 100 repetitions for each month. The RMSE can be useful in limited situations; other measures should be considered depending on how users intend to use the imputed datasets.

26. Suppose that we are interested in how accurately each model imputes individual earnings. Specifically, let Q_i denote the true reported earnings for individual i in a given month, $\hat{Q}^S_{i,j}$ be the corresponding estimate from our SRMI model (calculated using **mi**) and $\hat{Q}^T_{i,j}$ be the corresponding estimate obtained using **TEA**'s randomized hot-deck procedure. Each mean difference in Table 1 is calculated as

$$\frac{\sum_{j=1}^{r}\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Q}^S_{i,j}-Q_{i,j})^2}-\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Q}^T_{i,j}-Q_{i,j})^2}\right]}{r},$$

where r is the number of repetitions and n is the number of missing observations.

27. Table 1 displays the average difference in RMSE between the two procedures along with their standard errors. In each month, the SRMI model significantly outperforms the hot-deck procedure in RMSE; it indicates we obtain more accurate SIPP job-level earnings when imputing using the model-based imputation method when compared to the hot-deck imputes.We also evaluate each model's performance by comparing variances. Table 2 displays the average within-imputation, between-imputation and total variances and lower and upper 95% confidence intervals for each statistic by variable (mean earnings for each month) for each model. For each month, the SRMI estimates have significantly smaller variances than the hot-deck estimates, or the confidence intervals overlap.

TABLE 1. Average Difference in RMSE: (SRMI - Hot Deck)

Month	Mean_Diff	SE_Diff
Jan	-1257	141
Feb	-944	189
Mar	-2778	150
Apr	-1517	122
May	-2029	69
Jun	-2330	54
Jul	-2327	84
Aug	-2617	91
Sep	-2041	187
Oct	-4370	399
Nov	-1369	403
Dec	-1314	122

TABLE 2. Between-Imputation, Within-Imputation and Total Imputation Variance of Mean Monthly Earnings for Each Month (SRMI and Hot Deck)

Month	Variance	SRMI				Hot Deck			
		Mean	**SE**	**Lower**	**Upper**	**Mean**	**SE**	**Lower**	**Upper**
Jan	Between	12.30	1.02	10.29	14.31	115.05	11.52	92.45	137.65
	Within	576.06	8.98	558.45	593.67	664.64	39.28	587.57	741.70
	Total	591.44	9.11	573.57	609.31	808.45	45.75	718.69	898.21
Feb	Between	8.93	0.79	7.39	10.47	115.13	14.19	87.29	142.96
	Within	526.71	7.28	512.42	540.99	552.73	22.58	508.42	597.04
	Total	537.87	7.25	523.65	552.09	696.64	33.68	630.55	762.73
Mar	Between	14.10	1.63	10.90	17.29	471.01	43.44	385.78	556.23
	Within	1848.61	27.65	1794.36	1902.86	2275.61	47.76	2181.92	2369.31
	Total	1866.23	27.99	1811.31	1921.15	2864.37	91.06	2685.71	3043.03
Apr	Between	7.08	0.63	5.85	8.31	64.63	7.34	50.22	79.03
	Within	337.57	2.75	332.18	342.96	365.80	24.71	317.32	414.27
	Total	346.42	2.96	340.62	352.22	446.58	32.23	383.35	509.81
May	Between	7.38	0.67	6.07	8.68	58.63	4.76	49.28	67.97
	Within	292.76	1.06	290.68	294.85	320.46	20.01	281.20	359.73
	Total	301.99	1.35	299.34	304.63	393.75	23.13	348.36	439.14
Jun	Between	5.87	0.44	5.01	6.74	87.30	8.66	70.31	104.28
	Within	449.17	3.23	442.83	455.51	513.60	18.74	476.83	550.36
	Total	456.51	3.29	450.06	462.97	622.72	23.23	577.15	668.29
Jul	Between	6.99	0.52	5.97	8.02	56.86	5.60	45.88	67.83
	Within	404.85	1.28	402.33	407.37	407.72	18.13	372.15	443.29
	Total	413.59	1.42	410.80	416.39	478.79	22.33	434.98	522.60
Aug	Between	5.56	0.47	4.64	6.48	106.28	7.97	90.64	121.93
	Within	537.56	3.37	530.94	544.18	675.71	77.98	522.72	828.70
	Total	544.51	3.42	537.79	551.22	808.57	77.95	655.63	961.51
Sep	Between	5.69	0.54	4.64	6.74	71.97	5.57	61.05	82.89
	Within	454.17	7.81	438.85	469.48	452.18	20.35	412.25	492.11
	Total	461.28	7.85	445.87	476.69	542.13	21.18	500.57	583.69
Oct	Between	13.06	1.10	10.90	15.22	1075.56	141.02	798.88	1352.23
	Within	4866.71	157.37	4557.96	5175.46	8004.89	2264.23	3562.60	12447.18
	Total	4883.03	157.66	4573.71	5192.36	9349.34	2271.40	4892.97	13805.70
Nov	Between	6.98	0.57	5.86	8.10	88.24	8.34	71.89	104.60
	Within	595.11	45.34	506.15	684.07	605.85	72.62	463.38	748.33
	Total	603.84	45.31	514.94	692.74	716.16	75.12	568.77	863.54
Dec	Between	10.08	0.79	8.54	11.63	109.94	9.38	91.54	128.34
	Within	635.01	4.68	625.83	644.19	743.75	55.90	634.08	853.42
	Total	647.61	5.02	637.77	657.46	881.17	58.22	766.94	995.40

28. The SRMI model typically also produces better estimates of average monthly earnings, as shown in Table 3. That said, the randomized hot deck does a good job of estimating these statistics – the RMSE over the 100 repetitions ranges from $17.50 to $67.33, and the true average monthly earnings (as reported in this SIPP data) is near $2,500 in each month.

TABLE 3. RMSE of Mean Monthly Earnings (SRMI and Hot Deck)

Month	SRMI	Hot Deck	Difference
Jan	15.47	22.15	-6.67
Feb	17.90	19.95	-2.05
Mar	27.74	25.19	2.54
Apr	22.80	22.95	-0.15
May	18.64	20.87	-2.24
Jun	25.85	17.50	8.35
Jul	9.25	20.73	-11.48
Aug	23.08	24.27	-1.19
Sep	20.03	20.70	-0.67
Oct	33.66	67.33	-33.66
Nov	13.27	27.83	-14.55
Dec	38.47	26.70	11.76

V. CONCLUSION

29. The objective of this project is to offer alternative methods that improve the imputation of earnings in SIPP data. Currently, all SIPP variables are imputed by a deterministic hot-deck procedure whose shortcomings are discussed in Citro and Scholz [2009], and briefly in Section II of this paper. To address the shortcomings of the current hot-deck procedure for imputing SIPP earnings we propose using sequential regression multivariate imputation in a multiple imputation framework. Multiple imputation is particularly well suited for dealing with missing data because it allows for estimation of the variance due to imputation in terms of the within-imputation and between imputation components of the total variance. For this, we use the sequential regression multiple imputation model as implemented by Su et al. [2011]. Using this available R package for SRMI makes it possible to easily modify the model and apply it to other SIPP variables with minimal effort from the user. Model-based imputation methods have the advantage that they allow for easy incorporation of additional information into the model. Our results show that using a model-based approach to imputation is a feasible alternative to hot deck for imputing missing values in SIPP data and it has the potential to improve estimates of SIPP monthly earnings at the job-level.

30. We presented results of a simulation study for comparing the model-based method to hot-deck imputation, as it is traditionally used for imputing missing data in the SIPP. However, we do not compare with the hot-deck imputation as it is currently implemented for SIPP. SIPP data processing uses a deterministic hot deck to impute missing values while **TEA**'s randomized hot deck is not deterministic; that is, imputing values in repeated passes through the system will yield different imputations by simply changing the seed in the random number generator. **TEA**'s randomized hot deck has the added advantage of producing multiply-imputed data so we can estimate variances. Our results demonstrate that using model-based imputation for missing SIPP monthly earnings can result in average monthly

earnings that have smaller variances when compare to average monthly earnings imputed using the randomized hot deck imputation method. However, the randomized hot deck imputes plausible values from observed responses in the donor records and does a fair job of providing estimates of average monthly earnings.

31. There are two main advantages to using a model-based approach to missing data imputation in SIPP. First, there exists the possibility of incorporating any available auxiliary information into the model. Second, we can set up the model in a multiple imputation environment so we can estimate the variance introduced by imputation. On balance, we feel that model-based imputation for these data is feasible and should be further explored. This approach is not, however, without disadvantages. The SRMI is computationally intensive; run times for our computer program are extremely slow when compared with **TEA**'s randomized hot deck. For example, the R program implementing the SRMI model needs days to produce 400 mutliply-imputed SIPP datasets (from 100 simulated SIPP datasets). The same simulation runs using **TEA**'s randomized hot deck needs about 1 1/2 hours to produce the 400 replicates.

32. We want to close by noting there are limitations when using this type of simulation study. We started from a set of complete SIPP data, generated multiple repetitions of SIPP datasets with randomly imposed missing values for the variables in scope and multiply-imputed for each dataset using the two methods we described. This type of procedure allows us to measure how close the imputations are to the true missing values but provide no information on the frequentist properties of the estimators. To do so, Schafer and Graham [2002] recommend generating artificial data for which we know the target parameters and repeatedly (typically upward of 1,000 repetitions) draw samples and generate missing values on the samples. We could then impute using the two methods, compute estimates and intervals for the target parameter from each imputation method, and evaluate the quality of the point and interval estimates of the target parameter. Designing such a study is a major challenge because of the sheer size of the sample and the computational resources that are necessary to embed SRMI in a comprehensive simulation.

Acknowledgment

The authors would like to thank Jerry Maples, Joe Schafer, and Yves Thibaudeau for their valuable comments.

References

Gary Benedetto and Martha Stinson. *Testing New Imputation Methods for Earnings collected by the Survey of Income and Program Participation, FCSM 2009 Research Conference Papers*. 2009. URL http://www.fcsm.gov/09papers/Stinson_VII-C.pdf.

S.V. Buuren and K. Groothius-Oudshoorn. mice: Multivariate Imputations by Chained Equations in R. *Journal of Statistical Software*, 45(2):1–31, 2011. URL http://www.jstatsoft.org/v45/i02/.

Connie F. Citro and John Karl Scholz. *Reengineering the Survey of Income and Program Participation*. The National Academies Press, Washington, DC, 2009.

Brian Gough, ed. *GNU Scientific Library Reference Manual*. Network Theory, Ltd., 2003.

Ben Klemens. TEA for Survey Processing. UNECE Work Session on Statistiscal Data Editing, Oslo, 2012. URL http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/29_USA.pdf.

Ben Klemens. *Modeling with Data: Tools and Techniques for Statistical Computing*. Princeton University Press, 2008.

Trevillore E. Raghunathan, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodoloy*, 27(1):85–95, 2001.

Donald B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

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

Joseph L. Schafer and John W. Graham. Missing Data: Our view of the State of the art. *Psychological methods*, 7(2):147–177, 2002.

Yu-Sung Su, Andrew Gelman, Jennifer Hill, and Masanao Yajima. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.