Comparison of Imputation Methods in the Survey of Income and Program Participation

Sarah McMillan
U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC 20233

Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Abstract

The Survey of Income and Program Participation (SIPP) collects detailed information about income and program participation. These key questions can suffer from higher rates of nonresponse, so a sequential hot deck procedure is used to impute all items with missing data. Single imputation techniques like these do not incorporate the uncertainty of the imputation and can lead to underestimates of the true variance. However, if rates of item missing data are comparatively low, the underestimates may be slight with little impact on significance. Multiple imputation methods were used to impute missing data in the 2008 Panel of the SIPP to incorporate this uncertainty into the imputations. The imputations are made on a cross sectional basis (wave by wave) using several different multiple imputation methods and compared based on estimates of variance, as well as, Type I error, bias, mean square error, and fractions of missing information for key SIPP statistics. Additionally, we compare to similar estimates based on no imputation and based on various single imputation methods, like the current hot deck method.

Key Words: Imputation, hot deck, bias, multiple imputation

1. Introduction

Item nonresponse is a universal problem for survey analysts; however, “there is no single imputation method or statistical modeling technique that is optimal for all forms of item-missing data problems” (Heeringa et al, 2010). For this reason, many different methods are used in practice. Some of the most popular techniques include complete case analysis, weighting adjustments, and single or multiple imputation. Complete case analysis can lead to biased estimates if cases with missing data vary from cases with complete data in terms of the values of the analysis variables of interest. Weighting adjustments are generally used for unit nonresponse or monotone missing data that arise from respondents missing a complete phase of the questionnaire. Imputation, or filling in the missing values with observed or estimated values, provides a way to create a complete data set that can be analyzed and addresses some potential bias.

One single imputation method used, especially for those surveys at the U.S. Census Bureau, is hot deck imputation. Hot deck imputation involves filling in missing values with observed data from a similar respondent or unit. According to Andridge and Little (2010), hot deck imputation “can lead to gains in efficiency over complete case analysis … [and] reductions in nonresponse bias…” On the other hand, hot deck imputation, or any other single imputation method, treats imputed values as true values, which can “lead to serious underestimation of the true variance, when the proportion of missing values for an item is appreciable" (Rao and Shao, 1992).

Multiple imputation is fast becoming a preferred method of imputation. The general idea of multiple imputation is to create several completed data sets each with different imputations. This process adds uncertainty into the imputations and can lead to more accurate estimates of variance. According to Rubin (1996), "multiple imputation is substantially easier for the ultimate user than any other current method that can satisfy the dual objectives of reliance only on complete-data methods and general validity of inference." But multiple imputation in the context of complex sample designs is still being examined. Kim et al (2006) concluded that the multiple imputation variance estimator can be biased for certain domains for complex surveys.

The purpose of this project was to investigate the impact on variance estimates and bias using a multiple imputation method versus the current hot deck imputation method for the Survey of Income and Program Participation (SIPP). The SIPP uses a complex sample design and collects detailed information on income and program participation for the civilian non-institutionalized population living in the United States. Incorporating these complex design features, we compared means and standard error estimates calculated using three different methods of imputation: no imputation, hot deck imputation, and multiple imputation with all available observations. In addition, we analyzed the hot deck and multiple imputation methods on various metrics of bias using only the completed cases with an imposed missing data mechanism.

2. Data

1.1 The Survey of Income and Program Participation

The data used in this analysis were from the 2008 Panel of the Survey of Income and Program Participation[1]. The SIPP is designed to collect detailed information on income, assets, and program eligibility and participation for the civilian non-institutionalized population living in the United States. According to the SIPP Source and Accuracy Statement for 2008, a systematic selection was used to select housing units within 351 primary sampling units (PSUs) from the master address file created from the 2000 Decennial Census. In addition, households located in areas with a higher concentration of low-income households were oversampled by 44 percent to increase the accuracy of estimates for statistics of low-income households and program participation (U.S. Census Bureau, 2013).

When a respondent is interviewed, data is collected about the four preceding months. These four reference months comprise one wave. Only data from Wave 1, which covered the reference period from May 2008 to November 2008[2], were included in the analysis.

[1] Nonsampling errors in surveys may be attributed to a variety of sources, for further information on errors, statistical standards, and the computation and use of standard errors, go to http://www.census.gov/sipp/sourceac/S&A08_W1toW11(S&A-16).pdf

[2] The SIPP sample is divided into four equal groups, called rotation groups. A new rotation group is interviewed each month and asked about the four preceding months. Rotation group 1 of Wave 1 was interviewed in September 2008 and asked about May 2008 – August 2008. Rotation group 2 was interviewed in October 2008 and asked about June 2008 – September 2008. Rotation group 3 was interviewed in November 2008 and asked about July 2008 – October 2008 and rotation group 4 was interviewed in December 2008 and asked about August 2008 – November 2008.

The sample in Wave 1 consisted of approximately 65,500 households of which only 53,031 of the households were eligible for interview. Of those eligible households, 42,032 were interviewed, with a weighted response rate of 80.6% (US Census Bureau, 2013).

1.2 Analytic Variables

Differences in the three methods of imputation would be apparent for variables with a high rate of item-missing data. In general, missing data rates in the SIPP are low, but some sensitive questions, like those about income or assets, tend to have higher rates of missing data. Therefore, for the imputations, our main variable of interest was gross monthly wages, which is collected for each job reported in the four reference months. We summed the wages for all jobs as a monthly income variable. In addition, about 5% of respondents indicated they had a job in the reference month, but reported their wages as $0. Following a similar study by Stinson and Bennedetto (2009), we created a variable to indicate whether the respondent reported positive wages in the reference month. Additional variables of interest included the amount of income received from Social Security, Supplemental Security Income (SSI), and Supplemental Nutrition Assistance Program (SNAP) in each reference month. This resulted in twenty variables of interest to be imputed - four variables indicating positive wages, four wage variables indicating monthly wages for all jobs, and four variables for each of the specified programs. Table 1 summarizes the number of respondents in the analysis subpopulation, the number of respondents with observed values, and the frequency and rate of missing data.

Table 1: Subpopulation Totals and Rates of Missing Data

Variable of Interest	*Reference Month*	*Subpopulation Total*	*Frequency Observed*	*Frequency Missing*	*Percent Missing*
Gross Wages (All Jobs)	1	47,381	42,449	4,932	10.41
	2	47,381	42,562	4,819	10.17
	3	47,381	42,586	4,795	10.12
	4	47,381	42,506	4,875	10.29
Social Security	1	16,611	14,088	2,523	15.19
	2	16,611	14,087	2,524	15.19
	3	16,611	14,082	2,529	15.22
	4	16,611	14,063	2,548	15.34
Supplemental Security Income (SSI)	1	2,504	2,122	382	15.26
	2	2,504	2,128	376	15.02
	3	2,504	2,129	375	14.98
	4	2,504	2,125	379	15.14
SNAP	1	4,391	4,052	339	7.72
	2	4,391	4,066	325	7.40
	3	4,391	4,082	309	7.04
	4	4,391	4,098	293	6.67

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see *http://www.sipp.census.gov/sipp/source.html*.

3. Methods

3.1 Current Imputation Method

As described in the SIPP Users' Guide, the SIPP currently uses a hot deck imputation method that replaces individual missing data items with reported data from another person or household with similar characteristics. For each unit, edits and imputations are performed sequentially for each topical section: demographics, household characteristics, labor force, assets, general income, health insurance, and program participation. Sections are completely processed before moving to the next section. The hot deck arrays are created for each edited variable and stratified by age, race, sex, marital status, disability status and presence of own children. For imputation of income variables, industry occupation, sex, education level and number of hours worked are used to form the arrays. They are initialized with means of data from previous waves or similar surveys, but these are replaced with observed SIPP data of similar respondents on each pass through the data. Each pass contributes a new donor to the cell and each hot deck cell contains exactly one value at any point in the edit: either the initialized value or the most recently encountered good value meeting the same criteria for that cell - as defined by the stratifying variables. The hot deck imputation process, as currently implemented, is fully deterministic: subsequent re-processing using the same file and same edit program will result in identical imputations.

3.2 New Imputation Method

The general idea of multiple imputation, to create a number of complete data sets with different imputations, can be performed in a variety of ways. The approach advocated by Rubin (1987) is to create a multivariate normal regression model from the available data and draw from the posterior predictive model for the missing values to create the imputations for all variables. One can also take a simpler approach by repeating a hot deck imputation method to create several imputed data sets. For this analysis, we decided to use the sequential regression multivariate imputation (SRMI) model approach laid out in Raghunathan, et al (2001). The benefit of this method is the type of regression model varies with the type of variable being imputed, which is useful when imputing variables of different types. Because different types of variables were used in the analysis and our interest variables had skewed distributions, we thought this method would be easiest to implement.

For SRMI, imputation models are determined by stepwise regressions with a flat or non-informative prior distribution for the parameters in the regression model. More specifically, missing data are filled in with draws from a posterior predictive model that incorporates random variation. Predictors for the regression model can be any other variable that is completely observed, including those that have been previously imputed. Once all the variables have been imputed, the process is repeated to generate multiple complete datasets. The multiple imputations are interdependent and exploit the correlations among the variables.

3.3 Statistical Analysis

We performed two statistical analyses. The first analysis uses all respondent data, which allowed a comparison to the current imputations from the sequential hot deck. The second analysis focused only on those respondents with complete data which allowed us to estimate the bias associated with each method. Both analyses followed a similar process in forming the imputation model, performing the imputations, and estimating the key statistics.

3.3.1 All Respondent Data

When using multiple imputation it is better to include too many variables in the imputation model then to include too few (Rubin 1996). Possible predictors for the imputation model included demographic characteristics, job characteristics, geography, program participation, health related variables, final person weights, and stratification and clustering variables. Rubin (1996) also recommends including the complex sampling design features in the imputation model so that valid inferences are made when analyzing the multiply imputed data sets. To narrow down the extensive list of predictors, we looked at correlations and cross tabulations to test significant relationships. We also tested nonlinear relationships and interactions using residual plots and regression analysis. With the extensive list of predictors, only a few important interactions were tested and included in the imputation model. Following a study that involved family income variables by Schenker et al (2006), we performed a Box-Cox analysis on the monthly wage variables and examined residual plots to determine the best transformation to satisfy the normality assumption of the regressions before carrying out the imputations.

The MI IMPUTE command in STATA was used to perform the imputations of the key variables with missing data. We chose to perform twenty iterations and create five multiples.[3] For SRMI, you must identify the type of distribution and the model to be used. Continuous variables are imputed using either a linear regression model, a truncated linear regression for variables with a restricted range of values, or a predictive mean matching regression model. Binary and ordered variables are imputed using a logit regression model. The indicators of positive earnings were imputed first using a logit model, then based on these values the wage variables, were imputed using a normal linear regression.

Before any estimates could be made, the transformed wage variables were converted back to the original scale. The MI ESTIMATE command was used to estimate the mean personal monthly income for the employed population age 15 and over, as well as, the average monthly benefit for those receiving benefits from Social Security, SSI, and SNAP using the multiply imputed data sets. This procedure calculates weighted estimates and a variance estimate using Taylor Series Linearization for each individual data set and then combines the estimates to create one overall estimate of the mean and variance. We incorporated the final SIPP sampling weights and the complex sampling features using the stratification and cluster codes. The final sampling weights include adjustments for oversampling, adjustments for non-responding households, and a post stratification adjustment based on age, race, sex, Hispanic origin, and state. The final weights were the same for each process; they were not recalculated according to the new imputations.

In addition to these estimates, we computed the approximate fractions of missing information. Fractions of missing information are based on how much information is gained from the multivariate relationships in the imputation model over using observed

[3] The number of iterations determines the number of cycles missing values are imputed that build interdependence among imputed values and exploiting the correlational structure among covariates. The number of multiples determines how many complete datasets are created (Raghunathan et al, 2001).

values of the single variable. Fractions of missing information can range from zero to the rate of missing data for the variable. Higher values near the nonresponse rate indicate little information is gained from the additional variables in the imputation model. These were calculated according to the formula $FMI = \frac{\frac{M+1}{M} * B}{T}$ where M is the number of iterations of imputations, B is the between-imputation variance, and T is the overall variance.

We then calculated similar estimates for the complete case analysis, by excluding all cases with missing values, and for the data set containing the hot deck imputations. These calculations also included the complex design features and final sampling weights.

3.3.2 Complete Data

Many studies including Ziegelmeyer (2011) and Watson, et al (2011) compare different imputation methods using simulation studies restricted to the complete cases and impose a missing data mechanism to create the missing values. This allows for analysis of the bias since the true values are known. For this analysis, we only examined imputations for the wage variables. The population size of those receiving program benefits is extremely small, so estimates would be only for a very restricted population. We did not have access to the programs currently being used to form the hot deck imputations, so we used a weighted sequential regression technique in SAS with similar sorting and array variables (Ellis 2007). For the multiple imputation, we again used the SRMI method in STATA.

Our population of interest was all respondents that reported wages in all four reference months. We next had to impose a missing data mechanism on this population. This was done in two steps. First, we assumed the missing data mechanism is Missing At Random. Then we estimated a MAR mechanism using a logistic regression on a set of explanatory variables. From this regression we can obtain the likelihood p_i in (0,1) that the observation is missing [Ziegelmeyer]. To create multiple samples with missing data we had to incorporate an additional random process. An observation was coded as missing if $p_i > q_i$ where q_i was a random draw from a uniform distribution [0-*k*, 1-*k*) and *k* is the proportion of missing cases in the original dataset. This allowed us to create 25 different samples, each with an MAR mechanism and approximately 10% missing values for each of the wage variables. Each of these 25 datasets was then imputed using the two different methods and compared to the true values of the respondents.

Once these datasets were created, formation of the imputation model, performing the imputations, and the estimation of means and standard errors followed the same process as described in the previous analysis.

As another way of evaluating our methods we looked at several different metrics to examine different measure of bias. Following Ziegelmeyer (2011), we looked at the predictive accuracy and distributional accuracy.

The predictive accuracy measures how close the imputed value is to the true value. We used to following three measures:

i. Mean absolute deviation: $Abs\,Dev\left(\hat{Y}_\iota, Y_i^{true}\right) = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}_\iota - Y_i^{true}|$

ii. Square root of the mean square error: $RMSE\left(\hat{Y}_\iota, Y_i^{true}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_\iota - Y_i^{true})^2}$

iii. Mean relative deviation: $Rel\,Dev\left(\widehat{Y_\iota}, Y_i^{true}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{Y_\iota} - Y_i^{true}|}{Y_i^{true}}$

The distributional accuracy measures how close the distribution of the imputed values is to the distribution of the observed values.

i. Quartile 1 bias: $Q1bias\left(\widehat{Y_\iota}, Y_i^{true}\right) = \left(\widehat{Y_\iota}\right)^{Q1} - \left(Y_i^{true}\right)^{Q1}$

ii. Median bias: M$edian\ bias\left(\widehat{Y_\iota}, Y_i^{true}\right) = \left(\widehat{Y_\iota}\right)^{med} - \left(Y_i^{true}\right)^{med}$

iii. Quartile 3 bias: $Q3bias\left(\widehat{Y_\iota}, Y_i^{true}\right) = \left(\widehat{Y_\iota}\right)^{Q3} - \left(Y_i^{true}\right)^{Q3}$

Where *n* is the number of missing observations, $\widehat{Y_\iota}$ is the imputed value, and Y_i^{true} is the reported true value. For the multiple imputation estimates, $\widehat{Y_\iota}$ represents the average across the five imputations.

4. Results

4.1 All Respondent Data

4.1.1 Formation of the Imputation Model

Based on the correlation and regression analyses, we included 10 variables with significant coefficients in the imputation model for wage[4]. These included demographic variables such as age, race, sex, education level attained, Hispanic origin, marital status, and number of household members. We also included five job characteristics – number of hours worked per week, number of months in current occupation, size of current employer, occupational code[5], and type of work. Plots of residuals from regression analyses were used to test the linearity assumptions of the ordinal variables age, hours worked per week, and time in occupation. All of these parameters were significant in the model (and in a model incorporating the complex design features), but only the plot for age seemed to indicate a nonlinear trend. Interactions between age and time in occupation were included in the imputation model. Other interactions were tested but were not significant to include in the model.

The Box-Cox analysis determined that the correct power transformation for the gross wage variable was the logarithmic transformation. This transformation of gross wages for each reference month was done prior to any imputation and then transformed back before estimation. The transformation of the wage variable improved the normality and made it possible to only impute positive wage values.

4.1.2 Imputation

Because of the many steps needed to perform the imputations, we tried to re-examine our assumptions and determine the validity of each of our outputs. The first validity check for the imputations was to make sure convergence of all models was obtained. The MI package in STATA checks this criterion and will not create imputations if convergence isn't obtained. The next criteria we used were trace plots of the imputed values and standard errors. Trace plots graphically summarize the imputed values from the 20 iterations made before each multiple dataset is created. Patterns or trends seen in the trace plots could indicate misspecification of the imputation model, but no significant patterns

[4] All comparative statements in this report have undergone statistical testing, and, unless otherwise, noted, all comparisons are statistically significant at the five percent significance level.

[5] This is 4 digit code that was assigned according to the 2010 recode of the occupations for the American Community Survey 2007-2011 files.

or outliers were found. Additionally, we compared distributions of the imputed, observed, and complete data with imputations; these showed some lack of overlap for the imputed values at the higher end of the distribution.

Table 2 shows the different components of the variance – the between, within, and total variances - for each of the interest variables. It seems encouraging that for all of the key variables the results are consistent across the reference months. Other studies (Stinson and Benedetto 2009 and Erdman et al 2013) showed a wider range of values for the components of the variance. Table 2 also includes fractions of missing information which vary from 0.01 to 0.26. The order in which the variables were imputed could have an effect on the magnitude of these fractions, as variables imputed in the beginning of the process have fewer covariates to create imputations and so could have larger FMI values. It is interesting to note that items with the largest fractions of missing information also had some of the largest differences in standard error estimates between the methods.

Table 2: Summary of SRMI Imputation Variances

				Variance of the Mean			
Variable of Interest	*Reference Month*	*N*	*Mean*	*Within*	*Between*	*Total*	*FMI*
Wage	1	47,380	3,136.6	579.7	45.2	633.9	0.09
	2	47,381	3,155.0	520.0	59.2	591.0	0.13
	3	47,381	3,190.3	529.7	62.7	604.6	0.13
	4	47,381	3,213.2	504.5	31.5	542.3	0.07
Social	1	16,204	970.7	15.6	2.3	18.3	0.16
Security	2	16,204	973.7	15.4	1.9	17.7	0.14
	3	16,204	980.3	19.0	2.0	21.5	0.12
	4	16,204	983.5	17.1	3.5	21.4	0.22
SSI	1	2,362	507.0	58.9	14.8	76.6	0.26
	2	2,362	507.7	57.6	12.7	72.8	0.23
	3	2,362	510.2	57.6	14.5	74.9	0.25
	4	2,362	514.4	56.8	11.6	70.7	0.22
SNAP	1	4,397	194.4	9.0	0.3	9.3	0.04
	2	4,397	201.7	9.0	0.2	9.2	0.02
	3	4,397	212.0	9.2	0.2	9.4	0.03
	4	4,397	223.6	11.1	0.1	11.3	0.01

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

4.1.3 Estimation

Table 3 displays the overall means of wages and amounts received from Social Security, SSI, and SNAP, for each reference month and estimated standard errors for each of the three methods. For the overall means, only the hot deck estimates for wages were significantly different from the complete case estimates for reference months 1 and 2. All of the other estimates were not statistically different between the three methods. It is interesting to note that all methods had similar estimated standard errors. We would have expected hot deck estimates to be lower compared to multiple imputation, since a single imputation method underestimates the true variance.

Table 3: Comparison of Imputation Methods

Variable	*Ref*	*Complete Case*			*Hot Deck*			*SRMI*		
of Interest	*Month*	*N*	*Mean*	*SE*	*N*	*Mean*	*SE*	*N*	*Mean*	*SE*
Wages	1	42,448	3,122	24.5	47,380	3,190$^+$	23.7	47,380	3,137	25.2
	2	42,562	3,139	23.2	47,381	3,207$^+$	22.3	47,381	3,155	24.3
	3	42,586	3,173	24.3	47,381	3,235	23.4	47,381	3,190	24.6
	4	42,506	3,199	22.7	47,381	3,262	22.4	47,381	3,213	23.3
Social	1	13,677	975.1	4.31	16,203	977.2	4.24	16,204	970.7	4.28
Security	2	13,677	979.3	4.25	16,203	980.7	4.15	16,204	973.7	4.20
	3	13,672	986.5	4.64	16,203	986.5	4.46	16,204	980.3	4.63
	4	13,652	989.6	4.40	16,203	989.1	4.30	16,204	983.5	4.62
SSI	1	1,961	511.7	8.57	2,343	514.0	7.45	2,362	507.0	8.75
	2	1,967	512.1	8.56	2,343	514.6	7.41	2,362	507.7	8.53
	3	1,968	514.6	8.60	2,343	517.0	7.45	2,362	510.2	8.66
	4	1,964	519.2	8.48	2,343	520.3	7.30	2,362	514.4	8.41
SNAP	1	4,051	196.5	3.06	4,391	200.1	3.10	4,397	194.4	3.05
	2	4,065	204.4	3.09	4,391	206.8	3.06	4,397	201.7	3.03
	3	4,080	214.3	3.03	4,391	216.2	3.13	4,397	212.0	3.07
	4	4,094	226.6	3.38	4,391	226.7	3.39	4,397	223.6	3.36

$^+$ Denotes a statistically significant difference between the mean using the hot deck method as compared to the Complete Case methods.

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

Estimates for selected subgroups are summarized in Table 4. Following a study by Stinson and Benedetto (2009), we looked at the subgroup of single black females age 18-25, since this group has a high imputation rate. The SRMI estimate for mean monthly income for single black females 18-25 was statistically significant compared to the hot deck and the complete case estimate. The majority of the other subgroups examined had no significant differences in means. The standard errors for these subgroups do show a greater variation among the subgroups then the overall means, but the order depends on the subgroup. For example, the standard error for wages of white males was 34.12 for hot deck, 37.25 for complete case, and 40.43 for SRMI. However, the standard error estimates for wages of Asian males are 195.15 for hot deck, 220.13 for complete case and 214.20 for SRMI.

Table 4: Comparison of Imputation Methods On Wage for Particular Subgroups (Reference Month 1)

		Complete Case			*Hot Deck*			*SRMI*		
Subgroup of Interest		*N*	*Mean*	*SE*	*N*	*Mean*	*SE*	*N*	*Mean*	*SE*
White	Male	17,427	3,770.0	37.25	19,436	3,845.1	34.12	19,437	3,811.8	40.43
	Female	16,637	2,582.3	32.45	18,449	2,635.4	33.18	18,449	2,577.7	31.98
Black	Male	2,090	2,642.2	70.20	2,408	2,730.6	69.62	2,408	2,658.9	76.87
	Female	2,793	2,227.0	48.22	3,195	2,363.7	59.07	3,195	2,234.0	50.54
Asian	Male	978	4,678.9	220.13	1,100	4,684.4	195.15	1,100	4,618.3	214.20
	Female	938	3,193.4	96.11	1,063	3,266.8	82.06	1,063	3,154.5	95.08
Other	Male	807	2,923.6	93.30	874	2,941.7	89.56	874	2,910.2	123.56
	Female	778	2,182.2	110.79	855	2,224.3	100.69	855	2,166.9	103.33
BLK[1]	0	42,064	3,220.6	23.04	46,876	3,211.1	23.92	46,807	3,162.7	25.44
	1	442	1,223.4	56.79	504	1,265.6	68.86	574	1,005.9*	47.56

[1] BLK is in indicator representing the subpopulation of single black females aged 18-25.

* Denotes a statistically significant difference between the mean using the SRMI method as compared to the Hot Deck and Complete Case methods.

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

4.2 Complete Data

4.2.1 Formation of the Imputation Model

The imputation model for the completed cases was formed through the same process as for the previous analysis and similar results were achieved in the correlation and regression modeling. Demographic variables such as age, gender, race, and education, as well as, job related variables were included in the imputation model. Interactions and nonlinear relationships were also tested and significant relationships were included.

4.2.2 Imputation

Convergence for all models was obtained and we again examined the trace plots of the imputed values and standard errors to check for patterns in the imputation chains. No patterns were found. Distributions of the imputed values compared to the observed and complete data again showed lack of overlap for higher earners.

Table 5 shows the between, within, and total variances for each of the wage variables. As compared to using all the respondent cases, the between variances are much lower for this analysis. Also, note that the smallest variances are seen for month 4 and the largest for month 3 in both analyses. This could be due to the order of the imputations, since those that are imputed later can benefit from all previous imputations.

Table 5: SRMI Imputation Variances - Complete Data

Variable of					*Variance of the Mean*		
Interest	*Ref Month*	*N*	*Mean*	*Std Err*	*Within*	*Between*	*Total*
Wages	1	42,003	3,148.8	24.39	592.69	1.90	594.98
	2	42,003	3,171.6	23.52	551.96	1.17	553.36
	3	42,003	3,204.8	24.46	591.82	5.46	598.38
	4	42,003	3,219.2	23.24	539.34	0.70	540.17

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

4.2.3 Estimation

Table 6 shows the means and standard errors for each of the methods tested and for the true values. There were no statistically significant differences in the means; but unlike the last analysis, there does seem to be a slight underestimation of the overall true variance using the hot deck method.

Table 6: Comparison of Imputation Methods (Complete Data)

Variable	*Ref*		*True Values*		*Hot Deck*		*SRMI*	
of Interest	*Month*	*N*	*Mean*	*SE*	*Mean*	*SE*	*Mean*	*SE*
Wages	1	42,003	3,150.2	24.70	3,149.7	18.53	3,148.8	24.39
	2	42,003	3,171.3	23.50	3,170.7	18.00	3,171.6	23.52
	3	42,003	3,204.2	24.60	3,204.6	19.68	3,204.8	24.46
	4	42,003	3,223.5	22.99	3,223.5	19.69	3,219.2	23.24

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

The two following tables summarize the evaluation of bias. For predictive accuracy, the SRMI method has much lower estimates for absolute and relative deviation, as well as, root mean square error. This seems to indicate the SRMI method imputes values closer to true values. Significant improvement of around 80% can be seen for absolute and relative deviation for all reference months and improvements in root mean square error are around 60 to 70%.

Table 7: Comparison of Metrics (Predictive Accuracy)

Variable			*Hot Deck*			*SRMI*			*SRMI / Hot Deck*	
of	*Ref*	*Abs*		*Rel*	*Abs*		*Rel*	*Abs*		*Rel*
Interest	*Month*	*Dev*	*RMSE*	*Dev*	*Dev*	*RMSE*	*Dev*	*Dev*	*RMSE*	*Dev*
Wages	1	2,389.1	4,708.3	236.9	433.9	1726.5	63.8	18%	37%	27%
	2	2,348.6	4,682.2	210.1	365.2	1281.1	37.9	16%	27%	18%
	3	2,354.0	4,820.5	189.0	389.4	1440.5	38.8	17%	30%	21%
	4	2,368.1	4,821.3	184.3	511.0	1863.6	63.9	22%	39%	35%

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

Looking at the distributional metrics, it seems that for month 1, the weighted sequential hot deck performed significantly better at maintaining the original distribution. But, notice that for both methods we are mostly underestimating each of the percentiles, especially the 75th percentile. This could indicate our model needs improvement to fully capture the higher earners.

Table 8: Comparison of Metrics (Distributional Accuracy)

Variable	*Ref*	*Hot Deck*			*SRMI*			*SRMI / Hot Deck*		
of Interest	*Month*	*Q1*	*Median*	*Q3*	*Q1*	*Median*	*Q3*	*Q1*	*Median*	*Q3*
Wages	1	0.3	-1.0	-3.4	64.1	-49.5	-40.1	20031%	4855%	1178%
	2	-43.5	-7.7	-21.9	0.0	12.6	-34.9	0%	-163%	159%
	3	-79.3	-32.7	-38.4	-14.1	7.2	-56.9	18%	-22%	148%
	4	-78.6	-36.6	-46.8	61.8	-45.7	-56.8	-79%	125%	121%

SOURCE: U.S. Census Bureau, Survey of Income and Program Participation, Wave 1, 2008. For information on confidentiality protection, sampling and nonsampling error see http://www.sipp.census.gov/sipp/source.html.

5. Discussion

From both analyses, the majority of means showed no significant differences across the different methods. Even for the complete case analysis we do not see significant differences as compared to the two imputation methods. Especially for the second analysis of only complete data with imposed missingness, it seems that both methods do well at estimating the means of subgroups as compared to the true values for a range of subpopulations.

From our analysis of all respondent data, we did not see the underestimation of variance for the current hot deck method that we had expected. Like, Benedetto and Stinson (2009) we did find significant differences for a specific subgroup with a high imputation rate. Possibly expanding the imputation to all variables with missing data could make some differences more apparent, since multiple imputation builds off all the covariates in the model, different imputation in the covariates could lead to a wider range of imputed values.

For the analysis of complete data with imposed missingness, it was interesting to note that we did see a difference in standard errors for the hot deck method compared to the true values and the multiple imputation method. This could be due to a lack of richness in the hot deck arrays, since fewer values are available from the completed cases to populate the arrays. Further inspection into this reasoning is needed. It is also worth a closer examination as to why the magnitude of the between variance changed dramatically between the two analyses. Further examination may indicate the need for a better fitting model, or a different missing data mechanism then the one imposed. Similar simulation studies on SIPP wages by Erdman et al (2013) had a wide range of within and between variance estimates, so the data for some months may not be missing at random, leading to varying estimates of the different components.

The additional metrics helped us examine the bias associated with the two methods. The predictive metrics indicate that the SRMI method creates imputations closer to the true values than the hot deck method. Similar reductions in root mean square error were seen over hot deck imputation in Erdman et al (2013). As for the distribution accuracy, the hot deck method was closer to maintaining the original distribution, but both could use improvement, especially for the 75th percentile. We did also see this lack of fit when we examined the distribution of values after the imputations, possibly adding additional covariates or using a multi-level model would help improve fit.

Although these results help us in comparing the different methods of imputation, there are still several limitations and many areas of future research. One main area of improvement is our imputation model. As noted earlier, Rubin (1996) advocates erring on the side of too many variables than too few. The SIPP seems to have a very rich covariate structure which needs to be fully utilized when forming the imputation models. Exploration into these relationships could help improve the fit of our model. Also, the transformation of the wage variable may have an effect on the relationship between the variables used for the imputation or used in related analyses. Using a predictive mean matching model instead of parsing out zero earners from those with positive wages could also lead to a better fit and eliminate the need for transforming the wage variable, since predictive mean matching only imputes observed values. We did try to do several different diagnostic tests to determine the validity of our imputation and of our model, but new research into better methods seem to be forthcoming. Another area of potential improvement are exploring new methods being proposed to better incorporate the complex sample design features, as well as, examining the bias for different domains under a complex sample alluded to by Kim et al (2006).

6. Conclusion

As with other studies involving multiple imputation methods on wage variables for SIPP (Stinson and Benedetto (2009) and Garcia et al (2013)), there is evidence that the SIPP could benefit from improvements in predictive accuracy over the current hot deck methods by using a multiple imputation method. However, only small differences in estimates of means and standard errors were found. Further exploration into a wider range of variables and rates of missingness and their effect on estimates could be a beneficial next step.

References

Andridge, R.R. and Little, R.J.A (2010) “A Review of Hot deck Imputation for Survey Non-response,” *International Statistical Review*, 78(1), 40-64.

Benedetto, G. and Stinson, M. (2009) “Testing New Imputation Methods for Earnings collected by the Survey of Income and Program Participation” Proceedings for Federal Committee on Statistical Methodology, Washington DC, October 19, 2009. Obtained from http://www.fcsm.gov/09papers/Stinson_VII-C.pdf on June 3, 2013.

Erdman, C., Garcia, M. and Klemens, B. (2013) “A Comparison of Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation.” Center for Research and Statistical Methodology White Paper Series, U.S. Census Bureau.

Ellis, B. (2007) “A Consolidated Macro for Iterative Hot Deck Inputation” Proceedings from Northeast SAS Users Group (NESUG), Baltimore, MD, November 11-14,

2007. Obtained from http://www.nesug.org/proceedings/nesug07/po/po03.pdf on June 1, 2013.

Heeringa, S.G., West, B.T., and Berglund, P.A. (2010) *Applied Survey Data Analysis*, Boca Raton: Chapman & Hall/CRC.

Kim, J.K., Brick, J.M., Fuller, W.A., and Kalton, G. (2006) “On the Bias of the Multiple Imputation Variance Estimator in Survey Sampling,” *Journal of the Royal Statistical Society*, 68, 509-521.

Little, R.J.A, and Rubin, D.B. (2002) *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley.

Marchenko, Y (2011). “Chained equations and more in multiple imputation in STATA 12” Presentation at 2011 UK STATA Users Group Meeting. Obtained from http://www.stata.com/meeting/italy11/abstracts/italy11_marchenko.pdf on May 13, 2013.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J, and Solenberger, P. (2001), “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models,” *Survey Methodology*, 27, 85-95.

Raghunathan, T.E., Van Hoewyk, J, and Solenberger, P. (2002), “IVEware:Imputation and Variance Estimation Software User Guide,” The Regents of the University of Michigan. Obtained from ftp://ftp.isr.umich.edu/pub/src/smp/ive/ive21_user.pdf on September 10, 2012.

Rubin, D.B. (1996), “Multiple Imputation After 18+ Years” *Journal of the American Statistical Association*, 91, 473-489.

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

Schenker, N., Raghunathan, T.E., Chiu, P., Makuc, D.M., Zhang, G., and Cohen, A.J. (2006) “Multiple Imputation of Missing Income Data in the National Health Interview Survey,” *Journal of the American Statistical Association*, 475, 924-933.

US Census Bureau Memorandum (2013), *Survey of Income and Program Participation (SIPP) 2008 Panel: Source and Accuracy Statement for Wave 1 to Wave 11.* Obtained from http://www.census.gov/sipp/sourceac/S&A08_W1toW11(S&A-16).pdf on April 15, 2013.

Watson N. and Starick, R. (2011), “Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey” Journal of Official Statistics, Vol. 27(4), 693-715.

Ziegelmeyer, M. (2013). "Illuminate the Unknown: Evaluation of Imputation Procedures Based on the SAVE survey," AStA Advances in Statistical Analysis, Springer, Vol. 97(1), 49-76.