Westat®

Improving lives
through research™

Feasibility of Imputing Assets and Liabilities Account Values in the Consumer Expenditure Surveys

Author

Ismael Flores Cervantes
J. Michael Brick

October 6, 2021

Prepared for:

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Table of Contents

Chapter **Page**

Executive Summary

The goal of this research was to evaluate whether it was feasible to impute all items related to Asset and Liability (A&L) data collected in the Consumer Expenditure (CE) Surveys and to produce point and confidence interval estimates that would appropriately reflect the imputation of these missing data. We developed and evaluated three multiple imputation methods for this purpose and compared the performance of these methods in a simulation study. The methods are 1) Imputation based on the income imputation currently implemented in the CE, 2) a parametric imputation method based on chained equations, and 3) and statistical learning imputation method based on random forests and chained equations. We produced quarterly and annual estimates of the assets and liabilities (A&L) variables collected in the CE using the restricted-use data files for the three methods. The estimates were produced following the approach used in the production of estimates in the CE, which required imputing missing values by quarter and producing annual estimates for the entire year. We compared estimates of proportion and totals of A&L indicators (i.e., flags that indicate if the consumer unit (CU) holds the account and means and totals for A&L account values (variables with the dollar amount of the account for those CUs who report the account).

The performance of the multiple imputation methods was evaluated using repeated sampling in a Monte Carlo simulation. The simulation population and design were constructed to be consistent with the Consumer Expenditure Surveys design to the extent possible (using the actual population was not feasible for a number of reasons discussed later in the report. In particular, the values for the A&L data and the missing data patterns in the simulation were created using empirical distributions of these quantities from the actual data files, although some changes were implemented to facilitate the simulation process.

The key findings of the simulation study were:

- All three imputation methods performed well and showed that all the methods are feasible for imputing A&L data. The estimates of proportion and totals of A&L indicators were nearly unbiased, with confidence intervals close to the nominal level. The estimates of means and totals for A&L account values and their confidence interval were less accurate, largely due to bias in the point estimates. However, for the most part, the biases were small, and the confidence intervals were only slightly lower than the nominal level, but there were exceptions with some estimates that had big biases and lower than desired confidence intervals.

- None of the three multiple imputation methods was clearly the best in terms of statistical criteria. As a result, we recommend criteria for choosing an imputation method based on other criteria such as operational efficiency and convenience.

- Based on the statistical properties of the estimators produced by the proposed methods, the findings of the simulation, the programing language characteristics, and the efficiency of the implementation in production, we recommend the multiple imputation method based on the income imputation methodology in the CE with some important modifications. The rationale for recommending this method is detailed in the final section of the report. The changes suggested are the result of the evaluation of the other methods examined. In particular, we suggest implementing additional steps to address accounts with zero balances. Because of the number of A&L variables to impute, other suggestions include developing diagnostics to determine that the imputed values are generated within an expected range of values and without errors. We also suggest evaluating the software currently used for income imputations since much has changed since this method was first implemented in the CE.

- Since the emphasis of this research was on feasibility, we also recommend that additional research be undertaken to improve upon the inferences from the imputed data once a method is chosen. For example, if domain estimates are very important, then more research on how to reduce the bias of the imputed domain estimates might be very useful. Similarly, additional research for dealing with imputing missing values for those holding an account with a zero balance might be useful.

1. Research Goals

The objective of the research under the Task Order 13, GS-00F-009-DA, SIN 874-1, BPA No. 1625DC-20-F-00033 is the evaluation of the feasibility of the imputation of all items related to Asset and Liability (A&L) data collected in the Bureau of Labor Statistics (BLS) Consumer Expenditure Surveys (CE). The task consists of implementing and evaluating three imputation methods consistent with how the CE Interview Survey data is collected (i.e., quarterly), weighted, and released (annually). Westat performed the following activities:

- Reviewed the previous research conducted by the BLS for the imputation of assets and liabilities variables.

- Reviewed and explored the restricted use CE file, codebook, instrument, and the information related to the survey and A&L data (Section 1). The review included the analysis of the missing value patterns of the A&L variables.

- Proposed three methods for imputing the missing items of the A&L accounts: (1) a method based on the imputation procedure currently used by the BLS for imputing Income-related items; (2) a method based on Multiple Imputation using Chained Equation (MICE) using linear regression; and (3) a method based on MICE using Random Forests (RF), a more recent approach form Statistical Learning (SL) (Section 2).

- As an example of the methods, imputed the A&L data items using 2019 restricted-use data with the three imputation methods mimicking the BLS operational procedure where missing data are imputed quarterly, and annual estimates are produced (Section 2).

- Evaluated the three imputation methods using simulation after creating synthetic data that exhibited the type of patterns of missing data consistent with the restricted data set (Section 3).

The details of these efforts are presented in remaining sections this report.

1.1 The Consumer Expenditure Surveys

The CE program provides data on expenditures, income, and demographic characteristics of consumers in the United States. The US Census Bureau collects the CE for the BLS in two surveys, the Interview Survey for major and recurring items and the Diary Survey for minor or frequently purchased items. The source data for this research is the Interview Survey. These surveys are the only federal government surveys that provide information on the complete range of consumers' expenditures and their financials by demographic characteristics.

The CE data are primarily used to revise the relative importance of goods and services in the market basket of the Consumer Price Index. As in most current surveys, the surveys are subject to unit and item nonresponse. Nonresponse has been addressed for some of these items in the past, and CE users and stakeholders use imputed data to make valid statistical inferences from the data for these items. For example, multiple imputation of missing data values for income data have been available since 2004. The imputed income data allows the publication of income data for all consumer units instead of just those who responded to all the sources of income.

1.1.1 The CE Sample Design

The CE is a stratified multi-stage sample where the geographic areas or primary sampling units (PSUs) are drawn with probability proportional to size in the first stage within strata. The PSUs consist of geographic areas or clusters of counties based on the 2012 "core-based statistical areas" (CBSAs) defined by the Office of Management and Budget (OMB). The CE first stage sample consisted of 91 PSU drawn from 3 strata beginning in 2015. The number of PSUs per stratum and stratum definitions are

- Stratum S with PSUs in metropolitan CBSAs with a population of over 2.5 million persons with a sample of 23 PSU (self-representing urban PSUs/strata)

- Stratum N with the PSUs in remaining metropolitan CBSAs and the micropolitan CBSAs with a population under 2.5 million persons with a sample of 52 (non-self-representing urban PSUs/stratum)

- Stratum R with of the PSUs in non-CBSA areas with a sample of 16 PSUs (rural PSUs/stratum)

The 23 PSUs in Stratum S correspond to the largest CBSAs in the country and were selected with certainty (i.e., the probability of selection of the PSU in the first stage is 1). Each self-representing PSU can be considered a sampling stratum. In contrast, the PSUs in Strata N and R covered smaller CBSAs and were sampled with probability proportional to the population in the PSUs.

In the second stage, quarterly rotating samples (i.e., a panel) of consumer units (CU) are selected, retained, and then replaced after one year. Lists of addresses in the selected PSUs are compiled for the second stage sampling using the Census Bureau's Master Address File (MAF) with the residential addresses identified in the 2010 census updated twice per year with the U.S. Postal Service's Delivery Sequence File and supplemental lists of housing units owned or managed by organizations for residents in group arrangements such as college dormitories and retirement communities.

The total sample that is interviewed each period is divided into 4 panels with approximately one-fourth of the sample being introduced each calendar quarter and one-fourth of exiting the survey, after completing the fourth interview, in a rotating panel design.

Each quarter file has approximately 1,500[i] responding CUs during the 2017 to 2019 time period we examined, but the number varies slightly from one quarter to the next. Each quarter, a sample of addresses needed to produce approximately 1,500 responding CUs is drawn to form a new panel that replaces an existing panel of CUs contacted in the previous year. Each panel is retained in the sample for one year. The A&L-related items are collected during the last interview of the panel. As a result, in each quarterly data file, the A&L items are available for one-fourth of the CUs in the file. For the creation of an annual estimate of totals from any single quarter of data, a weighting adjustment of a factor of 4 would need to be applied. In production, annual estimates of the A&L variables are created by combining the data files of the 4 quarters for the year.

To produce estimates using the CE data, sampling weights are created as the product of the inverse of the probability of selection of the PSU, the inverse of the probability of selection of the address in the sampled PSUs, and the nonresponse adjustment for those CUs that did not respond to the survey. Any estimate of totals produced using the quarterly files needs a weighting adjustment (a factor of 4) since these items are asked on a fourth of the CUs in the file. However, such adjustment is not needed for annual estimates where the four quarterly files are combined.

1.1.2 Estimation of Multiple Imputed variables in the CE

In the CE, the missing values of income-related variables are imputed using a linear regression procedure described in User's Guide to Income Imputation in the CE (bls.gov). The estimates use multiple imputation (MI) (Rubin, 1987, 2014, and 2004) to reflect the uncertainty from the sample design and the imputed missing values due to item nonresponse. The MI procedures in the CE generate *m*=5 imputed values for each missing item. The imputed values are used to produce both the estimates and the variances estimates using Rubin’s rules presented in Table 1-1.

Table 1-1. Rubin's formulas for multiple imputation estimators.

Estimator	Expression_
MI estimate of Q	$\hat{Q}=\frac{1}{m}\sum_{l=1}^{m}\hat{Q}_l$
MI estimate of the variance of $\hat{Q}$	$\hat{V}\left(\hat{Q}\right)=\hat{\bar{U}}+\left(1+\frac{1}{m}\right)\hat{B}$

where $\hat{Q}$ is the statistic, $\hat{V}\left(\hat{Q}\right)$ is the variance estimate of $\hat{Q}$, where $\hat{\bar{U}}$ is the estimate of the within-imputation variance computed as the average of the variance estimates $\hat{U}_l$ of $\hat{Q}_l$ for $l=1,\ldots,m$ where m is the number of repeated imputed values, $\hat{\bar{U}}=\frac{1}{m}\sum_{l=1}^{m}\hat{U}_l$, and $\hat{B}$ is the between-imputation variance estimate of $\hat{Q}$ calculated as $\hat{B}=\frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{Q}_l-\hat{Q}\right)^2$.

We refer to MI as the method developed by Rubin (e.g., Rubin’s rules) to compute estimates and variance estimates that incorporate the uncertainty from the imputation using a set of repeated imputations. For example, the variance estimate of income in the CE uses the five imputed income values of those income-related variables of respondents who did not provide a valid response. In general, MI does not fully specify the method for how the repeated imputations are generated. Methods range from hot-deck (provided care is taken in drawing repeated donors from the donor set) to parametric methods where an assumed distribution for the item to be imputed is used to generate the repeated values.

1.2 A&L Account Variables in the CE

The CE collects 48 variables related to the CU’s A&L holdings (e.g., accounts). The first 16 variables listed in Table 1-2 corresponds to the flags (e.g., yes or no) that indicate if the CU held the account or not at the time of the interview and in the previous year.

Table 1-2. Asset and Liability Indicators the Consumer Expenditure Surveys.

Description	Time	Variable	Question
Assets			
IRA, retirement accounts such as 401(k)s, IRAs, Thrift Savings Plans	Current	IRA	Do you have any retirement accounts such as 401(k0s, IRAS, Thrift Saving Plans?
	Last year	IRAYR	Did you have any retirement accounts such as 401(k)s, IRAs, Thrift Savings Plans ONE YEAR AGO TODAY?
Stocks, bonds, or mutual funds	Current	STOCK	Do (you/you or any members of your household) have any directly held stocks, bonds, or mutual funds (Not in Retirement accounts)? Include US savings bonds
	Last year	STOCKYR	Did you have any directly held stocks, bonds, or mutual funds one year ago?
Checking accounts, Savings accounts, money market, CDs	Current	LIQUID	Do you have any checking, saving, money market accounts, certificates of deposit, or CDs?
	Last year	LIQUIDYR	Did (you/you or any members of your household) have any checking, savings, money market accounts, or certificates of deposit or CDs ONE YEAR AGO TODAY?
Whole life insurance or other life insurance policies that can be surrendered for cash or borrowed against prior to the death of the person insured	Current	WHOLIF	(Do/Does) (you/your household) own any whole life insurance or other life insurance policies that can be surrendered for cash or borrowed against prior to the death of the person insured? Also, include universal life and variable life insurance. Do NOT include term life insurance or other policies that only have a benefit upon death or disability
	Last year	WHLFYR	Did you own any whole life insurance or other life insurance policies that can be surrendered for cash or borrowed against prior to the death of the person insured one year ago today?
Any other financial assets, such as annuities, trusts, and royalties	Current	OTHAST	Do/Does) (you/your household) have any other financial assets, such as annuities, trusts, and royalties?
	Last year	OTHSTYR	Did you have any other financial assets, such as annuities, trusts, and royalties on a year ago today?

Table 1-2. Asset and Liability Indicators the Consumer Expenditure Surveys. (continued)

Description	Time	Variable	Question
Liabilities			
Any credit cards, including store cards and gas cards	Current	CREDIT	Could you tell me which range that best reflects the total amount owed on all major credit cards, including store cards and gas cards?
	Last year	CREDTYR	Did you have any credit cards, including store cards and gas cards, one year ago today?
Any student loans	Current	STUDNT	(Do/Does) (you/your household) have any student loans?
	Last year	STDNTYR	Did you have student loans one year ago today?
Any other debt such as medical loans or personal loans	Current	OTHLON	As of today, do you have any other debt, such as medical loans or personal loans?
	Last year	OTHLNYR	Did (you/your household) have any other debt such as medical loans or personal loans ONE YEAR AGO TODAY? Do not include mortgages, home equity loans, or vehicle loans

The second set of A&L-related variables listed in Table 1-3 describe the value (in dollars) at the time of the interview and in the previous year if the CU holds such account (the values of the account are conditional on the flags indicating that the CUs have the account). In addition to the 32 variables in these two tables, there are 16 bracket variables (8 for the current year and 8 for the previous year) that are used when the respondent has an account but does not give the exact value of the account holding (i.e., provides partial information on the range of the value).

Table 1-3. Asset and Liability Account values in Consumer Expenditure Surveys.

Type	Description	Time	Variable	Question
Assets	IRA, retirement accounts such as 401(k)s, IRAs, Thrift Savings Plans	Current	IRAX	As of today, what is the total value of all retirement accounts, such as 401(k)s, IRAs, and Thrift Savings Plans that you own?
		Last year	IRAYRX	What was the total value of all retirement accounts one year ago today?
	Stocks, bonds, or mutual funds	Current	STOCKX	As of today, what is the total value of all directly-held stocks, bonds, and mutual funds?
		Last year	STOCKYRX	What was the total value of all directly-held stocks, bonds, and mutual funds one year ago today?
	Checking accounts, Savings accounts, money market, CDs	Current	LIQUIDX	As of today, what is the total value of all checking, savings, money market accounts, and certification of deposit or CDs you have?
		Last year	LIQUIDYRX	What was the total value of all checking, savings, money market accounts, and certificates of deposit or CDs ONE YEAR AGO TODAY?
	Whole life insurance or other life insurance policies that can be surrendered for cash or borrowed against prior to the death of the person insured	Current	WHOLIFX	As of today, what is the total surrender value of these policies?
		Last year	WHLFYRX	What was the total surrender value of these policies one year ago today?
	Any other financial assets, such as annuities, trusts, and royalties	Current	OTHASTX	As of today, what is the total value of these other financial assets?
		Last year	OTHSTYRX	What was the value of these other financial assets one year ago today?
Liabilities	Any credit cards, including store cards and gas cards	Current	CREDITX	What is the total amount owed on all cards?
		Last year	CREDTYRX	What was the total amount owed on all cards one year ago today?
	Any student loans	Current	STUDNTX	What is the total amount owed on all student loans?
		Last year	STDNTYRX	What was the total amount owed on all student loans one year ago today?
	Any other debt such as medical loans or personal loans	Current	OTHLONX	What is the total amount owed on all other loans?
		Last year	OTHLNYRX	What was the total amount owed on all other loans ONE YEAR AGO TODAY? Do not include mortgages, home equity loans, or vehicle loans

1.2.1 Data collection of A&L related variables and generation of missing values

Figure 1-1 shows the protocol for the collection of the A&L variables in the CE: (1) the respondent is asked if the CU currently has the A&L account (e.g., at the time of the interview); if the answer is yes, the numeric value of the account is asked in (2). If the respondent does not provide a valid value, then the CU is asked to identify one out of five ranges that contain the A&L value in (3). The range values differ depending on the type of account. If either a range or a value is provided, then questions about the value of last year's A&L holding are asked in (4). If the respondent does not provide a valid value, then the CU is asked to identify one out of five ranges that contain last year's A&L account value in (5).

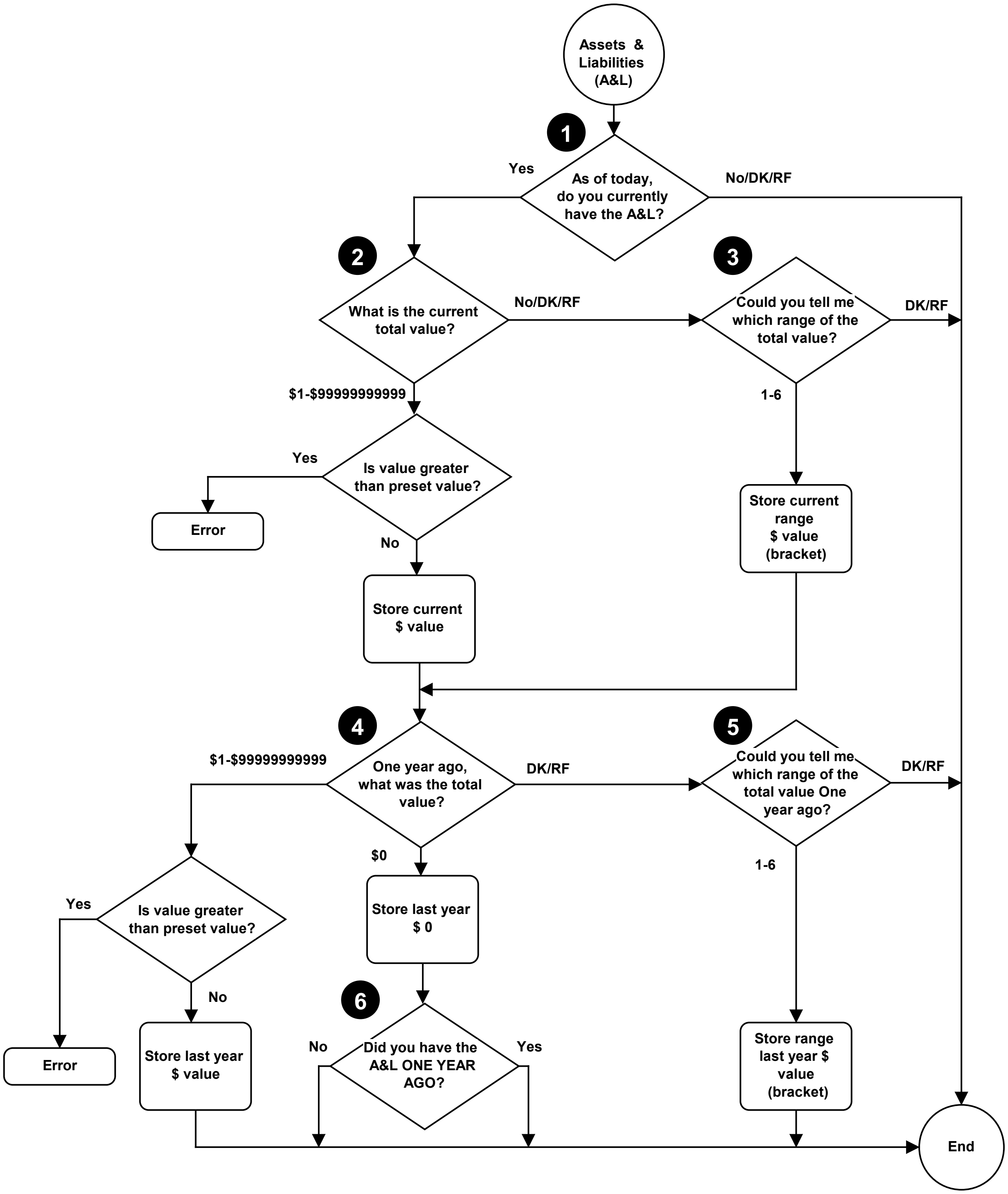

Assets & Liabilities (A&L)
1
As of today, do you currently have the A&L?
Yes
No/DK/RF
2
What is the current total value?
No/DK/RF
3
Could you tell me which range of the total value?
DK/RF
$1-$99999999999
1-6
Is value greater than preset value?
Yes
Error
No
Store current $ value
Store current range $ value (bracket)
4
One year ago, what was the total value?
$1-$99999999999
DK/RF
5
Could you tell me which range of the total value One year ago?
DK/RF
$0
1-6
Is value greater than preset value?
Yes
Error
No
Store last year $ value
Store last year $ 0
6
Did you have the A&L ONE YEAR AGO?
No
Yes
Store range last year $ value (bracket)
End

Figure 1-1. Collection of asset/liability holding data as the time of interview and as the previous year in the Consumer Expenditure Surveys.

Depending on the responses to the questions in the protocol, different missing values are generated. For example, if respondents do not confirm holding an A&L account currently in (2) and do not provide a range (bracket) for the A&L account value in (3), missing values are generated for the current and previous year's A&L account values. The missing pattern for the last year's A&L account values is called monotonic since not responding to having the account or providing the range in the current A&L value generates the missing values for the previous year's account indicator and account values automatically. In this case, the missing previous' year variables are nested within the current year variables.

1.2.2 Issues in the data collection protocol of the A&L variables

Two important issues arise from the data collection protocol for the A&L variables: The first issue is how the nesting of the missing values of the previous year variables impacts the population of inference. This problem is illustrated in Figure 1-2, which shows the theoretical status of the accounts at the time of the interview and the previous year. Theoretical means that these statuses could exist, but they are not all captured in the interview protocol.

	Did the CU hold the A&L account in the previous year?	
Does the CU hold the A&L account now?	Yes	No
Yes	D_{11}	D_{10}
No	D_{01}	D_{00}

Figure 1-2. Asset/liability holdings at the time of the interview and in the previous years.

In addition to the groups of CUs, with the same status during the current and previous year (e.g., D_{11} and D_{00} , CUs hold or do not hold the A&L account at the two points in time), the figure shows those with changing status, D_{10} and D_{01}. These refer to CUs who acquired a new account or lost it within the previous year (e.g., CU held the account at the time of the interview but did not in the previous year, and CUs did not hold the account at the time of the interview but held it in the previous year). Data related to the previous year's A&L account variables are not complete because data from group D_{01}, CUs who lost the account within the previous year are not being collected.

The data collection protocol in Figure 1-1 shows data from group D_{01} are not collected, and this introduces an undercoverage for estimates related to the previous year's A&L account variables. Furthermore, no imputation method can be used to replace the missing data in D_{01} because there are no CUs that are ever asked about this in the survey. For example, any total estimates for the previous year's A&L accounts are always less or equal to the corresponding estimate for CUs who had the account at the time of the interview. Table 1-4 illustrates this situation for IRA. Row 11 of the table shows CUs without any assets related to retirement accounts, IRAs, or Thrift Savings Plans at the interview time[1]. In all these cases, it is unknown whether the respondent had the account in the previous years because this question is not asked.

Table 1-4. Distribution of variables for assets related to retirement accounts such as 401(k), IRA, or Thrift Savings Plan (Variables IRA, IRAX, IRAB, IRAYRX, IRAYRB, and IRAYR).

Row	IRA	IRAX	IRAB	IRAYRX	IRAYRB	IRAYR	Count
1	Blank	Blank	Blank	Blank	Blank	Blank	1,000
2	1=Yes	Missing	Blank	Missing	Blank	Blank	1,300
3	1=Yes	Missing	Range	Missing	Blank	Blank	100
4	1=Yes	Missing	Range	$0	Blank	1=Yes	N<15
5	1=Yes	Missing	Range	$0	Blank	2=No	40
6	1=Yes	>$0	Blank	Missing	Blank	Blank	80
7	1=Yes	>$0	Blank	Missing	Range	1=Yes	400
8	1=Yes	>$0	Blank	$0	Blank	1=Yes	N<15
9	1=Yes	>$0	Blank	$0	Blank	2=No	100
10	1=Yes	>$0	Blank	>$0	Blank	Blank	3,400
11	2=No	Blank	Blank	Blank	Blank	Blank	8,100

The second issue is the presence of A&L accounts with zero balances that we call "zero value" accounts. Accounts with zero balances are sensible for A&L accounts such as credit (e.g., CU's have a credit card with a balance of $0). For others, the zero accounts are not easy to conceptualize; for example, a life insurance account with a zero value might be considered the same as not having a life insurance account. If the CU acknowledges holding the account and responds with the bracket value instead of providing a value, then it is not possible to indicate a zero account because the brackets in the instrument do not include $0. An implicit assumption is that all accounts where the CU provided a bracket value are positive accounts. This assumption may be sensible for most A&L variables but not for credit cards. In this research, we modified the lower value of bracket for credit to include $0. The modification acknowledges that zero account values can be used in the imputation.

Another issue affecting the previous year's A&L variables is differentiating zero account values and nonexistent accounts. For the previous year's A&L variables, the instrument's protocol specifically

[1] Row 1 in Table 1-1 may include some CUs without a current account but has the account in the previous year.

confirms if the account exits when a value of zero is provided. However, confirming the status of the A&L is not possible if a bracket value is provided since the brackets for all A&L variables do not include $0. The same modification made to the current year's credit card bracket was implemented to the previous year's credit card accounts.

1.3 Variables to Impute in the CE Survey and Restricted Use Files

Imputation is the process that produces values that replace missing data due to nonresponse, attempts to preserve the relationships in the data, and provides tools for measuring the uncertainty about these relationships (van Buuren, 2018). This research evaluated the imputation of 24 variables for A&L holdings. These variables are the 8 indicators of flags for the CU having the account the previous year at the time of the interview shown in Table 1-2 and the 8 current and 8 last year's A&L account values listed in Table 1-3. Note that as discussed above, the 8 indicator variables for the current year account holdings are not imputed because they are deduced following the assumptions given in the next section.

For this research, the BLS provided 17 restricted-use files listed in Table 1-5[2]. The restricted-use files included data from the first quarter of 2017 to the first quarter of 2020. The table shows the number of records, sums of weights[3], and Kish's design effects (1 + the squared coefficient of variation of the weights) for all the records in the file and those who received the A&L questions identified by the variable INTERI=4.

The analysis of the missing patterns in Section 1.5 is based on all the files in Table 1-5. On the other hand, the imputation for the 2019 annual estimates computed in Chapter 2 excludes the file FMLY201[4].

[2] We also received the file FMLY1971 but due to changes of the questionnaire, this file does not have all the set of A&L variables. It was decided to exclude them from the research.

[3] The variable FINLWT21 contains the final CU nonresponse adjusted weights

[4] We mirrored the production with sequential imputations of the A&L variables by quarter and the production of annual estimates for 2019. The file FMLY201 with data for 2020 Q1 data was excluded because is extemporaneous to the 2019 process.

Table 1-5. Restricted-Use files used in this research.

Files	Year	Quarter	Reported Months	Complete file			Where INTERI=4		
				Number of records	Sum of weights	Kish's DEFF	Number of records	Sum of weights	Kish's DEFF
FMLY172	2017	2	Apr-May-Jun	6,200	130,000,000	1.10	1,600	33,250,000	1.09
FMLY173	2017	3	Jul-Aug-Sep	6,100	129,900,000	1.11	1,600	33,110,000	1.12
FMLY174	2017	4	Oct-Nov-Dec	6,000	130,000,000	1.12	1,500	31,610,000	1.12
FMLY181	2018	1	Jan-Feb-Mar	5,900	130,600,000	1.11	1,500	31,770,000	1.11
FMLY182	2018	2	Apr-May-Jun	5,900	131,100,000	1.10	1,500	32,700,000	1.10
FMLY183	2018	3	Jul-Aug-Sep	5,800	131,400,000	1.11	1,500	32,970,000	1.09
FMLY184	2018	4	Oct-Nov-Dec	5,600	131,600,000	1.12	1,400	32,660,000	1.11
FMLY191	2019	1	Jan-Feb-Mar	5,600	131,800,000	1.12	1,400	33,820,000	1.13
FMLY192	2019	2	Apr-May-Jun	5,500	131,700,000	1.13	1,400	33,160,000	1.12
FML Y193	2019	3	Jul-Aug-Sep	5,300	132,100,000	1.12	1,300	31,310,000	1.11
FMLY194	2019	4	Oct-Nov-Dec	5,200	132,500,000	1.14	1,300	32,820,000	1.15
FMLY201	2020	2	Jan-Feb-Mar	5,200	131,900,000	1.13	1,400	34,610,000	1.15

The type of response in the restricted use files are indicated by SAS special missing values listed in Table 1-6. Special care is needed for some cases that have a SAS missing value (e.g., .A for valid blanks for items not asked or skipped), so these cases are not counted as missing values in the missing pattern analysis in a later section.

Table 1-6. Values for the associated variables for type of response in public use microdata of the Consumer Expenditure Surveys*.

Special `values	Description
.A	Valid Blanks
.B	Valid Blanks
.C	Illegal Nonresponse
.D	Illegal Nonresponse
.E	Other Nonresponse
.F	Don't Know/Refusal
.G	Illegal Response (BLS-derived Family Characteristic Illegal Response (BLS-derived Family Characteristic

*The table includes only the special missing values found in the A&L variables

1.4 File Processing and Data Cleaning

Before imputing the missing values, the data from the restricted used files in Table 1-5 were processed and cleaned to resolve data inconsistencies. The restricted-use data files were created using an older version of SAS with a restriction on the length of the variable names to 8 characters. As a result, special rules were developed to name the A&L variables when the variable's name would have more than 8 characters. As part of the files' pre-processing, the A&L variable names were standardized using a common root with different suffixes, as indicated in Table 1-7. The standardization simplified the development of the programs as the names of the related variables can be created by adding the required suffix to the common root. This procedure enabled us to focus more resources on the analysis of the methods instead of the implementation of the code. As an example, Table 1-8 shows the standardization of the variables related to the student loan account.

Table 1-7. Roots of A&L variables.

Root	Description
CREDIT	Any credit cards
IRA	IRA, retirement accounts
LIQUID	Checking accounts, savings accounts, money market, CDs
OTHAST	Other financial assets
OTHLON	Other debt
STOCK	Stocks, bonds, or mutual funds
STUDNT	Any student loans
WHOLIF	Whole life insurance

Table 1-8. Standardization of A&L variable names.

Root	Suffix	Variable	
STUDNT	none	STUDNT	Flag for the CU is currently holding a student loan at the time of the interview
	YR	STUDNTYR	Flag for the CU holds a student loan at the time of the interview last year
	X	STUDNTX	The student loan account value at the time of the interview
	YRX	STUDNTYRX	Last year's student loan account value at the time of the interview
	B	STUDNTB	Bracket containing the student loan account value at the time of the interview
	YRB	STUDNTYXB	Bracket containing last year's student loan account value at the time of the interview

In the pre-processing of the files, we also perform logical imputations that are not part of the imputation process described earlier. These logical imputations were the following:

- For known current account values, we filled out the bracket indicator

- For cases with the CU's indicator for currently having or not the A&L account flagged as valid blanks, we filled out the corresponding account value and bracket indicator as valid blanks

- For cases with the CU's indicator of currently having or not the A&L account coded as 2 (e.g., the CU does not hold the account), we filled out the account value and bracket with valid blanks

- The current account indicator of the CUs who refused or answered don't know was recoded as 2 (CU does not have the account). This assignment assumes that those CUs with a known current account indicator do not have the account. This decision was discussed with BLS, and it was agreed that the missing values result from CUs not being familiar with the type of A&L account. A consequence is that the 8 current A&L indicators in Table 1-2 are assumed to be fully reported. Therefore, the imputation methods evaluated in this research do not pertain to these indicator variables.

- The bracket variables were filled out using the account value for all CUs with valid account values. This type of missing value occurs when the respondent reports the account value, and the question for the bracket is skipped, as shown in Figure 1-1.

- The cases with all valid blanks, that is, CUs where no questions related to the A&L were asked, were considered as valid but not subject to imputation. These arise when the CU is not eligible for such questions (i.e., CUs consisting of emancipated minors). In other words, these cases are not imputed nor can be used in any imputation model fitting.

Although the instrument's protocol collects the previous year's A&L accounts with a minimum value of $1, there are some instances on the data files where the account value was $0. In this situation, the previous year's account indicator is used to determine if the CU held such an account, differentiating accounts with no balance from those without an account. This classification is not only important for the production of estimates (e.g., the total number of accounts vs. total value of the A&L holding) but also for the imputation process since we need to determine if the zero value accounts can be used to fit the imputation model.

As an example, Table 1 9 shows the distribution of the zero value accounts for IRAYR (assets related to retirement accounts, IRAs, or Thrift Savings Plans) at the time of the interview and last year.

Table 1-9. Distribution of account value of assets related to retirement accounts such as 401(k), IRA, or Thrift Savings Plan by the time of the interview and last year's account.*

		Last year's account value (reported value or from range)			
		$0		>$0	Missing
Respondent currently has an IRA account	Current account value (reported or from range)	IRAYR		IRAYR	IRAYR
		Yes	No	Blank	Blank
Yes	>$0	N<15	80	2,300	90
Yes	Missing			N<15	550

*** Counts in the table include those cases where holding the IRA =Yes**

- All the previous year's indicators were logically imputed based on the previous year's account values or brackets. An example of this assignment is shown in Table 1-10 for IRAYR. The table shows the cross-tabulation of the variables IRAYRX (What was the total value of all retirement accounts ONE YEAR AGO TODAY?), IRAYRB (could you tell me which range on CARD D best reflects the total value of all retirement accounts ONE YEAR AGO TODAY?), and IRAYR (Did (you/you or any members of your household) have any retirement accounts such as 401(k)s, IRAs, or Thrift Savings Plans ONE YEAR AGO TODAY?) The table shows that the values of IRAYR do not reflect the information of the cases with either a positive value of IRAYRX or a valid range IRAYRB. In this example, the value of IRAYR of these cases (highlighted in the table) was set to "1:Yes."

Table 1-10. Cross-tabulation of variables IRAYRX, IRAYRB, and IRAYR.

		IRAYR		
IRAYRX	**IRAYRB**	**1 = Yes**	**2 = No**	**Missing**
$0	Missing	N<15	80	
>$0	Missing			1,600
>$0	Valid Range			N<15
Missing	Missing			4,400
Missing	Valid Range			700

1.5 Missing Pattern Analysis

Table 1-11 shows the number of valid cases, missing values, unweighted mean, standard deviation, and sums of the current and previous year's A&L account values after the data cleaning and logical imputation assignments described above.

Table 1-11. Statistics after preprocessing the files.

VARIABLE	N	NMISS	MEAN	STDDEV	SUM	MIN	MAX
CREDITX	8,900	9,900	4,300	8,700	38,000,000	$0	250,000
CREDITYRX	5,800	13,000	6,000	11,000	35,000,000	$0	250,000
IRAX	5,500	13,000	230,000	510,000	1,300,000,000	$0	15,000,000
IRAYRX	5,100	14,000	200,000	450,000	1,000,000,000	$0	12,000,000
LIQUIDX	9,600	9,200	28,000	150,000	270,000,000	$0	12,000,000
LIQUIDYRX	9,100	9,700	26,000	130,000	240,000,000	$0	10,000,000
OTHASTX	300	18,000	280,000	920,000	96,000,000	$30	10,000,000
OTHASTYRX	300	18,000	270,000	850,000	87,000,000	$0	9,000,000
OTHLONX	900	18,000	16,000	65,000	15,000,000	D*	1,000,000
OTHLONYRX	900	18,000	12,000	52,000	11,000,000	$0	1,000,000
STOCKX	1,400	17,000	230,000	900,000	320,000,000	D*	23,000,000
STOCKYRX	1,300	17,000	220,000	860,000	290,000,000	$0	22,000,000
STUDNTX	2,300	16,000	40,000	53,000	94,000,000	$20	600,000
STUDNTYRX	2,200	17,000	39,000	52,000	87,000,000	$0	550,000
WHOLIFX	1,100	18,000	170,000	3,000,000	200,000,000	$0	100,000,000
WHOLIFYRX	1,100	18,000	150,000	3,000,000	160,000,000	$0	100,000,000

*Not Disclosed due to Privacy concerns

One of the important characteristics of the A&L data is the high correlation between the A&L account value at the time of the interview and the previous year's value. Figure 1-3 shows the scatter plots based on the nonmissing cases of the A&L account value variables. The plots with red borders show the scatter plots of the same account at the two points in time (e.g., at the time of the interview and the previous year).

In addition to the high correlation between the account values at the two points of time, there is also a high degree of association between having the account at the two time periods (i.e., the likelihood of having the A&L account last year given the hold the account at the time of the interview). An example of the high association between the A&L account indicators is shown in Table 1-12. The table shows that most respondents who reported holding the IRA account at the interview also reported having the account the previous year.

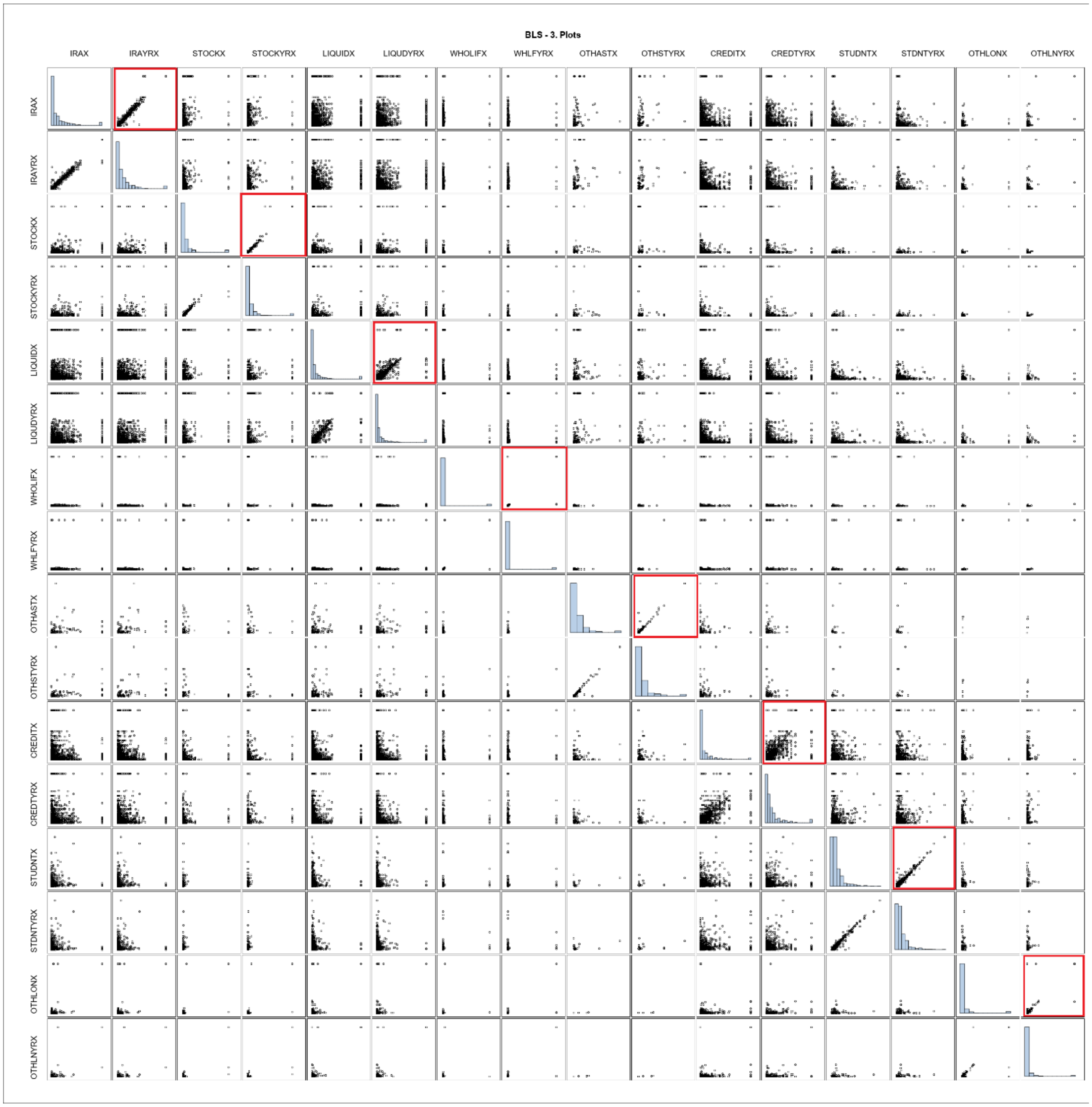


Figure 1-3. Scatter plots of the A&L variables in the Consumer Expenditure Surveys.

Table 1-12. Distribution of type of response for the account value of assets related to retirement accounts such as 401(k), IRA, or Thrift Savings Plan.*

		Respondent had an IRA account last year		
Respondent currently has an IRA account	Current account value	Yes	No	Missing
Yes	Valid value or range	2,300	80	90
	Missing value or range	N<15		600
No	Not applicable			3,400

* Counts in the table include those cases where holding the IRA account is known.

The plots suggest that imputation methods where the account values at the time of the interview are used to predict the last year's account value in the imputation model or the reverse, previous year's value as a predictor of the current account value, may be more likely to preserve the high correlation between the variable than those methods that reflect this correlation only though the independent variables.

We also examined the seasonality of the missing values from the second quarter of 2017 to the first quarter of 2020 by creating a variable with values of 1 if the variable was missing or 0 otherwise. We computed the autocorrelation function (ACF) to determine if there is stationarity in a time series and to identify lags with significant correlations. In the plots, each bar represents the size and direction of the correlation. The horizontal lines indicate the values where the autocorrelation function is statistically different from 0 with $\alpha = 0.5$. The autocorrelations are statistically significant if the bars extend across the horizontal line in the ACF plots. None of the autocorrelations in the plots were significant, and therefore, there is no evidence that there is a seasonal effect in the number of missing values. However, the ACF analysis is limited because it is based on relatively few quarters.

1.6 Predictor or Explanatory Variables

With the help of the BLS, we identified categorical and continuous variables as potential explanatory or predictor variables for the imputation models. All these variables are defined at the CU level (A&L data are not defined at the person level). Most of these variables are used in the current imputation method for Income in the CE. These variables are critical in the next section, where we discussed imputation models that rely on having good predictors of the missing A&L data.

2. Imputation Methods

There are many methods for imputing missing data, but few produce statistics that can be used to produce valid statistical inferences. We propose and evaluate three imputation methods for the A&L that meet the following requirements determined by the Task Order and conversations with BLS staff that are intended to support producing valid statistical inferences:

- The imputation methods should follow the current production process in the CE. These include the release of quarterly data and the production of annual estimates based on the quarterly files. Since data are released quarterly, the imputation method for the A&L variables should be done quarterly.

  The estimates for A&L variables with multiply imputed data should use the same analysis weights available in the quarterly files. Furthermore, estimates of variances should be computed using Balanced Repeated Replication (BRR) using the same 44 replicate weights used for other analyses (Wolter, 2017). This restriction determined the types of imputation, as methodologies such as fractional imputation (Yang & Kim, 2016) require a different data configuration in the data files and the addition of records for fractional weights and imputed values.

- Similar to the income imputation method currently used for the CE, the proposed methods for the A&L variables should use the non-missing data from the previous quarters. Specifically, the income procedure uses data from the previous 20 quarters, including the imputed quarter, to impute the current quarter's missing values.

- One of the imputation methods to be evaluated should follow the general imputation process currently used to impute income in the CE.

All the methods we propose and evaluate use multiple imputed values (Rubin, 1987, 2014, and 2004) to reflect the uncertainty due to the missing values as currently done for income in the CE. As in the income imputation, the methods generate m=5 imputed values for each missing value, which are used to compute the estimates and variance estimates.

2.1 Definitions and Concepts

We first introduce definitions and concepts used in the methodology before describing the properties of the methods proposed.

- Model donor pool: Set of cases with non-missing values of the variable being imputed used for fitting a statistical model. The model donor pool consists of complete cases

$\left(Y_i, \mathbf{x}_i\right)$ used to establish the relationship between the variable being imputed, Y_i, and p auxiliary variables $\mathbf{x}_i = \left(x_0, x_1, \ldots, x_p\right)$. We note that the term donor pool is sometimes used to describe the set of values that can be used in a hot-deck imputation to replace the missing value. We use it more generally to describe the set of values used for modeling. In most implementations, the model donor pool consists of the non-missing cases found in the current sample. However, mainly due to the small sample size of the A&L data in a quarter, the missing donor pool is expanded to include the non-missing cases from the last 20 quarters. The goal of expanding the model pool is to provide enough cases to produce robust models. On the other hand, since the data from the additional quarters include data collected at different times, this process may dampen any temporal effects compared to those from donor pools based on a single quarter[5]. However, for the A&L variables in this research, only data beginning with the first quarter of 2017 are available (e.g., a total of 8 quarters when imputing the A&L variables for 2019 Q1). The reduced number of available cases in the donor pool may reduce the variability of the imputed value from estimating the parameters of the models as the number of quarters increases (see Section 1).

- Imputed model and imputed values. To simplify the description of the imputation model, we consider a univariate model. Let Y_i be the variable to be imputed with an assumed distribution, for example $N\left(\mu_i, \sigma_i^2\right)$. Let $\mathbf{x}_i = \left(x_{io}, x_{ip}, \ldots, x_{ip}\right)$ be a vector of auxiliary variables for Y_i (data from the interview describing the CU; for example, sociodemographic, geographic, and economic variables. Assume a linear model $Y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{\beta}^{\mathrm{t}} = \left(\beta_{io}, \beta_{ip}, \ldots, \beta_{ip}\right)$ and t is the transpose operator, then we can produce the prediction $\hat{\mu}_i = \overset{\square}{E}\left(Y_i\right) = \mathbf{x}_i \hat{\boldsymbol{\beta}}$. These predicted values $\hat{\mu}_i$ do not reflect the uncertainty from two sources: the error: 1) ε_i, the error from the assumed model distribution and 2) the uncertainty from estimating the model's parameters (i.e., the estimates of regression coefficients $\hat{\boldsymbol{\beta}}$). The value to impute is created by adjusting these predictions as

$$\hat{\mu}_i^* = \mathbf{x}_i \left(\hat{\boldsymbol{\beta}} + \mathbf{b}^*\right) + e_i^*,$$

where $\mathbf{b}^*$ is a random noise added to the parameters $\hat{\boldsymbol{\beta}}$ drawn from the multivariate normal distribution $MN\left(\mathbf{0}, \hat{V}\left(\hat{\boldsymbol{\beta}}\right)\right)$ that reflects the uncertainty of estimating the number of parameters and the parameter values of the model, and the random noise e_i^* drawn from the distribution $N\left(0, \overset{\square}{V}\left(\hat{Y}_i\right)\right)$. The *m* multiple imputed values are then created using *m* independent draws of $\mathbf{b}^*$ and e_i^*. We also refer to $\hat{\mu}_i^*$ as the shocked

[5] To reduce the dampening effect, a variable for the quarter is used as a predictor in the models.

predictor because its value depends on the random numbers of $\mathbf{b}^*$ and e_i^*. We called this *parametric imputation* because the parameters of an assumed model are estimated.

The above simple example assumes a linear relationship between Y_i and $\mathbf{x}_i$, but may not be appropriate for many variables, especially binary variables such as the A&L indicators. In this case, a different model is used, for example, generalized linear models (GLM) with an appropriate link function (i.e., logit).

The parametric imputation methods are designed to work under the assumption that the relations within the missing parts are similar to those in the observed data. More technically, there is an assumption of ignorability conditional on the auxiliary variables (this assumption is also referred to as MAR, or missing at random). Note this assumption is not related to the assumptions about the missing observations that are taken into account in nonresponse weighting adjustments. As a result of these assumptions, the model that generates the variable Y_i can be identified using the observed data $\mathbf{x}_i$.

The use of well-defined models clarifies the tasks for imputing any variable (Rubin 1987). First, we have the modeling task that determines a specific model for the data. Then the estimation task estimates the parameters of the assumed distribution given by the model. The last task, the imputation task, generates the values that replace the missing data by drawing successively from the parameter and data distributions.

All the proposed imputation methods use well-defined, and in most cases parametric, models. The details of the type of model for the A&L indicators and account variables are described in the following sections.

- Blocking. We refer to blocking as the procedure for identifying the cases for the donor pool and the cases to impute. Blocking is mainly a data reduction tool that keeps only the eligible cases for the imputation analysis of a given A&L variable. Blocking relies on two characteristics of the A&L data – not all sampled cases report having the same A&L accounts, and the missing A&L values are nested within those CUs with a current account[6].

  As an example, consider blocking for the variables related to IRA accounts; that is, IRA, IRAX, IRAYR, IRAYRX). First, the initial file is consolidated by the cases for both the quarter being imputed and the previous quarter. The variables to impute are IRAX and IRAYR. The first block is created with all the cases where IRA=1, and within this group, the donor pools for both IRAX and IRAYR are created using the cases where IRAX or IRAYR are not missing, respectively. The beggars, or cases to impute, are those where IRAX and IRAYR are missing in the current quarter (i.e., the quarter being imputed). After imputing for IRAYR, a new block is created using the cases where

[6] We assume that there are no missing values for current account indicators (variable IRA). See Section 1.3.

IRAYR=1 for the imputation of IRAYRX. The model donor pool is defined by these cases where IRAYRX is not missing.

Blocking is most useful when the same type of A&L accounts are imputed because smaller files are created and processed as in the first imputation method evaluated in this research. In this case, the final file with all sampled cases in the quarter being imputed is recreated by merging the imputed values and non-missing cases from the different blocks. On the other hand, the other two imputation methods proposed here require maintaining a complete quarter file during the imputation process. In this case, blocking is more complex and involves creating special flags to identify the model donor pool and cases to impute in these blocks.

2.2 Imputation Methods

We implemented the three imputation methods listed in Table 2-1. The table shows the name of the method, the models used for the A&L variables, and the software used for the implementation. The table shows that the methods use a mixture of parametric imputation models that depend on the type of A&L variable. For example, all A&L indicators are imputed with models that produce either 0 or 1 as the imputed value (e.g., the CU holds the account or not). In contrast, for imputing the A&L values, the imputation models depend on the availability of the bracket with the range of the account value. When the bracket is known, all methods use the same model imputation called mean bracket imputation (we sometimes refer to this as mean imputation as a short-hand). This model ensures that the imputed values are contained within the lower and upper value of the bracket. A detailed description of the mean imputation is given in the next section. One important difference is that Method 1 does not multiply impute for the account indicators as discussed below.

When the bracket is missing, then the methods use different imputation models to estimate the distribution of Y_i. Method 1 is based on a univariate model but transformed values of Y_i are modeled using linear regression. In contrast, Method 2 and 3 are based on a Fully Conditional Specified (FCS) model (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006; van Buuren, 2007), which can be used to impute multivariate missing data on a variable-by-variable basis. In other words, the imputation models in Method 2 and 3 take advantage of the correlation among all variables to be imputed $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{iq}\right)$ in addition to the auxiliary variables $\mathbf{x}_i$.

Instead of specifying a joint multivariable distribution for all variables, the FCS approach specifies the conditional distributions for each variable to impute Y_{ir} conditional on the remaining variables

$\left(\mathbf{Y}_{i,\bar{r}},\mathbf{x}_i\right)=\left(Y_{i1},Y_{i2},\ldots,Y_{ij\neq r},\ldots,Y_{iq},x_{i0},x_{i2},\ldots,x_{ip}\right)$ from which the draws are made. In this approach, there is no need to specify the full multivariate model for the data.

There are several ways to implement the imputation of missing values under FCS, including the Multiple Imputation Chained Equations (MICE) algorithm (van Buuren & Groothuis-Oudshoorn, 2000; van Buuren & Groothuis-Oudshoorn, 2011). The MICE algorithm takes a random draw from the observed data, and then it imputes the incomplete data variable-by-variable through repeated iterations. One iteration consists of one cycle through all Y_j. The number of iterations is generally low (between 5 and 10). Repeated imputed values are generated by executing the MICE algorithm as many times as the number of repeated imputed values (*m*). This process can be time-consuming, and there are computing approaches to reduce the time required to reach a solution. More details on the implementation of the MICE algorithm are found in Section 2.5.

Another difference mentioned above is the additional step in Method 2 and 3 for imputing zero accounts (the CUs hold an account, but it has a zero balance). This additional imputation is not implemented in Method 1 because we wanted to mirror the process for the income imputation that does not include this step.

Although Methods 2 and 3 both use FCS and the MICE algorithm, the difference is in the imputation model. Method 2 is based on a linear regression of the transformed data that is similar to Method 1 in this regard, while Method 3 uses MICE with models based on fits of random forests using untransformed data. Details of these methods are found in Sections 2.5 and 2.6.

Table 2-1. Details of evaluated methods.

Method	Description	Type of A&L variable	Model	Software
1	Linear regression similar to that used in income imputation	Account Indicators	Probit (i.e., logistic regression), single imputed value, no multiple imputation	SAS
		Account values	Linear regression of transformed account value	SAS
			Mean bracket	SAS
2	MICE with linear regression	Account Indicators	Probit (i.e., logistic regression) with multiple values	R, modified package MICE
		Zero value account indicator	Probit (i.e., logistic regression) with multiple values	R, modified package MICE
		Account values	Linear regression of transformed account value (best transformation)	R, modified package MICE
			Mean bracket	R, modified package MICE
3	MICE with random forests regression	Account Indicators	Random forest for binary variables with multiple values	R, modified package MICE
		Zero value account indicator	Random forest for binary variables with multiple values	R, modified package MICE
		Account values	Random forest for continuous variables	R, modified package MICE
			Mean bracket	R, modified package MICE

2.3 Estimates of Multiple Imputed A&L Variables for 2019

The software programs for the three methods were written for simulating the imputation for the 2019 estimates. In this chapter we describe an illustration of the methods for imputing the annual 2019 values. We begin by pretending that the 2019 Q1 is available (as would be the case in production with the first quarter being available before the others) and the imputation is carried out for all A&L variables in this file. The expanded model donor pool consists of all the data from 2017 Q2 to 2019 Q1 files. Although it is not done in the production process, we produce the estimates for the A&L variables for this quarter. The sampling weights from the quarterly estimates were not adjusted, so the totals are missing a factor of 4 since the A&L data were collected in approximately one-fourth of the sample.

A similar process was repeated for the estimates for 2019 Q2 to Q4. After the last quarter was imputed, a file containing all the 2019 quarters was used to produce the annual estimates. No weighting adjustment was required since the combined file represents the total CU population in the US.

Unlike other variables in the CE that refer to account values in the previous three months from the time of the interview, all A&L variables referred to account status and value on the day of the interview and one year ago from the day of the interview. As a result, there is no need to prorate or identify the cases for a specific quarter for the 2019 annual estimate as it is done with income imputation.

The estimates were produced using SUDAAN Release 11.0.3 (RTI, 20XX) with the option for estimates with repeated values. We use the procedure PROC DESCRIPT for totals, means, and proportions for account values and indicators after recoding the latter into 0 and 1 values. For estimates that include more than one set of multiple imputed variables (e.g., IRAYR1 to IRAYR5 and IRAYRX1 to IRAYRX5), the file needs to be restructured from a horizontal layout with the five repeated values into five separate files, each with a single imputed value. The estimates were computed using the option BRR with the full sample weight FINLWT21 and the 44 replicate weights WTREP01-WTREP44. Missing values of the replicate weights were set to 0 as SUDAAN excludes weights without a numeric value.

2.4 Method 1: Income Methodology

The first method for imputing the missing values of the A&L accounts, Method 1, follows the current methodology closely to impute the components of income as described in the document “I_IncomeImputation.docx” provided by BLS. Method 1 is based on a univariate missing model (van Buuren, 2018) where the A&L variables (indicators and account values) are separately imputed, assuming a univariate model for each variable. Although the methods are based on a univariate model, the variables related to the same type of account (e.g., IRA, IRAX, IRAYR, and IRAX) are imputed sequentially.

Method 1 uses a linear model for the A&L account values when the bracket is unknown. Because of the skewness of the data, Model 1 does not assume a model for the Y_i variables in the block but does so for the transformed values, where the transformation is based on the order statistics. The ordered variable Z_i is defined as the Normal Score of $Y_i + u_i$, where the term u_i is a $U\left(0,1/1000\right)$ random number to prevent tie values. The imputed Z normal score value is generated as $\hat{\mu}_{zi}^{*} = \mathbf{x}_i\left(\hat{\boldsymbol{\beta}} + \mathbf{b}^{*}\right) + e_i^{*}$ and the imputed value Y_i^{*} is the untransformed value of $\hat{\mu}_{zi}^{*}$ [7].

The linear model for $\hat{\mu}_{zi}^{*}$ is identified by backward stepwise regression starting with the full set of auxiliary variables $\mathbf{x}_i$. It is desirable to use an automatic procedure for variable selection with minimum intervention. The set of auxiliary variables consist of all the dummy variables generated by each level of categorical variables in addition to the continuous variables in $\mathbf{x}_i$ (see Section 1.6). The fitted model after the backward regression is evaluated to determine if there is overspecification according to pre-determined criteria for the bias and variance of $\hat{\mu}_{zi}^{*}$. These methods are nearly the same as used in income imputation.

The model for imputing account values when the bracket value is missing is very different because the bracket contains so much information about the missing value. The imputed value is generated by

[7] Special adjustments are dome when the value of $\hat{\mu}_{zi}^{*}$ are outside the range of the Z_i in the donor pool.

$$\hat{\mu}_{ib}^{*}=\begin{cases} B_{lb}+e_{2i}\left(\overline{y}_b-B_{lb}\right) & \text{if } \dfrac{B_{ub}-\overline{y}_b}{B_{ub}-B_{lb}}<e_{1i} \\ \overline{y}_b+e_{2i}\left(B_{ub}-\overline{y}_b\right) & \text{Otherwise} \end{cases}$$

where $\overline{y}_b$ is the mean of the account values of the donor pool in the bracket b , B_{lb} and B_{lb} are the upper and lower bounds of bracket b ; and e_{1i} and e_{2i} are two independent random numbers from uniform distribution, $U\left(0,1\right)$, respectively. If the bracket only contains the lower bound, then the imputed value is generated as

$$\hat{\mu}_{ib}^{*}=\begin{cases} B_{lb}+e_{2i}\left(\overline{y}_b-B_{lb}\right) & \text{if } \dfrac{B_{up}^{*}-\overline{y}_b}{B_{ub}-B_{lb}}<e_{1i} \\ \overline{y}_b+e_{2i}\left(B_{up}^{*}-\overline{y}_b\right) & \text{Otherwise} \end{cases},$$

where B_{up}^{*} is the data-based upper range computed as $B_{up}^{*}=\overline{y}_b+z_{0.95}\sigma_{y_b}$, where σ_{y_b} is the standard deviation of the account values in the model donor pool in bracket b . The mean bracket imputation is a prediction plus random noise $\hat{\mu}_{zi}^{*}=\hat{\mu}_{ib}+e_i$, without any variation from the parameters from a model because none is estimated. As a result, the mean imputed values have lower variability compared to the regression imputation.

Depending on the number of donors in the bracket, the mean imputation method relied on the auxiliary variable CFAM_TYPE2 in Table 2-2 to identify imputation subgroups within the bracket. The same rules are used except that the mean and standard deviation are computed in the group defined by the bracket and auxiliary variable. The variable CFAM_TYPE2 was created under the guidance of BLS using the indicators for the type of CU.

Table 2-2. Auxiliary Variable CFAM_TYPE2 for mean imputation.

C_FAM_TYPE2	Definition	Frequency
1	Single Male	2,200
2	Single Female	2,600
3	Single father	200
4	Single Mother	600
5	Other	10,500

An issue arises when subgroups defined by C_FAM_TYPE2 and the bracket (or by the bracket) have one or no model donors. In this case, the value is imputed using linear regression of the normal scored value in the newly defined block.

Figure 2-1 shows the imputation for Method 1 for one type of account, IRA. The same process is repeated for A&L account types. We describe the steps for imputing all IRA account-related variables. The same steps are done for all the A&L variables.

A. Creation of the block for IRA=1. By assumption, there is no imputation for the current year indicator IRA, so we start by creating the set with the eligible cases for imputing the current IRA account value.

Normal Score transformation (NST). The normal scores $Z_{j(i)}$ for $i \in \{1,...,n\}$ (or normal Z -score values) are created for the n non-missing observations of the IRAX as $Z_{j(i)} = \Phi\left(\frac{R_i - 0.5}{n}\right)$ where R_i is the rank of $y_{j(i)}$, defined as the (i) order statistic of $\{y_1,...,y_n\}$ and $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal distribution defined as $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-t^2/2\right) dt.$ The NST creates a set of transformed values $\left\{Z_{j(1)},...,Z_{j(n)}\right\}$ that would be expected had the original set of data values arisen from a normal distribution. The goal of the NST is to minimize heteroscedasticity. The median value of the A&L variable in the regression is transformed to zero, and the other values in the distribution are assigned values corresponding to the Z-score distribution.

We separate the cases into two groups depending on the missing status of the current year bracket (IRAB). Steps B to G apply to the cases where IRAB is missing. Step H applies to the remaining cases (i.e., non-missing IRAB values).

B. If the number of cases in the model donor pool for fitting the model is fewer than 30 CUs, then we fit the mean model $Z_i = \beta_0 + \varepsilon_i$ that includes only the parameter for the intercept, β_0.

C. If the number of cases in the model donor pool is 30 or more, the method posits an initial model $Z_i = \mathbf{x}_\mathbf{i}\boldsymbol{\beta} + \varepsilon_i$ where the parameter $\boldsymbol{\beta} = \left(\beta_0,...,\beta_p\right)^{\mathrm{t}}$ are the regression coefficients of the auxiliary variables $\mathbf{x}_i = \left(1,...,x_{ip}\right)$ described in Section 1.6. These include dummy variables for the categorical variables in $\mathbf{x}_i$.

D. The model is fitted using Ordinary Least Squares (OLS) regression, and the variables in the model are identified with backward elimination. This selection starts with the full model, and the predictors with the smallest contribution to the model are deleted one at a time until a stopping condition is satisfied. The predictors remaining in the model are significant at a pre-specified stay significance level (SSL). The SSL for retaining the variable x_j in the model is the *p*-value of the estimate of the associated regression coefficient $\hat{\beta}_p$ set to $p \le 0.15.$ The high SSL retains as many variables as possible to preserve statistical relationships without imputing extreme values. After dropping a variable, the model is refitted, and the process is repeated until all the associated regression confident in the model have $p \le 0.15.$

E. For the pool of beggars (cases where missing IRAB and missing IRAX in the quarter being imputed), the imputed values are generated in the following steps, which are repeated 5 times to produce 5 imputed values

- First, the parameters of the final model identified in D, $\hat{\boldsymbol{\beta}}^*$, are shocked by adding noise as $\dot{\hat{\boldsymbol{\beta}}}^* = \hat{\boldsymbol{\beta}}^* + \mathbf{e}_1$. Then, the value of the added noise $\boldsymbol{\varepsilon}$ is drawn from the multivariate normal distribution $\mathbf{e}_1 \ \square\ MN\left(\mathbf{0}, \hat{V}\left(\hat{\boldsymbol{\beta}}^*\right)\right)$ where $\mathbf{e}_1$ is a vector computed as $\mathbf{e}_1 = \sqrt{\hat{V}\left(\hat{\boldsymbol{\beta}}^*\right)}\mathbf{Z} = \sqrt{\dfrac{\hat{\sigma}^2 (n-p)}{g}}\ \mathsf{L}\left(\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\right)\mathbf{Z}$ where $\mathbf{Z}$ is a random vector drawn from the standard multivariate normal distribution, $\hat{\sigma}^2 = \sum\left(Y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}^*\right)^2 / (n-p)$ is the mean squared error of the model with a donor pool of size n and p variables, g is a random variable drawn from a $\chi^2_{df=n-p}$, and $\mathsf{L}\left(\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\right)$ is the Cholesky decomposition of $\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}$.

- The imputed Z-scores for the beggar j is computed as $\dot{\hat{z}}_j = x^*_j \dot{\hat{\boldsymbol{\beta}}}^* + e_{j2}$ where e_{j2} is the random noise drawn for the distribution $N\left(0, \square V\left(\hat{z}_i\right)\right)$ where $\hat{z}_j = x^*_j\hat{\boldsymbol{\beta}}^*$ is the predicted Z-score is computed as $\hat{z}_j = x^*_j\hat{\boldsymbol{\beta}}^*$.

F. The model in 4 is evaluated for overspecification by computing the mean and variance of the Z-scores, $\overline{Z}$ and $V\left(\overline{Z}_m\right)$, for all the cases in the block (including the imputed Z-scores) separately by the $m = 5$ repeated values. The model is not overspecified if

$$\left|\overline{Z}_m\right| < 0.03 \text{ and } 0.95 < V\left(\overline{Z}_m\right) < 1.05\text{, for all } m = 1 \text{ to 5.}$$

If the model is overspecified, then 10% of the variables in the final model are dropped, and the steps D to F are repeated until there is no overspecification. The variables to drop are those with the smallest values of $\left|\rho_{x,Z}\right|$, the absolute value of the correlation between the auxiliary variable **x** and Z score

G. If the revised model in F is not overspecified then the inverse of the 5 imputed Z-scores are computed to produce the 5 imputed account values, and we proceed with step I

H. Mean bracket imputation. The model pool of donors is divided into imputation cells defined by the demographic characteristics of the CU described by CFAM_TYPE2 in Table 2-2 within brackets. For each beggar in a specific cell, five imputed account values are produced using the mean and standard deviation (if applicable) of the model donor cases in the cell as described above when there are two or more donors in the cell. If this is not the case, the cell is redefined as the bracket ignoring CFAM_TYPE2.

I. After imputing the current account value IRAX, the indicator for the previous account IRAYR is imputed using the same block created in A. Then, a logistic regression model with a dependent variable IRAYR recoded to 0 and 1 and all the auxiliary variables as dependent variables is posited as the initial model $\operatorname{logit}\left(p_i\right)=\mathbf{x}_i\boldsymbol{\theta}$ for the probability p of having the account in the previous year. Finally, the variables in the final model are identified with backward elimination as in D.

J. The final model for the previous year indicator IRAYR is used to predict $\hat{p}_i$, the probability of having the account in the previous year conditional on the auxiliary variables in the final model for all beggars for IRAYR. The imputed value is the random realization of the Bernoulli trial $BE\left(\hat{p}_i\right)$. Unlike the linear imputation for the current account value in steps C and D, the parameters or regression coefficients $\boldsymbol{\theta}^*$ in the final model are not shocked, and only one imputed value is created. Furthermore, there are no criteria to evaluate if the model is overspecified or not.

K. After imputing for the previous year's indicator (e.g., IRAYR), a block with the cases where IRAYR=1 is created to impute for the previous year's account value (e.g.., IRAYRX). Then, similar steps from A to H are repeated using the cases in this block, replacing the current year with the previous year variables (i.e., IRA by IRAYR, IRAX by IRAYRX, and IRAB by IRAYRX).

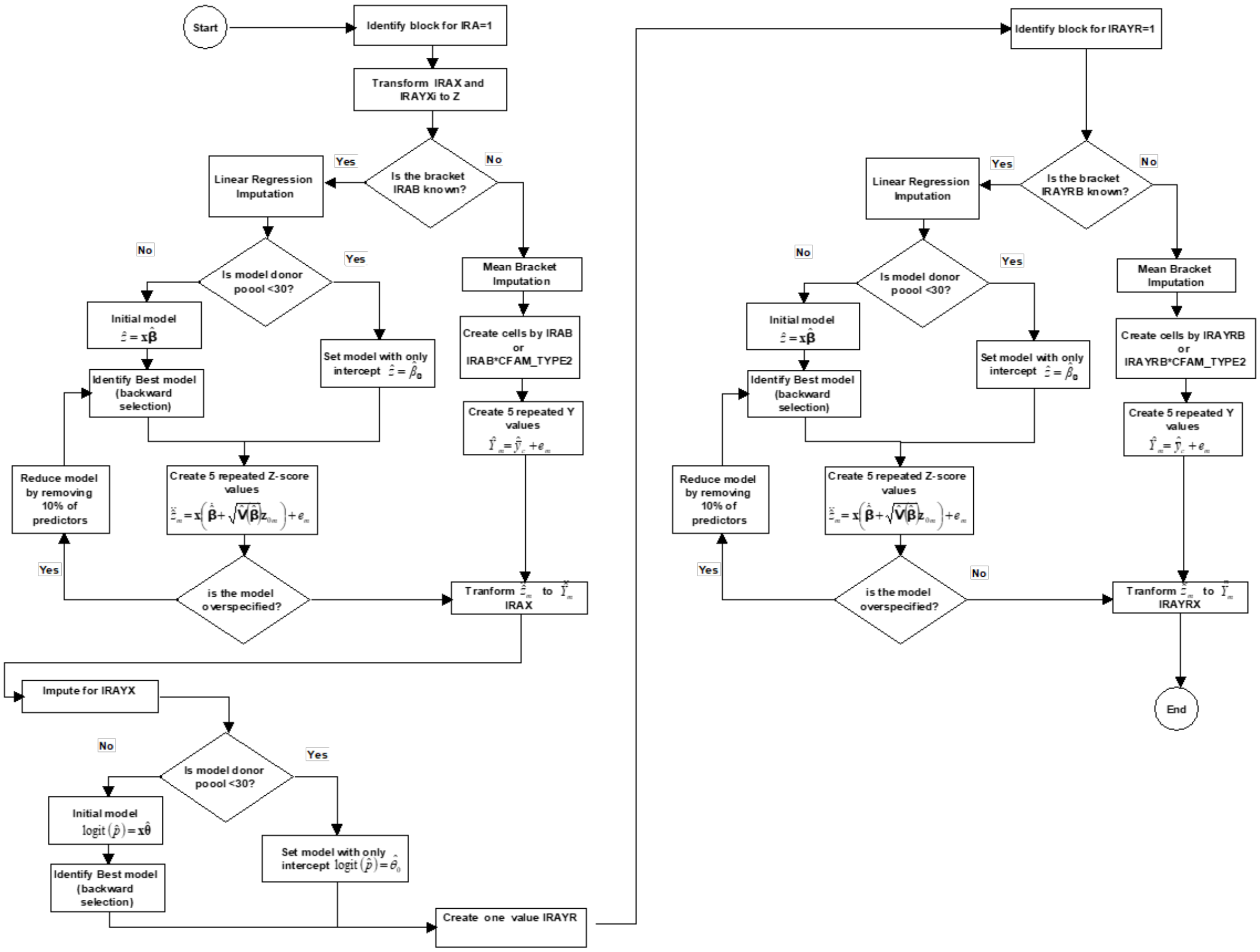

Start
Identify block for IRA=1
Transform IRAX and IRAYXi to Z
Is the bracket IRAB known?
Yes
No
Linear Regression Imputation
Is model donor poool <30?
No
Yes
Initial model $\hat{z} = \mathbf{x}\hat{\boldsymbol{\beta}}$
Set model with only intercept $\hat{z} = \hat{\beta}_0$
Identify Best model (backward selection)
Reduce model by removing 10% of predictors
Create 5 repeated Z-score values $\breve{z}_m = \mathbf{x}\left(\hat{\boldsymbol{\beta}} + \sqrt{\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})}\mathbf{z}_{0m}\right) + e_m$
is the model overspecified?
Yes
Mean Bracket Imputation
Create cells by IRAB or IRAB*CFAM_TYPE2
Create 5 repeated Y values $\hat{Y}_m = \hat{\bar{y}}_c + e_m$
Tranform $\breve{z}_m$ to $\breve{Y}_m$ IRAX
Impute for IRAYX
Is model donor poool <30?
No
Yes
Initial model $\text{logit}(\hat{p}) = \mathbf{x}\hat{\boldsymbol{\theta}}$
Set model with only intercept $\text{logit}(\hat{p}) = \hat{\theta}_0$
Identify Best model (backward selection)
Create one value IRAYR
Identify block for IRAYR=1
Is the bracket IRAYRB known?
Yes
No
Linear Regression Imputation
Is model donor poool <30?
No
Yes
Initial model $\hat{z} = \mathbf{x}\hat{\boldsymbol{\beta}}$
Set model with only intercept $\hat{z} = \hat{\beta}_0$
Identify Best model (backward selection)
Reduce model by removing 10% of predictors
Create 5 repeated Z-score values $\breve{z}_m = \mathbf{x}\left(\hat{\boldsymbol{\beta}} + \sqrt{\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})}\mathbf{z}_{0m}\right) + e_m$
is the model overspecified?
Yes
No
Mean Bracket Imputation
Create cells by IRAYRB or IRAYRB*CFAM_TYPE2
Create 5 repeated Y values $\hat{Y}_m = \hat{\bar{y}}_c + e_m$
Tranform $\breve{z}_m$ to $\breve{Y}_m$ IRAYRX
End

Figure 2-1. Method 1 Imputation process for A&L account values based on CE Income imputation.

Method 1 was implemented in SAS 9.4 (TS1M6) using the procedures PROC GMLSELECT for creating the dummy variables and backward selection, PROC IML for the estimation of $\sqrt{\hat{V}\left(\boldsymbol{\beta}^{*}\right)}$ and the computation of the 5 repeated values of Z-scores using matrix algebra. The large number of auxiliary variables in the model caused, on some occasions, numerical instability since $\sqrt{\hat{V}\left(\boldsymbol{\beta}^{*}\right)}$ could not be computed using the Cholesky decomposition as specified in the income imputation. The Cholesky decomposition or Cholesky factorization is used to generate a random vector from a multi-normal distribution with a specific variance-covariance matrix. The advantage of the Cholesky decomposition is that it produces an upper triangular matrix. With half of the elements of the matrix being zero, the matrix multiplication is simplified. For example, let suppose we want to shock the predicted Z-score $\hat{Z}_i = x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + x_3\hat{\beta}_3$ with a random vector error from a multivariate normal distribution $MN\left(0, \hat{V}\left(\boldsymbol{\beta}\right)\right)$. We can write the matrix with the standard error $\sqrt{\hat{V}\left(\boldsymbol{\beta}\right)} = \sqrt{\sigma^2}\mathsf{L}\left(\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\right)$ where $\mathsf{L}\left(\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\right)$ is the Cholesky decomposition of $\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}$. If we let $\mathsf{L}\left(\left(\mathbf{X}^{\mathrm{t}}\mathbf{X}\right)^{-1}\right) = \begin{pmatrix} c_{11} & c_{12} & c_{12} \\ 0 & c_{22} & c_{23} \\ 0 & 0 & c_{33} \end{pmatrix}$ then the expression for the shocked value $\dot{\hat{Z}}_i$ is

$$
\begin{aligned}
\dot{\hat{Z}}_i &= \mathbf{x}\left(\hat{\boldsymbol{\beta}} + \mathbf{e}_1\right) = \\
&= x_1\left(\hat{\beta}_1 + \sqrt{\sigma^2}c_{11}z_1\right) + x_2\left(\hat{\beta}_2 + \sqrt{\sigma^2}c_{22}z_2\right) + x_3\left(\hat{\beta}_3 + \sqrt{\sigma^2}c_{33}z_3\right) \\
&\quad + x_1\sqrt{\sigma^2}\left(c_{12}z_2 + c_{13}z_3\right) + x_2\sqrt{\sigma^2}c_{23}z_3
\end{aligned}
$$

The terms $\sqrt{\sigma^2}c_{11}z_{11}$, $\sqrt{\sigma^2}c_{22}z_{11}$, and $\sqrt{\sigma^2}c_{33}z_{33}$ are related to the variance of the parameters $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ while the last two terms are related to the covariance among the $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$[8]

[8] The documentation for income imputations does not include the terms for the covariance among the regression coefficients, and the expression does not correspond to $\sqrt{\hat{V}\left(\boldsymbol{\beta}\right)}$. We are not sure if this is intentional but it suggests that the income imputation assumes that the beta coefficients are independent.

Since the Cholesky decomposition could not always be computed, we used the Singular Value decomposition (SVD) to compute $\sqrt{\hat{V}(\boldsymbol{\beta})}$, as

$$\hat{V}(\boldsymbol{\beta}) = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\hat{V}(\boldsymbol{\beta})$. Since $\hat{V}(\boldsymbol{\beta})$ is a symmetric matrix, then $\mathbf{Q}$ is guaranteed to be orthogonal with $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}$, and
$\sqrt{\hat{V}(\boldsymbol{\beta})} = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^{T}$

If we define $\boldsymbol{\Lambda} = \begin{pmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{pmatrix}$ and $\mathbf{Q} = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{12} & q_{22} & q_{23} \\ q_{13} & q_{23} & q_{33} \end{pmatrix}$ then the expression for the shocked value $\dot{\hat{Z}}_i$ is much more complex:

$$\begin{aligned}
\dot{\hat{Z}}_i &= \mathbf{x}\left(\hat{\boldsymbol{\beta}} + \sqrt{\sigma^2}\mathbf{Q}\boldsymbol{\Lambda}^{1/2}\mathbf{Q}^{T}\mathbf{Z}\right) \\
&= x_1\left(\hat{\beta}_1 + \sqrt{\sigma^2}z_1\left(q_{11}^2 d_{11} + q_{12}^2 d_{22} + q_{12}^2 d\right)\right) \\
&\quad + x_2\left(\hat{\beta}_2 + \sqrt{\sigma^2}z_2\left(q_{12}^2 d_{11} + q_{22}^2 d_{22} + q_{23}^2 d_{33}\right)\right) \\
&\quad + x_3\left(\hat{\beta}_3 + \sqrt{\sigma^2}z_3\left(q_{13}^2 d_{11} + q_{23}^2 d_{22} + q_{33}^2 d_{33}\right)\right) \\
&\quad + x_1\sqrt{\sigma^2}\left\{z_2\left(q_{11}q_{12}d_{11} + q_{12}q_{22}d_{22} + q_{13}q_{23}d_{33}\right) + z_3\left(q_{11}q_{13}d_{11} + q_{12}q_{23}d_{22} + q_{13}q_{33}d_{33}\right)\right\} \\
&\quad + x_2\sqrt{\sigma^2}\left\{z_1\left(q_{11}q_{12}d_{11} + q_{12}q_{22}d_{22} + q_{13}q_{23}d_{33}\right) + z_3\left(q_{12}q_{13}d_{11} + q_{22}q_{23}d_{22} + q_{23}q_{33}d_{33}\right)\right\} \\
&\quad + x_3\sqrt{\sigma^2}\left\{z_1\left(q_{11}q_{13}d_{11} + q_{12}q_{23}d_{22} + q_{13}q_{33}d_{33}\right) + z_2\left(q_{12}q_{13}d_{11} + q_{22}q_{23}d_{22} + q_{23}q_{33}d_{33}\right)\right\}
\end{aligned}$$

Note that these results are equivalent but using the SVD is more complex.

2.5 Method 2: MICE Imputation with Linear Regression Models

As mentioned in Section 2.2, Method 2 is based on a FCS model, and the algorithm to estimate the parameters of the distribution is MICE. Mathematically, the MICE algorithm is a Markov chain Monte Carlo (MCMC) method, where the state space is the collection of all imputed values. More specifically, if the conditional distributions are compatible, the joint distribution exists and is unique (see Section 4.5.3 in van Buuren, 2018). Van Buuren gives a precise definition of compatibility of conditional distributions and shows that MICE algorithm is a Gibbs sampler. A Gibbs sampler is a Bayesian simulation technique for sampling observations from conditional distributions to obtain samples from the joint distribution (see Casella & George, 1992). While the Gibbs sampler's common applications have the full conditional distributions derived from the joint probability distributions (Gilks, 1996), the opposite is true in the MICE algorithm. In MICE, the conditional distributions set by the user are used to produce the joint distribution. However, under these conditions, there is no assurance that the joint distribution exists. As in any Gibbs sampler, the joint distribution identification requires the MCMC to converge to a stationary distribution. Since the distributions are under the user's control, they may not be consistent, and the MCMC may not converge to the joint distribution (i.e., the chain oscillates). Despite these potential problems, the method has been successfully applied in practice.

As in Method 1, the method for variable selection (i.e., model selection) and criteria for measuring model fit need to be specified. In Method 2, we depart from the backward linear regression in favor of greedy algorithms based on the Akaike information criterion or AIC (Akaike, 1981), which produces parsimonious models with a better model fit (i.e., better predictions). The AIC for variable selection evaluates the model's goodness of fit (overfitting risk) and the model's simplicity (underfitting risk). In contrast to the models in Method 1 that tend to include many more predictors, the models in Method 2 have fewer predictors.

Another difference between Methods 1 and 2 is the order of the imputations. While in Method 1, all the A&L related to the same time of account are imputed sequentially, in Method 2, all the types of A&L variables are imputed in the same run.

Method 2 also differs from Method 1 because it uses the variables to impute as predictors. Before the imputation process starts, the missing values are replaced by non-missing values. Then Method 2 follows these steps:

A. We start with the A&L indicators with the current account. In this step, the blocks for each A&L account type are created for the current year's A&L indicators. Since the blocks overlap, they are defined using separate flags by type of A&L account as the model donor pools and the set of beggars changes depending on the A&L account.

B. After creating the blocks for each current and previous year's A&L indicator, an analysis of the model donor pools is carried out to determine the best normalizing transformation of the dependent variable for the imputations that use the linear regression model. Unlike Method 1, where the normal Z-score of the account value is always used, Method 2 determines the best transformation that produces a normalized dependent variable. The transformations evaluated are the Normal Z-score, the Box-Cox transformation, the Yeo-Johnson transformation, three types of Lambert WxF transformations, and the ordered quantile normalization transformation. These transformations were available in the R-package used to normalized the data.

C. Imputation of the current account value. Since by assumption there is no imputation of the current account indicators, Method 2 begins with the next set of variables. In this case, two imputation methods are implemented depending on the status of the A&L bracket containing the account value.

D. Mean Bracket Imputation. If the bracket is not missing, the cases are imputed using the mean bracket imputation exactly as in Method 1 (see Section 2.4). Note that the mean bracket imputation does not use the transformed dependent variables.

E. Imputation of zero account values for cases with missing bracket values. The imputation of the current account value in Method 2 is done in two steps.

- Imputation of zero account values. We first model the probability of the A&L account having a zero balance for those CUs that reported having a current account. Method 2 posits an initial logit model for the probability p of a zero balance account. The best model is identified using forward regression with stopping rules based on the AIC. Once the model with the best fit is identified, the model is used to produce the shocked predictors of the beggars as $\text{logit}\left(\hat{p}_i^*\right) = \mathbf{x}_i\left(\boldsymbol{\theta}^* + \mathbf{q}^*\right)$ where $\mathbf{q}^*$ is a random vector with a multivariate normal distribution $MN\left(\mathbf{0}, \hat{V}\left(\boldsymbol{\theta}^*\right)\right)$ where $\hat{V}\left(\boldsymbol{\theta}^*\right)$ is adjusted by the uncertainty in the number of parameters in the final model. The random vector $\mathbf{q}^*$ is computed using the Cholesky decomposition or the SVD to compute $\sqrt{\hat{V}\left(\boldsymbol{\beta}\right)}$ the same way as in Method 1. The imputed value is the random realization of the Bernoulli trial $BE\left(\hat{p}_i^*\right)$.

- Imputation of account values conditional on the non-zero accounts. After identifying those accounts with a positive balance, a linear regression model is posited to impute the current account values with a new block defined by those cases where the current account value is greater than zero. The best linear model is identified using forward regression based on the AIC with the transformed dependent variable derived in 2. Once the best model is identified, the model is used to predict the imputed values of for the beggars. The imputed values are produced in the same way as in Method 1; that is, the parameters as $\hat{\mu}^*_{zi} = \mathbf{x}_i \left(\hat{\boldsymbol{\beta}} + \mathbf{b}^*\right) + e^*_i$. The random noise $\mathbf{b}^*$ and e^*_i used to shock the predictor are computed in the same way as in Method 1. The imputed value Y^*_i is calculated as the untransformed value of $\hat{\mu}^*_{zi}$.

There is no criterion to determine if the models for zero account or positive account values are overspecified, as in Method 1. However, this could be implemented following the same procedure used in Method 1. We suspect the forward selection procedure avoids this problem in most cases.

F. After imputing the current account value, Method 2 proceeds to impute the previous year's indicators. The procedure is similar to the one used for the imputation of zero accounts. It is based on a logit model for the probability of having an account in the previous year. The best model is identified using the AIC, and the imputed value is generated by shocking the regression coefficients and the realization of the Bernoulli trial as described above.

G. In the next step, the previous year's account values are imputed in blocks defined by the CUs with the A&L account in the previous year. The procedure follows Steps C to E but replacing the current account values and brackets with those from the previous years.

H. Steps A to G are repeated using the imputed variables, replacing the initial values of the missing variables at the beginning of the process. These imputations are repeated using the updated values 10 times to attempt to get the Markov Chain close to its equilibrium distribution so that the fitted model converges to the conditional distribution of the model for the data.

Each instance of steps A through H in Method 2 produce only one imputed value. Therefore, these steps are 5 repeated five times to produce five imputed values for a total of 50 rounds for each variable.

Method 2 generates 5 imputed indicators for the previous year's indicators in addition to the 5 imputed account values (current and previous years). In contrast, Method 1 does not produce multiply imputed values for these indicators. Thus, Method 2 reflects the uncertainty of model parameters and distribution of having the account in the previous year. The additional multiply imputed values require a different file structure when more than one set of multiple imputed values are used in estimation. For example, the last year’s CU mean account value for CUs that reported having an account (see Section 2.3) needs to be structured differently.

Method 2 was implemented in R version 3.6.2 using the R-packages *mice* version 3.1.13.0, *bestNormalize* Version: 1.7.0, *glm2* version 1.2.1, and *MASS* version 7.3-51.4. The code of the package *mice* was heavily edited to address the mixture of imputation methods, imputation of zero accounts, imputation of account conditioned on positive accounts, the transformation of the dependent variables, and multiple blocking because none of these features could be handle using the original code. The imputation models not included in mice were added using new objects linked to the mice function. In other instances, the internal code of the mice function was replaced by custom functions using the *assignInNamespace* function. Modifying the code of the package mice meant that we did not have to write a new package for Method 2.

2.6 Method 3: MICE Imputation with Random Forests

With the development of machine learning algorithms, there has been an explosion of computer-intensive methods for variable selection such as random forests, gradient boosting, LASSO, in addition to the classical methods such as stepwise regression (forward, backward, and both) based on a stopping rule. Our experience is that the difference among these methods is minimal when the goal is to produce a parsimonious model. For imputation, a parsimonious model may have an advantage since the model has a better fit, but the possible disadvantage is that variance estimates of the coefficient may be underestimated.

We suspect that this potential underestimation is the reason for shocking both the estimated coefficient and the model's predictions in the income imputation. On the other hand, the two shocks may not be needed in other imputation methods since the regression coefficients' values are drawn from the empirical distribution of the parameters. The simulations in the next chapter will shed some light on these issues.

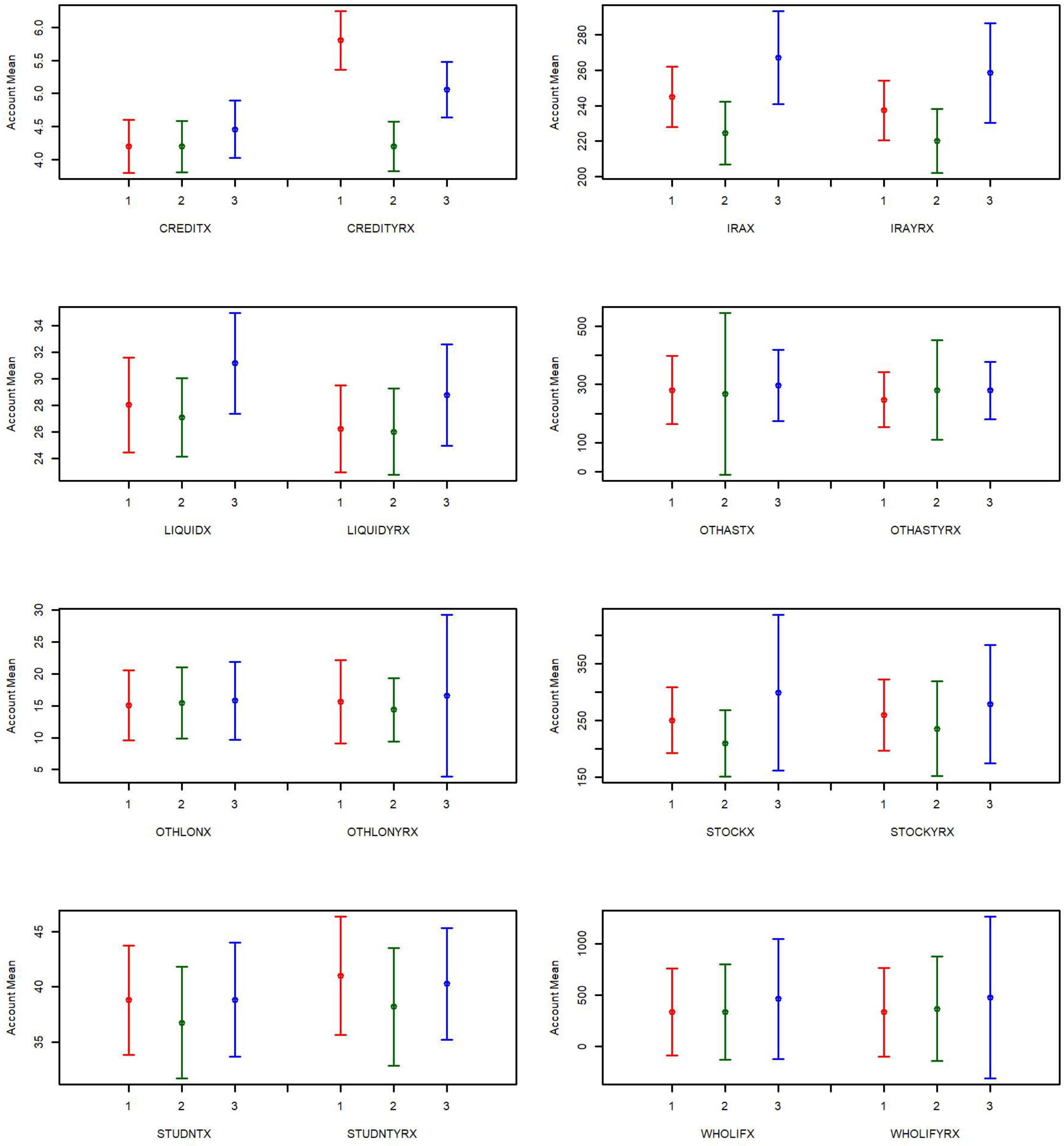
Method 3 is very similar to Method 2. Like Method 2, it is based on an FCS model and the algorithm is MICE. Where it differs from Method 2 is in the approach for variable selection (i.e., model selection) and the prediction of the imputed value. Method 2 uses a greedy algorithm based on the AIC for variable selection and predicts based on the transformed data. Method 3 uses random forests (random forests are a series or ensemble of decision trees based on random subsamples of the data) for variable selection on the untransformed values. The predicted value is the mean of the predictions of the various trees.

As in Method 2, all the types of A&L variables are imputed in the same run, and like Method 2, it uses the imputed variables as predictors. In other words, all the steps given in the previous section with respect to imputation are followed exactly for both Methods 2 and 3. The only differences are: (1) the data are not transformed in Method 3, (2) the variable selection in Method 3 is random forests rather than based on the AIC, (3) the fit for Method 3 is based on the average of the random tree fits rather than from the generalized linear models in Method 2. As with Method 2, if bracket information is provided, mean bracket imputation is done rather than the random forest.

Method 3 was implemented in R version 3.6.2 and the R-package *mice* version 3.1.13.0. This package uses the package *randomForest* version: 4.6-14 to create the random forest for classification (e.g., the indicators for A&L account and zero account) and regression (account values). The same modifications described in Section 2.5 were made to the package mice to address the mixture of imputation methods, imputation of zero accounts, and multiple blocking.

2.7 Comparisons of 2019 Quarterly and Annual Estimates

In this section, we compare the estimates produced by the methods described in Sections 2.4, 2.5, and 2.6. A key point is that most of the estimates are not statistically different for the three methods (although we have not done any statistical testing since they are meant as examples). The 2019 annual estimates and their 95% confidence intervals are shown in Figure 2-5 for A&L mean account values and 2-6 for the proportion of CUs with an account for all the A&L variables for the three methods. The figures show that most of the 95% confidence intervals overlap; there are no statistical differences among the 2019 annual estimates produced by the methods. Figures 2-7 and 2-8 compare the same estimates and the 95% confidence interval produced using the non-imputed cases (in black) with those produced using the imputed cases for the three methods. Figure 2-8 only shows the estimates of proportions computed using the previous year's A&L indicators because, by definition, no current year account indicator was missing and therefore imputed.



*acc

Figure 2-5. Comparisons of estimates of mean account value (in thousands) and their 95% confidence intervals for the A&L variables by imputation method.

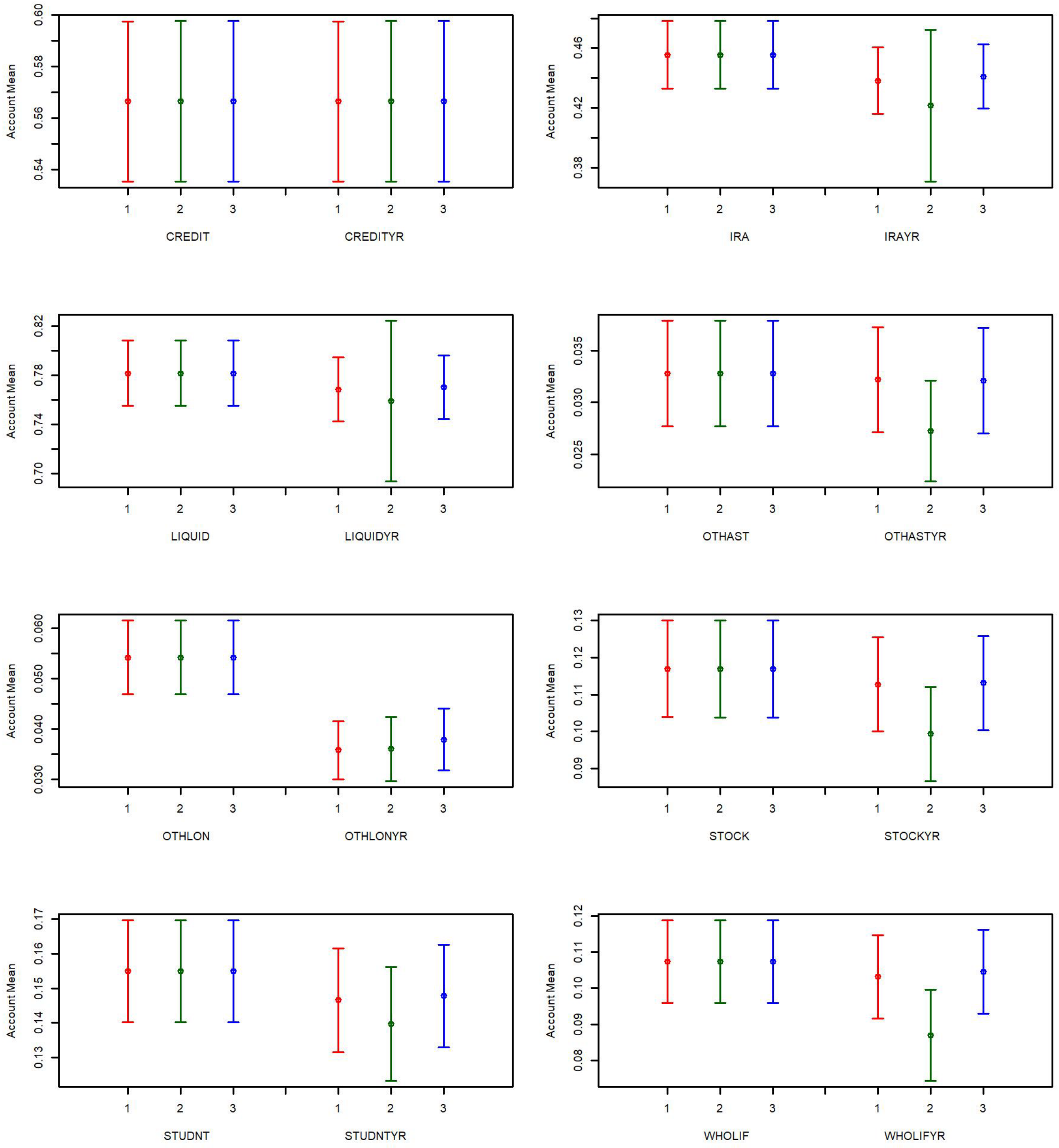


Figure 2-6. Comparisons of estimates of mean account value (in thousands) and their 95% confidence intervals the estimates of the proportion of CUs with an account for the A&L variables by imputation method.

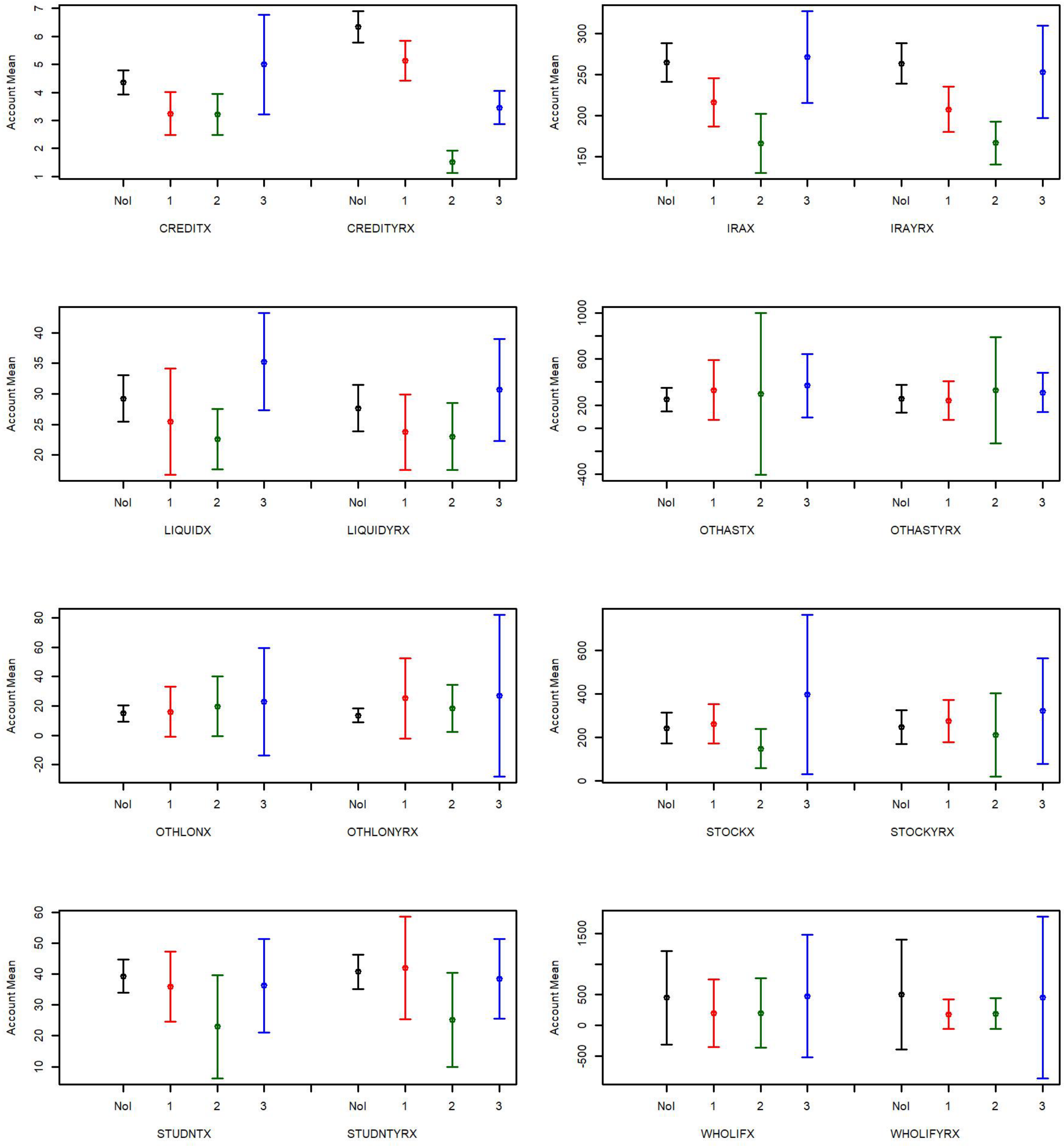


Figure 2-7. Comparisons of estimates of mean account value (in thousands) and their 95% confidence intervals for the A&L variables by imputed and not imputed (Noi) cases by imputation method.

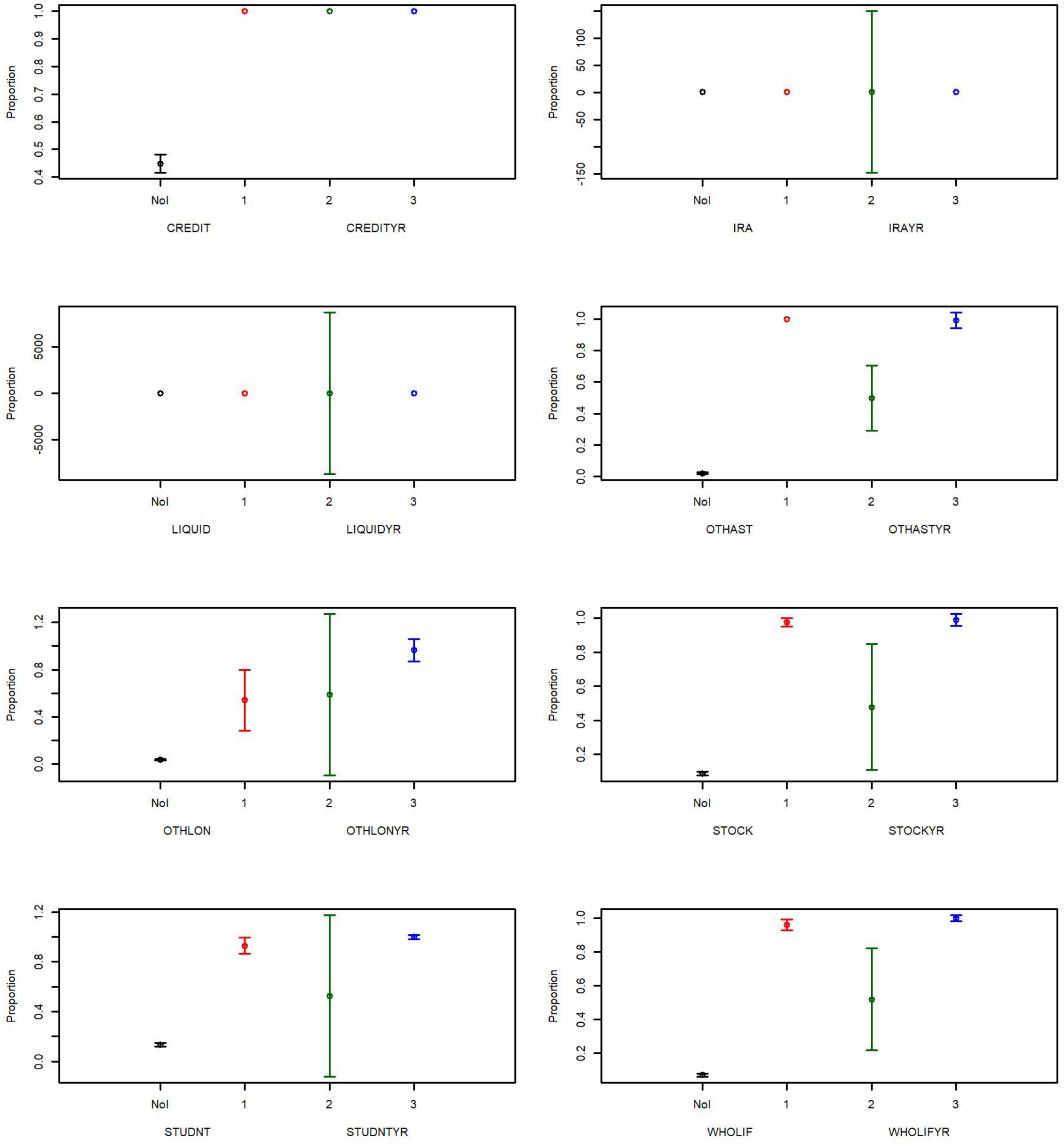


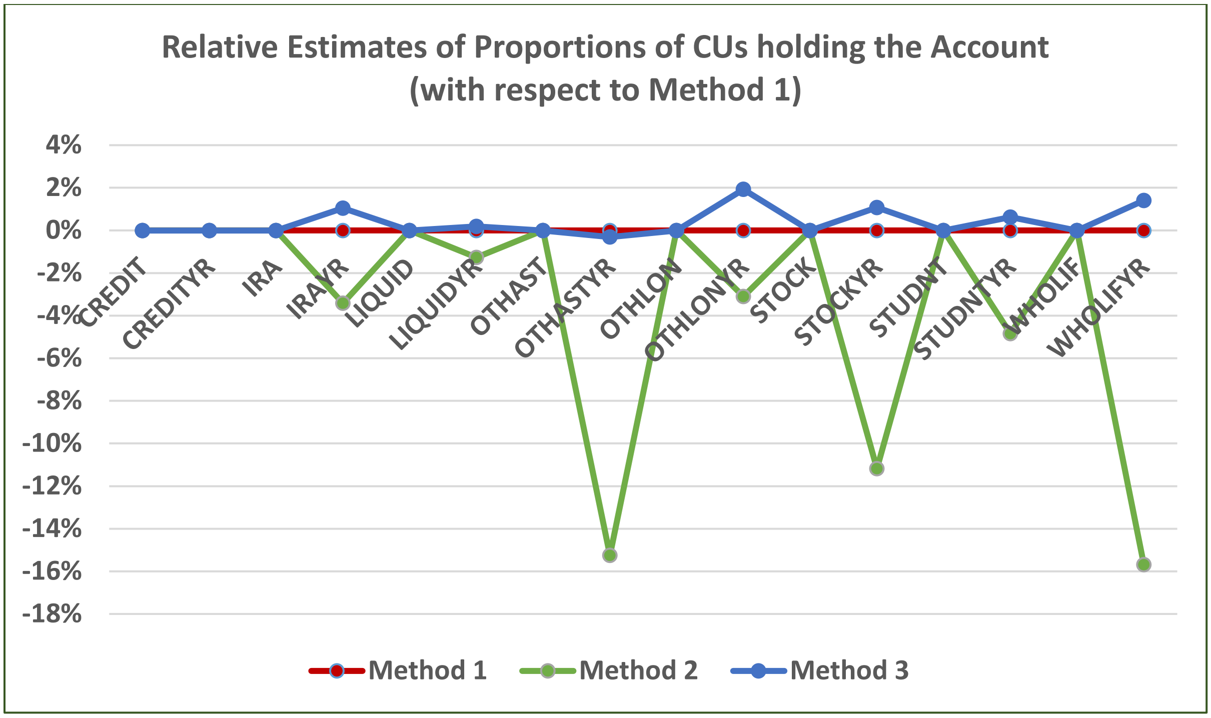
Figure 2-8. Comparisons of estimates of the proportion of CUs with an account and their 95% confidence intervals for the A&L variables by imputed and not imputed (Noi) cases by imputation method.

As noted earlier, these are one random realization of the selected sample and the imputed values and also there is no way to evaluate which of the methods might produce estimates that are closer to the true value (which is unknown). The evaluation of the methods is done through Monte Carlo simulation in Section 3. In this section, we compare the relative size of the estimates and their standard errors defined in Table 2-3 with respect to Method 1 for illustration.

Table 2 3. Definition of the A&L totals, means and proportions statistics.

Estimates	Definition
$\hat{Y}_{Mi}$	**Estimate of the total account value (i.e., the estimate of the sum of all accounts in the US) computed using method *i***
$\hat{N}_{A,Mi}$	**Estimate of the total number of CUs with the account computed using method *i***
$\hat{\bar{Y}}_{Mi} = \dfrac{\hat{Y}_{Mi}}{\hat{N}_{A,Mi}}$	**Estimate of the mean account value (i.e., average account value per CU) computed using method *i***
$\hat{P}_{Mi} = \dfrac{\hat{N}_{A,Mi}}{\hat{N}_{Mi}}$	**Estimate of the proportion of CUs with the account where $\hat{N}_{Mi}$ is the estimate of the total number of CUs in the US.**

Figure 2-7 shows the relative ratios of the proportions and their standard errors of the CUs that hold the account computed as $R_{\hat{P}i} = \frac{\hat{\bar{P}}_{Mi}}{\hat{\bar{P}}_{M1}}$ and $R_{SE(\hat{P})} = \frac{SE\left(\hat{\bar{P}}_{Mi}\right)}{SE\left(\hat{\bar{P}}_{M1}\right)}$. Figure 2-8 shows the relative ratios of the estimate of the total of CUs as $R_{\hat{N}Ai} = \frac{\hat{N}_{A,Mi}}{\hat{N}_{A,M1}}$ and $R_{SE\left(\hat{N}_{A,Mi}\right)} = \frac{SE\left(\hat{N}_{A,Mi}\right)}{SE\left(\hat{N}_{A,M1}\right)}$. The relative size of the proportions in Methods 1 and 3 are closer than those in Method 2. However, the estimates of proportions are very small, so even small differences are large when measured in relative terms. Figure 2-9 and 2-10 show the same relative ratios for $\hat{\bar{Y}}_{Mi}$ and $\hat{Y}_{Mi}$, the mean account value and total account value, respectively. The figures show what appear to be large differences among the methods, but these are all for estimates with small sample sizes



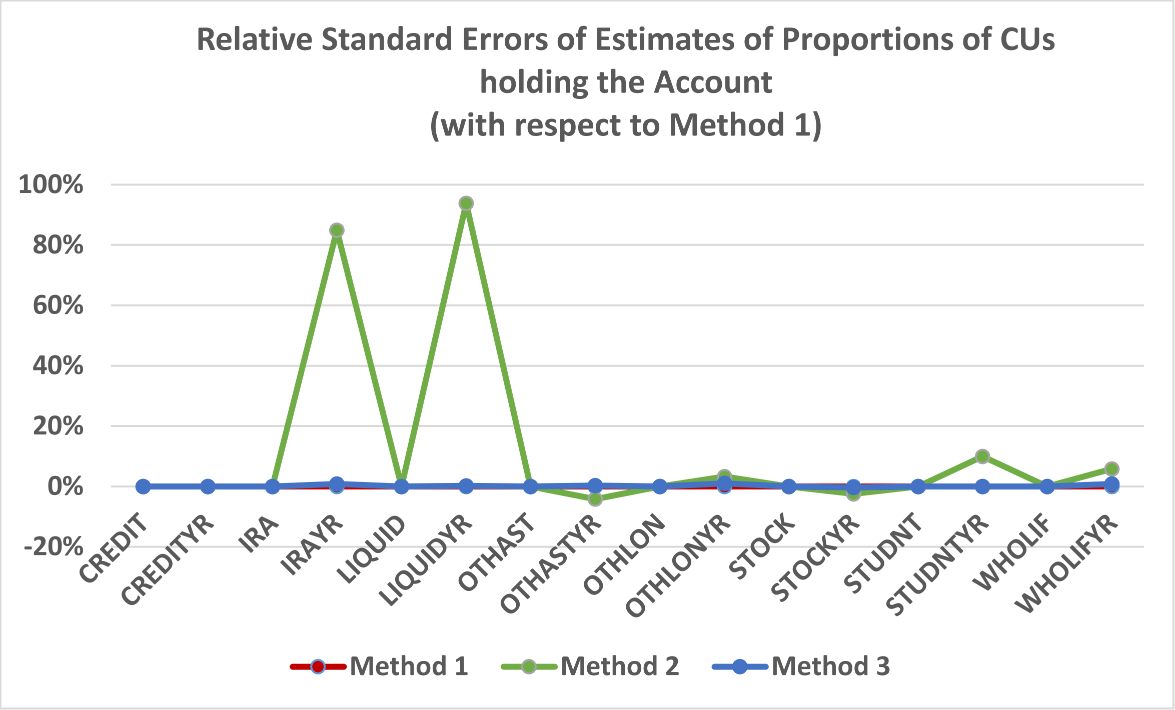
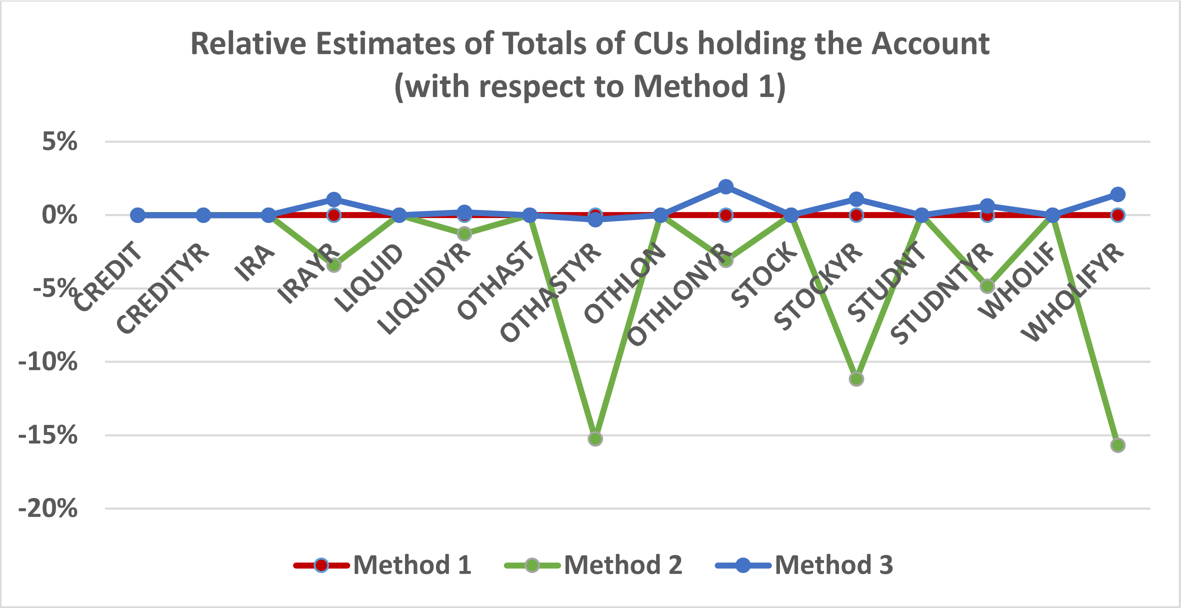


Figure2-7. Comparisons of the relative size of estimates of proportions and their standard errors by imputation method.



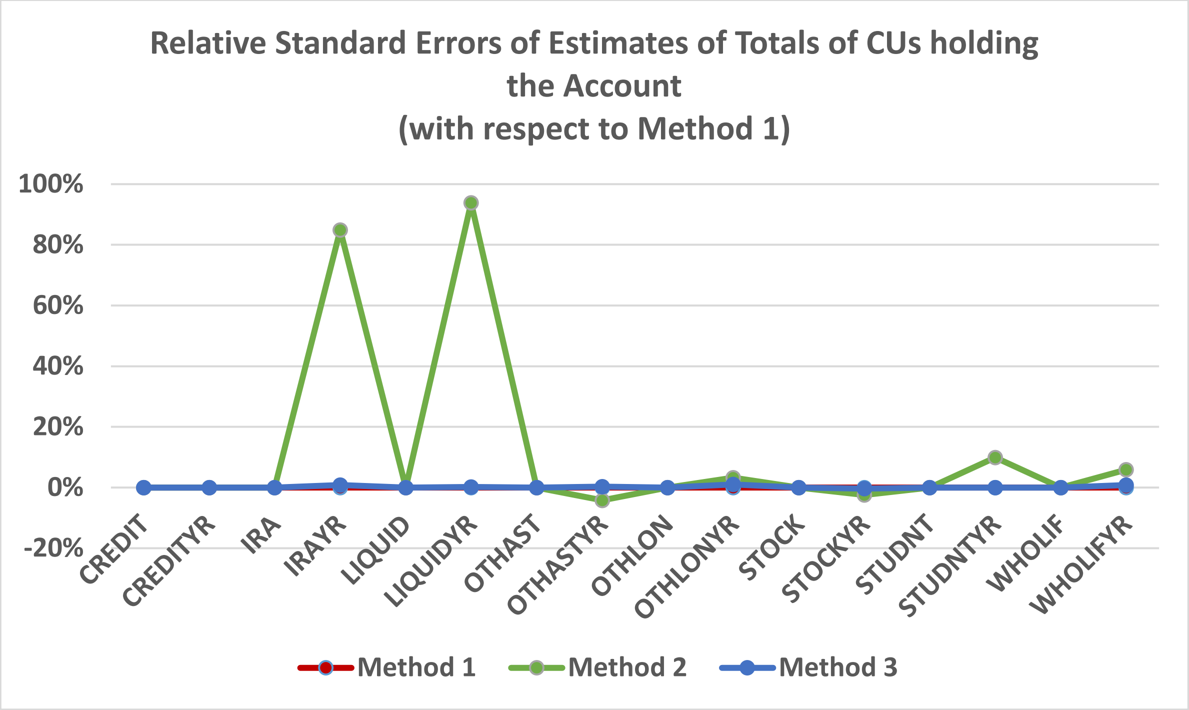
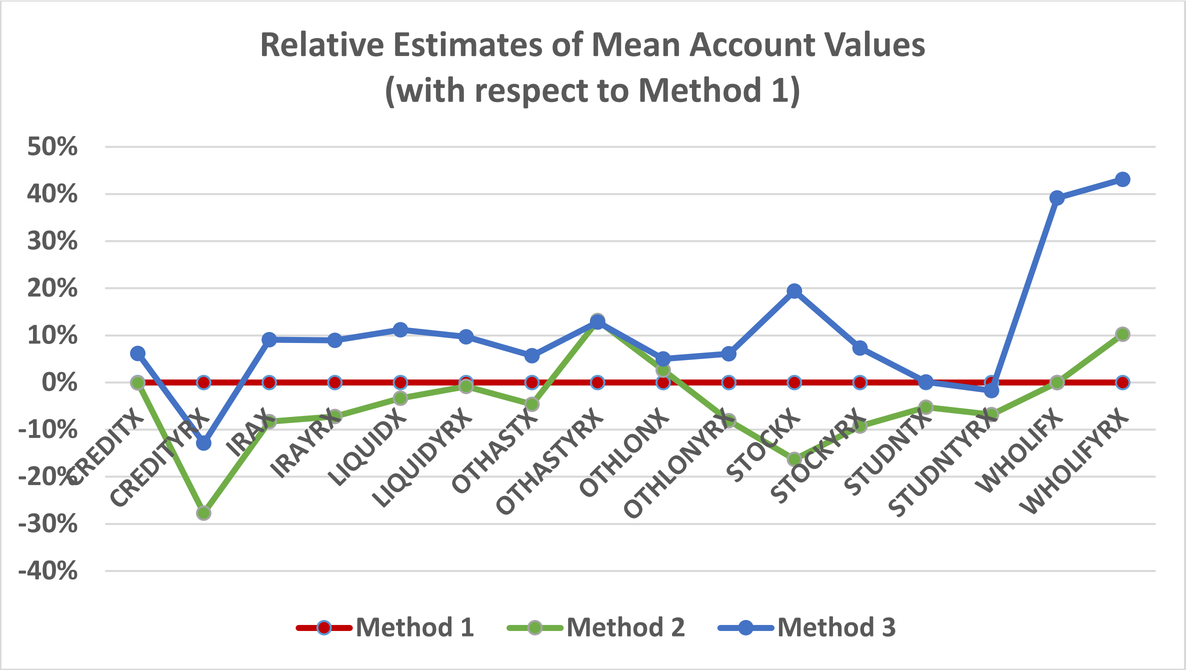


Figure2-8. Comparisons of the relative size of estimates of totals of CUs with the account and their standard errors by imputation method.



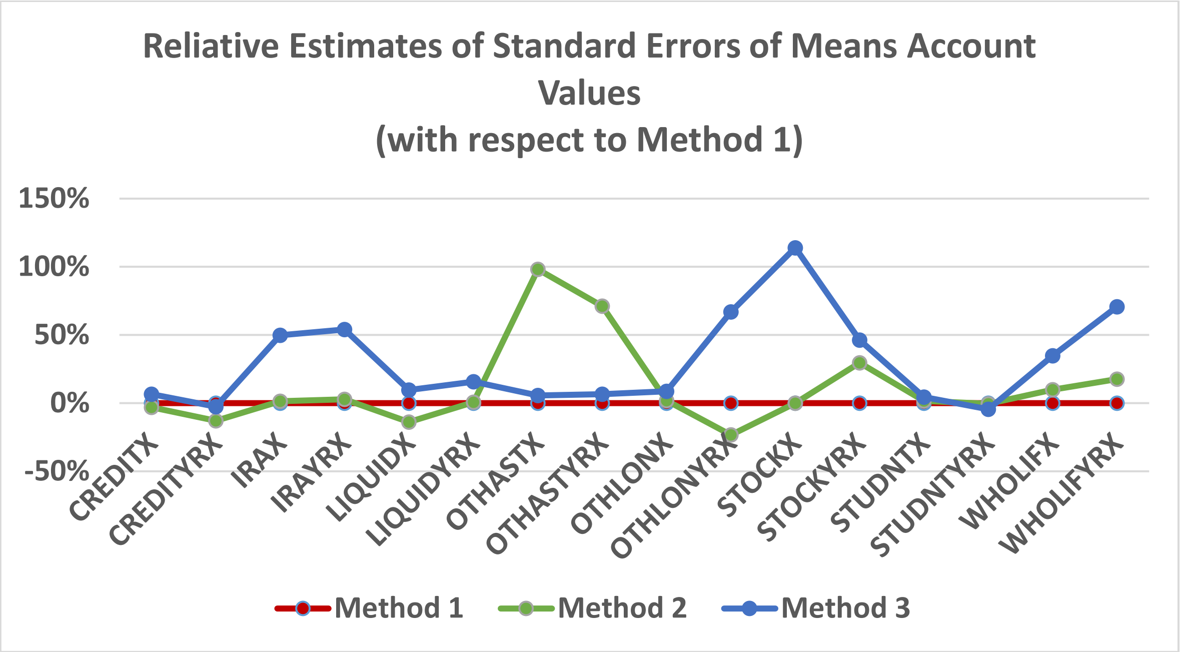
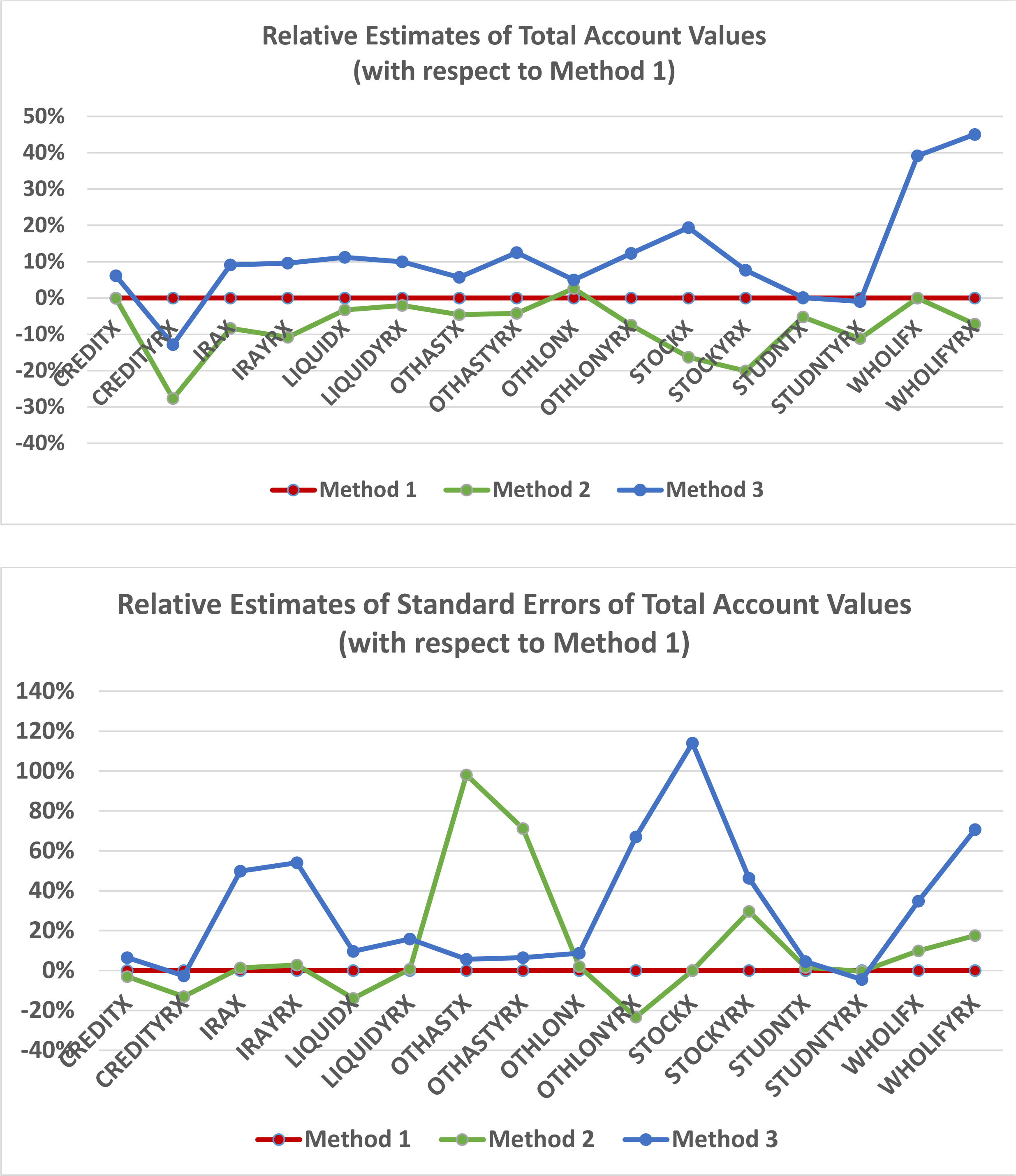


Figure2-9. Comparisons of the relative size of estimates of the mean account value and their standard errors by imputation method.


Relative Estimates of Total Account Values (with respect to Method 1)

Relative Estimates of Standard Errors of Total Account Values (with respect to Method 1)

Figure 2-10. Comparisons of the relative size of estimates of total account value and their standard errors by imputation method.

3. Evaluation of the Imputation Methods

This chapter evaluates the three imputations methods described in Chapter 2 using a Monte Carlo simulation study with repeated sampling of a synthetic population. We use a synthetic population to evaluate the methods because the population parameters (i.e., means proportion and total for all A&L variables) are known in this file. An evaluation based on the 2019 Quarterly and Annual estimates presented in Chapter 2 is impossible because those estimates are single random realizations from the assumed distribution for the sample selection and imputation.

A fixed synthetic population is sampled according to a pre-determined sample design, and empirical estimates are computed using the average of the estimates from each sample. The population and the sample design mimic many of the features of the CE population and sample. The repeated samples were used for producing empirical estimates of selected statistics by the methods. Although we implement a sample design close to the one used in the CE, the evaluation has important limitations. First, due to the computational properties in all the methods, the simulations require much longer running times than a single imputation and analysis. This caused us to limit the number of simulation runs to 1,000 for each method. The second limitation is related to the creation and size of the synthetic population. For purposes of the simulation, we need to create a population sampling frame where the study variable values are known so that we could examine the statistical properties of the estimates. We also need to create missing values when they are selected in the sample in a way similar to the patterns of missing data in the CE. The missing flag indicators were needed for the current and previous year's A&L variables (account indicators, account values, brackets). Furthermore, the relationships among the A&L indicators and the independent variables needed to be replicated as closely as possible.

For all these reasons, we created a synthetic population of CUs much smaller than the frame in the CE. Another difference is that the simulation produces estimates for one quarter instead of annual estimates as done in the previous chapter since the annual estimates require producing four separate quarterly imputations. The other main difference is that the model donor pool is not expanded to include non-missing values for previous quarters in the simulation. Instead, the donor pool is restricted to the complete cases for the quarter being imputed. As a result, the simulation does not reflect the effect of the rotating panels on the model donor pools and removes the effect of averaging donor pools from previous quarters.

The simulation targeted the key elements among the imputation methods, despite these limitations. In particular, the key elements include

- Different models were examined that range from those that keep most of the predictors as in Model 1, to more parsimonious models in Method 2, and random forests, a Statistical Learning or Machine Learning in Method 3.

- Different variable selection methods were tested, with backward linear and logistic regression in Method 1 based on p-values of the regression coefficients, to forward linear and logistic regression using the AIC in Method 2, and the average of an ensemble of decision trees outputting the mode of the classes (classification) or mean prediction (regression).

- The use of transformed variables as the dependent variables was tested, where the dependent variable is the Z-score of the account values in Method 1, and the transformed account value uses the best normalizing transformation in Method 2, and untransformed account values in Method 3.

- The use of imputed variables as predictors in the models was explored. The variables to impute were excluded as predictors in Method 1, while these variables were allowed as predictors in the models in Method 2 and 3 based on the MICE algorithm. These predictors had an impact on the order of variables to impute. The same account type of variables (i.e., all IRA-related variables) can be imputed separately, without any specific order of imputations by account types (e.g., other assets, credit variables). In contrast, Method 2 and Method 3 required a specific order.

- The imputation of either one value or five values of the previous year's A&L account indicators. In Method 1, only one imputed value is produced for the missing values of the previous year's account indicators similar to the CE income imputation. For the other two methods, five imputed values were automatically produced by the MICE algorithm. The 5 repeated values were expected to reflect the additional uncertainty of predicting if the CU had the account the previous year in the same way the 5 repeated values reflected the uncertainty of the account value.

- Different approach to handle zero values account were tested. In Method 1, there is no separate step to identify the proportion of accounts with a zero balance. Methods 2 and 3 specifically include a modeling step that predicts the accounts with zero balances conditional on having the account for the current and previous year.

3.1 Simulation Sample Design

The population distribution, sample design, sample size, and allocation were designed to be close approximations to the CE sample design. Table 3-1 shows the distribution of CU's by sampling strata, sampled PSU number, the average number of CUs in sample PSUs, and the distribution in the sample for 2019 in the restricted-use files.

Table 3-1 Sampling stratum in the CE design.

Sampling strata	Estimated Number of CUs	Percent	Sampled PSUs	Sampled CUs	Percent	Sampled CU/ PSUs
1: Certainty Urban	47,000,000	36%	20	600	39%	30
2: Non-certainty Urban	75,000,000	57%	50	800	53%	16
3: Non-certainty Rural	8,500,000	6%	20	90	6%	5

* Counts in excluded panels where the A&L items were not asked.

The design for the synthetic population is shown in Table 3-2. The goal was to draw close to the same number of PSU and CUs by strata as in the CE in a given quarter. For example, the percentage of PSUs and CUs from the stratum 2 (non-certainty urban) for the synthetic population design is 57% and 53%, respectively, compared to 57% and 53% in the 2019 CE files.

Table 3-2 Sample design of for the synthetic population.

Sampling strata	Number of PSU to sample	Percentage	Number of CUs to sample	Percentage	Average number pf CU/PSU
1: Certainty Urban	20	25%	600	39%	30
2: Non-certainty Urban	50	57%	800	53%	16
3: Non-certainty Rural	20	17%	100	6%	5

In stratum 1, the 20 PSUs were selected with certainty. In the second stage in this stratum, we allocated the 600 CUs to the PSUs using proportional allocation with the PSUs as sampling strata and the squared root of the number of CUs in the PSU as the stratum size. We used the square root of the number of CUs to reflect errors in the number of CUs in the PSUs in the allocation as the exact number of CUs in a PSU is unknown (e.g., out-of-date) before sampling. The incorrect stratum sizes produce a less efficient design avoiding a self-weighting design. After the initial allocation, the sample size of CUs was modified to produce an even number of CUs within each certainty PSU to simplify the pairing for the BRR replicate weights.

In strata 2 and 3, the PSUs were selected in the first stage using probability proportional to the square root of the number of CUs in the PSU as the measure of size. We also use this measure to reflect errors in the number of CUs in the PSUs. Once the PSUs in strata 2 and 3 were selected, a fixed number of CUs; within stratum: 20 CUs were selected from each of the sampled PSUs in stratum 2, while (N<15) CUs were selected from the 20 sampled PSUs in stratum 3. Since the primary sampling unit in strata 2 and 3 is the PSU for replicate pairing, we ensured an even number of sampled PSUs in strata 2 and 3 to facilitate the creation of the BRR replicate weights.

The next step was the creation of the synthetic population. As mentioned earlier, the synthetic population is much smaller than the full population with a distribution of CUs by sampling strata as close as possible to the CE frame (see the third column in Table 3-1) but large enough to be sampled repeatedly.

3.2 Creation of Synthetic Population Frame

The main goal in creating a synthetic population was to include the relationships among the variables in the restricted files (A&L and predictors). This task is very complex. Some approaches rely on defining statistical models that generate all A&L variables, their relationships among themselves and the predictors, and their missing patterns, but this approach was beyond the scope of this research. Instead, we used the restricted files to generate the population, thus ensuring the synthetic population would reflect the relationships among A&L variables found in A&L restricted-use files. In addition, the variables with missing patterns are already identified, so there is no need to generate them. One disadvantage is that the mechanism that generates the missing values is unknown. As a result, this approach prevents describing how the imputation methods worked with respect to a specified known missing value distribution.

The synthetic population was created as follows:

- The restricted us quarterly files from 2017 Q2 to 2019 Q4 were combined to create the initial synthetic population file.

- After combining the files, all CUs with invalid data were removed. The invalid records include those where the current year A&L account indicators were missing (e.g., there is no current-year indicator imputation), cases with both bracket and previous year's account values, cases with zero account values with missing indicators, and cases where the A&L were not asked.

- The procedures for the models used in Method 2 were used to create known A&L variables (indicators and account values) with some modifications. First, the logarithm of the account values was used as the transformed dependent variables in the linear regression models. Second, the predictions of these models were used to create the population A&L variables; that is, these predictors replaced the cases with and without missing values in the restricted used files. Although these synthetic values were correlated with the A&L values, they were not the same. All restricted-use values were then removed from the population file. Other missing A&L population variables such as the current and previous year's brackets were filled out using the synthetic values. Since the population A&L variables are known for all cases in the frame, the population A&L account totals and the number of CUs with a type of A&L account are known.

- No special procedure was implemented to model the CUs with missing A&L values. Instead, the flags for missing AL& variables were the same as the A&L variables in the restricted used file. For example, if a CU in the restricted file had IRAYR, IRAYRB, and IRAYRX missing, the same CU in the population file were assigned to be missing for these variables.

In the next step, we create the PSU and stratum indicators following the distribution of the frame in the CE. We expanded the cases in the final file as follows:

- Since the CE is a two-stage sample where geographic areas (or PSUs) are sampled in the first stage, we retained the original PSU indicators from the restricted files.

- For the PSUs in stratum 1, the original PSU indicator was used to group all the CUs from all quarters. Since the CU sample selection was made within certainty CU, no modification of these PSU indicators was needed, and they contain all the CUs in certainty urban PSUs from all the quarters.

- For strata2 and 3, pseudo-PSUs were defined by the cross-tabulation of the original PSU identifier and the year/quarter indicators. Although this process created many PSUs in these strata, most of them were relatively small to support the number of sampled CUs in the design. Therefore, the pseudo-PSUs in Strata 2 and 3 were combined to produce the final PSU indicator in these strata. The final PSUs were created by combining quarters within the initial PSUs.

- We expanded the synthetic population to reflect the US distribution of CUs from the CE in the last step. First, we doubled the file and recreated the PSU indicators and CU indicators. Then we computed the distribution of CUs by stratum and the average number of CUs per stratum and compared them to the CE stratum distribution. For example, the proportion of CUs in stratum 2 in the double synthetic population file was

53% compared to 57%, the estimated proportion of CUs in stratum 2 in the CE. Next, using the average number of CUs per PSU, we randomly selected additional PSUs from strata 2 and 3 and added the CUs from the sampled PSUs to the file to match the CU's distribution in the synthetic population to those estimated proportions in the CE. The final distribution of PSUs and CUs in the final synthetic population file is shown in Table 3-3.

Table 3-3 Final distribution of PSUs and CUs in the Synthetic Population.

Sampling strata	Number of PSUs in Frame	Percentage	Number of CUs in Frame	Percentage	Average number of CU/PSU
1: Certainty Urban	20	2%	17,500	37	875
2: Non-certainty Urban	600	77%	2,000	56	3
3: Non-certainty Rural	150	20%	1,500	6	10

Table 3-4 the synthetic population A&L account totals and the number of CUs with the A&L. The table also shows the number of cases in the population with missing A&L indicator and account values.

Table 3-4 A&L account statistics and number of missing values in the synthetic population.

Period	Account	Number of CUs	Missing Account indicator	Total account value	Missing account values
Current year	CREDIT	18,000	0	38,000,000	2,200
	IRA	14,000	0	2,000,000,000	5,700
	LIQUID	24,000	0	260,000,000	7,300
	OTHAST	1,100	0	110,000,000	450
	OTHLON	1,800	0	10,000,000	200
	STOCK	4,100	0	260,000,000	1,500
	STUDNT	4,700	0	110,000,000	600
	WHOLIF	3,600	0	95,000,000	1,600
Previous year	CREDITYR	18,500	6,800	15,000,000	9,000
	IRAYR	14,000	2,800	1,800,000,000	6,900
	LIQUIDYR	24,000	4,600	230,000,000	8,900
	OTHASTYR	1,000	350	120,000,000	650
	OTHLONYR	1,300	100	8,800,000	300
	STOCKYR	4,000	800	250,000,000	1,900
	STUDNTYR	4,600	350	110,000,000	1,000
	WHOLIFYR	3,400	1,000	79,000,000	2,100

3.3 BRR Replicate Weights

In each simulation run, a sample of 1,500 CUs was drawn using a two-stage sample design described in Table 3-2. The CU sampling weight w_{hji} is

$$w_{hji} = \frac{1}{P_{hj}} \frac{1}{C_{hji}},$$

where P_{hj} is the probability of selection of PSU j in stratum h computed as $P_{hji} = 1$ if the PSU is in stratum $h = 1$ or $P_{hj} = m_h \frac{M_{hj}}{M_h}$ if the PSU is in stratum $h = 2,3$, where m_h is the number of PSUs drawn in stratum h, M_{hj} is the measure of size of the PSU j in stratum h, C_{hji} is the probability of selection of CU i in PSU hj where $C_{hji} = \frac{n_{hj}}{N_{hj}}$ and n_{hj} is the number of CUs drawn in the PSU hj and N_{hj} is the number of CUs in the PSU hj . We assumed there was no unit nonresponse in the simulation, so the sampling weights did not require any additional adjustments.

Following the sample design in the CE, we created 44 replicates weights for variance estimation[9]. A total of 44 replicate weights were created using BRR with a Hadamard matrix of size 44. The BRR replicate weights were created using the first randomly selected units; in other words, in stratum 1, the first randomly selected units were the CUs within the certainty strata, while for strata 2 and 3, the selected units were the PSUs. The BRR replicate weights were created as follows .

- The first selected random units were paired and numbered to create the variance strata. In each certainty PSU, the pairs consist of two CUs, while in Strata 2 and 3 the pairs consist of two PSUs. The number of variance strata in the certainty PSU varied depending on the number of CUs sampled in the PSU; sometimes, the number of variance strata was greater than 44. In contrast, there were always 26 and 8 variance strata in strata 2 and 3, respectively.

- In the next step, in the strata with more than 44 replicates, the sequences of pairs were folded so that the maximum number of variance strata is 44. This process is done by replacing the values greater than 44 with a new sequence starting with 1. This process is equivalent to an additional pairing of the sampling units.

- The replicates of these strata minimize the overlap across strata. This assignment is illustrated in Figure 3-1.

[9] The process for creation of replicates differs somewhat from the CE process. One issue is that we did not have the groups used in the CE to form pairs within the certainty stratum so we randomly sample CUs for pairing. The other issue is that our analysis of the file from BLS with the pairs for the CE suggested that the replicate groups from strata 2 and 3 overlapped with each other but not with stratum 1. To increase the degrees of freedom for our analysis, we completely overlapped stratum 1 with the other strata.

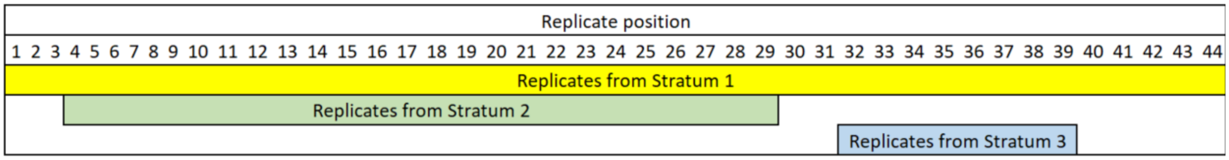


Figure 3-1 Variance stratum and replicate assignment for the simulation study.

- In the last step, each element in the variance strata is assigned a variance unit with values 1 or 2. The 44 replicate weights are created using a Hadamard matrix and matching the variance strata with the rows of the matrix and multiplying the weights for the cases in the variance unit by 0 or 2 depending on the entry of the matrix (i.e., 1 or -1).

3.4 Evaluation Statistics

In the simulation, we conducted $B =$ 1,000 simulation runs for each method and computed the summary measures of bias and accuracy of the estimates as listed in Table 3-5.

Table 3-5 Evaluation statistics.

Statistics	Description
$RB\left(\hat{E}_M\right)\% = 100 \times \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{E}_{Mb}-E}{E}$	Relative Bias
$RRMSE\left(\hat{E}_M\right) = 1000\sqrt{\frac{MSE\left(\hat{E}_M\right)}{E^2}}$	Relative Mean Squared Error
$RCI\left(\hat{E}_M\right) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}_{\lvert\hat{E}_{Mb}\rvert<z_{0.95}\sqrt{\hat{V}\left(\hat{E}_{Mb}\right)}}(x)$	95% Confidence Interval
$RRCI\left(\hat{E}_M\right) = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{U}\left(\hat{E}_{Mb}\right)-\hat{L}\left(\hat{E}_{Mb}\right)}{E}$	Relative range of the 95% Confidence Interval

where $\hat{E}_{Mb}$ is the estimator based on imputation method M of the population parameter E for the simulation run b for $b \in \{1,\ldots,B\}$, $MSE\left(\hat{E}_M\right) = \frac{\sum_{b=1}^{B}\left(\hat{E}_{Mb}-E\right)^2}{B}$ is the empirical mean squared error, $\mathbf{1}_{\lvert\hat{E}_{Mb}\rvert<z_{0.95}\sqrt{\hat{V}\left(\hat{E}_{Mb}\right)}}(x)$ is the indicator faction with a value of 1 if $\left|\hat{E}_{Mb}\right| < z_{0.95}\sqrt{\hat{V}\left(\hat{E}_{Mb}\right)}$ or

0 otherwise, $z_{0.95} = \Phi(0.95)$ is the 0.95 critical value for the standard normal distribution, $\hat{U}\left(\hat{E}_{Mb}\right) = \hat{E}_{Mb} + z_{0.95}\sqrt{\hat{V}\left(\hat{E}_{Mb}\right)}$ and $\hat{L}\left(\hat{E}_{Mb}\right) = \hat{E}_{Mb} - z_{0.95}\sqrt{\hat{V}\left(\hat{E}_{Mb}\right)}$ are the upper and lower limits of the 95% confidence interval of the estimate $\hat{E}_{Mb}$ for run b. For example, the empirical statistics for the estimate of the mean retirement account value computed using Method 1, $\hat{\bar{Y}}_{IRAX,1}$, are

$$RB\left(\hat{\bar{Y}}_{IRAX,1}\right)\% = 100 \times \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{\bar{Y}}_{IRAX,1} - \bar{Y}_{IRAX}}{\hat{\bar{Y}}_{IRAX,M}}$$

Relative Bias of $\hat{\bar{Y}}_{IRAX,1}$

$$RRMSE\left(\hat{\bar{Y}}_{IRAX,1}\right) = 1000\sqrt{\frac{MSE\left(\hat{\bar{Y}}_{IRAX,1}\right)}{\bar{Y}_{IRAX}^{2}}}$$

Relative Mean Squared Error $\hat{\bar{Y}}_{IRAX,1}$

$$RCI\left(\hat{\bar{Y}}_{IRAX,1}\right) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}_{\left|\hat{\bar{Y}}_{IRAX,1}\right| < z_{0.95}\sqrt{\hat{V}\left(\hat{\bar{Y}}_{IRAX,1}\right)}}(x$$

95% Confidence Interval of $\hat{\bar{Y}}_{IRAX,1}$

$$RRCI\left(\hat{\bar{Y}}_{IRAX,1}\right) = \frac{1}{B}\sum_{b=1}^{B}\frac{\hat{U}\left(\hat{\bar{Y}}_{IRAX,1}\right) - \hat{L}\left(\hat{\bar{Y}}_{IRAX,1}\right)}{\hat{\bar{Y}}_{IRAX,1}}$$

Relative range of the 95% Confidence Interval of $\hat{\bar{Y}}_{IRAX,1}$

3.5 Simulation Results

Although there are differences in performance among the estimators within the imputation methods, we begin by looking at the means and medians of the estimates for all A&L variables, separately for the whole population and by domains, for each imputation method. These results are summarized in Table 3-6 and Table 3-7. Both tables show the means and medians of the empirical bias, absolute bias, relative root mean squared error (RRMSE), the 95% confidence intervals, and the relative range of the 95% confidence interval described in Table 3-4. Table 3-6 shows the means and medians of the statistics related to the population proportions and totals of CUs with the A&L account (i.e., A&L account indicators). Table 3-7 shows the empirical statistics related to the population totals and means of A&L account values. The mean in the tables was computed as the average of 1,000 empirical estimates for the simulations for each statistic separately by population and domain parameters. The median was computed as the middle value of the distribution of 1,000 empirical estimates within the same groups.

Since the main goal was the feasibility to produce data for making inferences using the imputed data, we focus on the coverage of the empirical 95 percent confidence intervals (CI) of estimates using the Monte Carlo simulations. The goal is to achieve the nominal 95 percent CIs[10] in repeated sampling. We examined the mean and median in Columns 10 and 11 in Table 3-6, which showed no differences in coverage of the 95% confidence interval for estimates of both the entire population and the domains. Furthermore, the means and medians achieved the nominal value in all methods.

For population totals and means of A&L account values, the results in columns 10 and 11 in Table 3-7 show small differences in the 95% confidence intervals among these methods[11]. In this table, the coverages of the 95% confidence intervals are not nominal but still have medians at the 90% level or higher. These results do not vary much by the imputation method, either for the entire population or domains.

Comparing the columns for the bias in Table 3-6 with those in Table 3-7 shows the imputation methods did a better job attributing the status of CUs A&L account compared to the imputed A&L account value. These results were expected as it is easier to impute an A&L account indicator (e.g., CU holds or does not hold the A&L) than the A&L account value conditional of the CU holding the A&L account. However, no method of imputation is clearly better for imputing either A&L account means or totals based on the simulation results.

[10] In a 95% nominal confidence interval, the coverage probability or probability that the interval includes the true population is 0.95.

[11] These differences are not statistically significant for the number of simulations.

Table 3-6 Means and medians of evaluation statistics for estimates of total and percentage of CUs with the A&L account for the population and domains by imputation method.

Type	Method	Metric	Bias		Absolute Bias		RRMSR		95% Confidence Interval		Relative Range of 95% CI	
			Total	Propor-tion	Total	Propor-tion	Total	Propor-tion	Total	Propor-tion	Total	Propor-tion
All	1	Mean	-0.55	-0.57	0.60	0.61	7.08	6.86	0.95	0.95	0.28	0.29
All	1	Median	-0.13	-0.15	0.22	0.22	6.85	6.71	0.96	0.95	0.28	0.29
All	2	Mean	-2.02	-2.06	2.09	2.11	7.81	7.62	0.94	0.94	0.30	0.31
All	2	Median	-0.60	-0.63	0.72	0.71	6.86	6.71	0.95	0.95	0.28	0.29
All	3	Mean	0.03	-0.01	0.42	0.40	7.11	6.84	0.95	0.95	0.28	0.29
All	3	Median	0.19	0.15	0.29	0.25	6.86	6.65	0.95	0.95	0.28	0.28
Domains	1	Mean	-0.55	-0.58	0.67	0.68	11.07	10.52	0.95	0.94	0.43	0.45
Domains	1	Median	-0.13	-0.14	0.31	0.29	8.37	7.69	0.95	0.95	0.31	0.34
Domains	2	Mean	-2.09	-2.14	2.23	2.24	11.54	11.03	0.94	0.94	0.45	0.60
Domains	2	Median	-0.75	-0.82	0.80	0.91	8.89	8.21	0.95	0.95	0.34	0.36
Domains	3	Mean	0.16	0.10	0.55	0.52	11.12	10.53	0.95	0.95	0.43	0.45
Domains	3	Median	0.24	0.22	0.33	0.30	8.25	7.59	0.95	0.95	0.31	0.34

Table 3-7 Means and medians of evaluation statistics for estimates of total and mean A&L account values for the population and domains by imputation method.

Type	Method	Metric	Bias		Absolute Bias		Absolute RRMSR		95% Confidence Interval		Relative Range of 95% Confidence Interval	
			Total	Mean	Total	Mean	Total	Propor-tion	Total	Mean	Total	Mean
All	1	Mean	0.00	0.54	3.32	3.39	14.66	12.51	0.87	0.89	0.40	0.50
All	1	Median	0.25	0.64	2.64	1.56	13.30	11.77	0.90	0.91	0.32	0.39
All	2	Mean	0.00	0.54	3.32	3.39	14.66	12.51	0.87	0.89	0.40	0.50
All	2	Median	0.25	0.64	2.64	1.56	13.30	11.77	0.90	0.91	0.32	0.39
All	3	Mean	-1.52	0.60	3.71	4.52	17.75	15.81	0.89	0.89	0.59	0.63
All	3	Median	-1.61	-0.62	2.98	3.81	12.39	10.30	0.92	0.91	0.31	0.40
Domains	1	Mean	3.43	4.01	6.28	6.44	22.64	19.14	0.88	0.89	0.65	0.78
Domains	1	Median	0.53	0.73	3.36	3.27	17.42	14.01	0.92	0.91	0.49	0.59
Domains	2	Mean	0.76	3.02	5.08	6.16	24.06	21.03	0.90	0.90	4.36	1.27
Domains	2	Median	-0.65	0.43	4.14	4.23	17.50	15.43	0.93	0.92	0.51	0.63
Domains	3	Mean	0.89	0.70	4.13	4.07	19.82	15.93	0.90	0.91	0.60	0.75
Domains	3	Median	0.02	0.01	2.23	1.89	16.62	14.14	0.93	0.93	0.47	0.59

To get a fuller picture of the simulation results, we examined the distribution of the 1000 simulation estimates separately by A&L variable (account indicator and value) for the current and previous year for estimates of proportions of CUs with the A&L account and the mean account value. We also repeated the examination by looking at domain estimates. For example, the upper plot in Figure 3-2 shows the boxplots of the 1,000 estimates of the proportion of CUs with a credit card at the time of the interview (CREDIT) and the previous year (CREDITYR) by imputation method. The lower plot shows the boxplots for the estimates of the mean of credit card balance at the time of the interview (variable CREDITX) and the previous year (CREDITYRX). The plots in Figures 3-3 and 3-4 are similar boxplots for the domains estimates for education attainment in CU, HS_1=1 (high school or less) and household tenure, OWN_1=1 (own household).

There are also differences by periods (current and previous year) for the same account. For Method 1, the only method that does not implement a special step for zero accounts, the mean of the previous year estimate (CREDITYRX), has a positive bias. In contrast, this bias is not present for either Methods 2 or 3. This bias is observed only for CREDIT. One hypothesis is that the proportion of zero accounts for the other variables is very small, so excluding this step in Method 1 does not result in a substantial bias. Notice that the bias for this variable could be addressed with Method 1 but would require an additional step similar to that used in the other methods.

Another observation is the larger variability in the estimates of means for Method 2 in the box plots for OTHASYRX and OTHLOANYRX. These results may be a side effect of the limited simulation as the model donor pool only includes cases from the current quarter. Expanding the model pool for the A&L variables with donors from previous as is proposed for production will reduce this variability with more robust models.

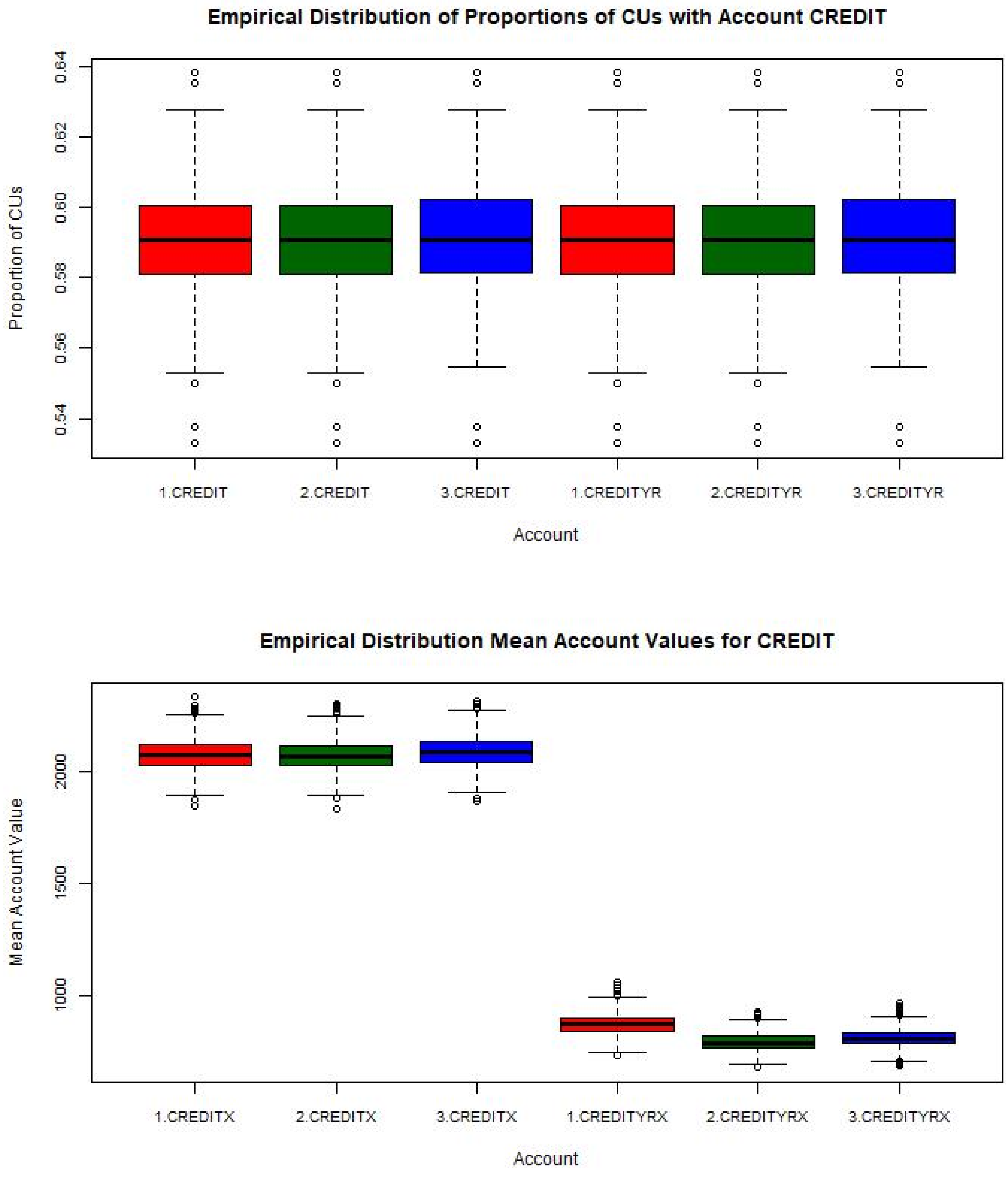


Figure 3-2 Empirical distribution of estimates for CREDIT for 1,000 simulation runs.

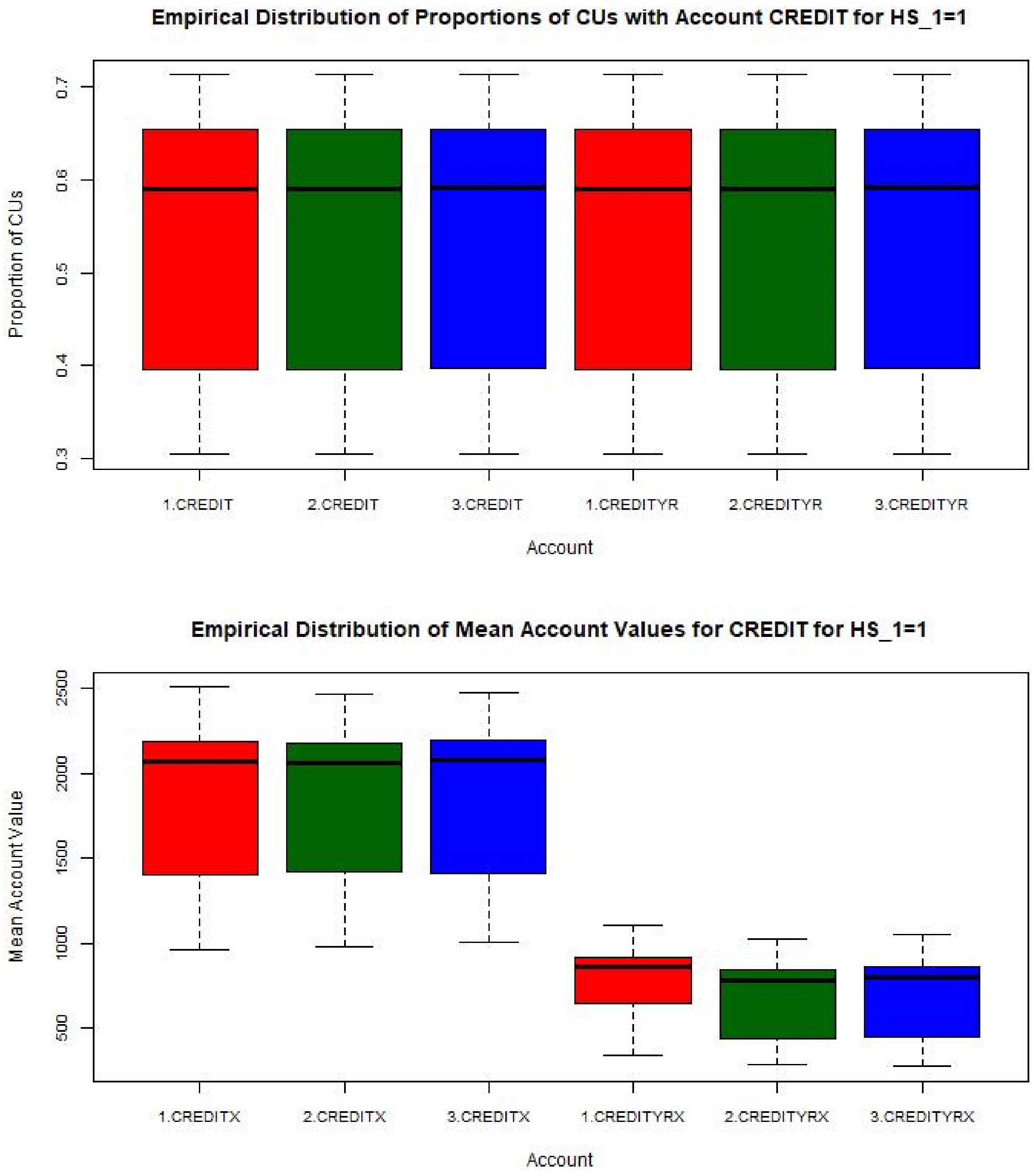


Figure 3-3 Empirical distribution of estimates for CREDIT for 1,000 simulation runs.

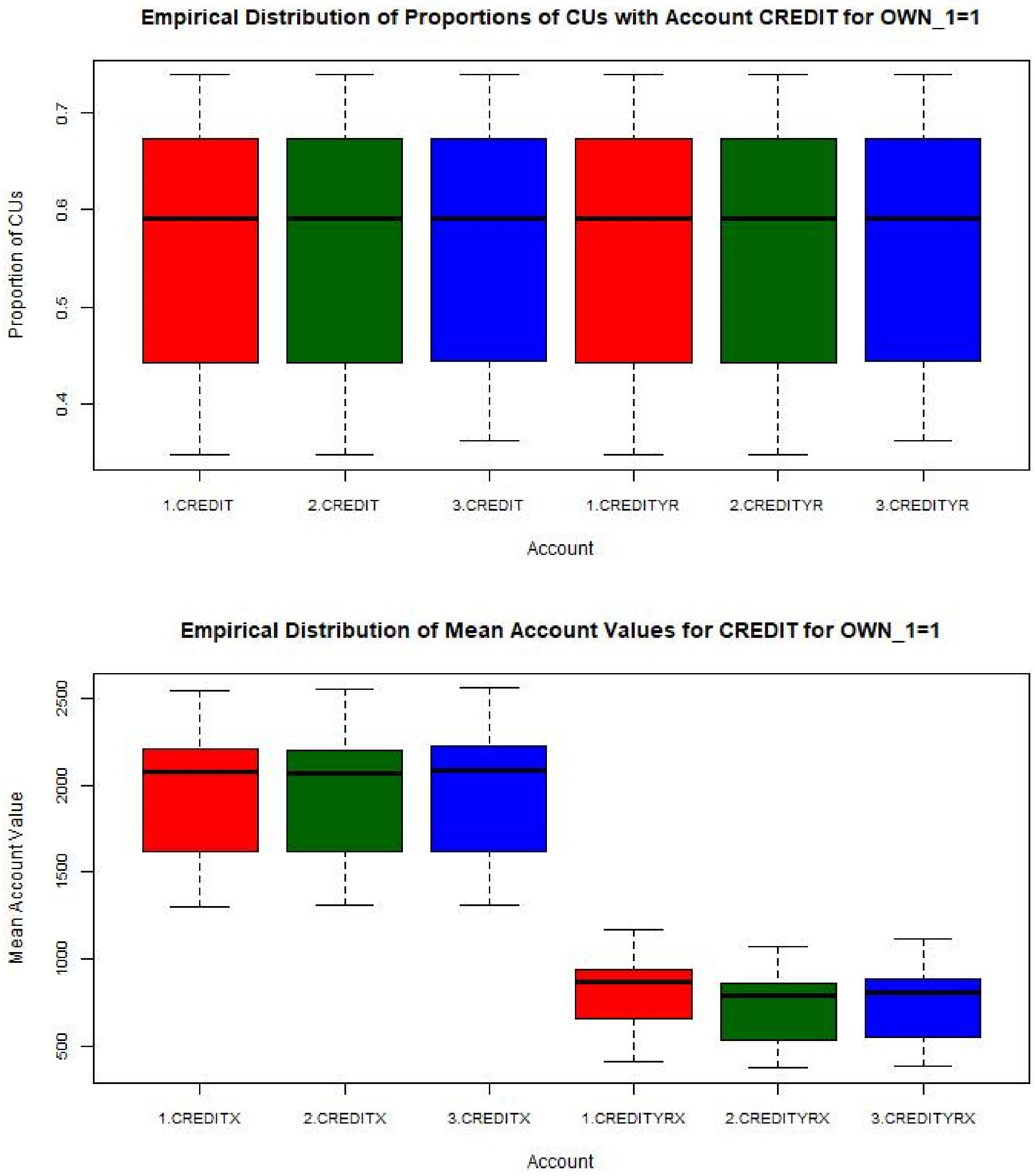


Figure 3-4 Empirical distribution of estimates for CREDIT for 1,000 simulation runs.

4 Conclusions and Recommendations

The simulation findings clearly show that it is feasible to impute missing A&L variables from the CE within the constraints imposed by the data collection production protocols. We examined the performance of the methods based on two types of variables when their estimates are jointly examined: 1) the A&L account indicators used to produce totals and proportions of CUs that hold the account and 2) the A&L account values used to produce estimates of means and totals of account values conditional on the CUs who hold the account. For the A&L account indicators, all three imputation methods performed well and produced nearly unbiased estimates of proportions and totals of account indicators with confidence intervals close to the nominal level for estimates. For A&L account values, the mean and total estimates and their confidence intervals were less accurate than those of the indicator estimates due to the relatively small biases in the point estimates. These biases caused the confidence intervals to cover at lower than the nominal levels, but the confidence intervals for most account values estimates were only slightly lower than the nominal level. Nevertheless, there were some exceptions with bigger biases and lower than desired confidence intervals.

Our recommendation for the multiple imputation method for production considers the statistical results (i.e., inferences), programing language, implementation, and the relationship with current CE procedures. From the statistical point of view, the results show that the three imputation methods, which rely on different algorithms and assumptions on how the data are generated, produce similar inferences. Therefore, no superior method can be identified. In this situation, there are no statistical advantages from using more complex procedures. One possible reason for the equivalence of the methods is the abundance of highly correlated auxiliary variables that are not commonly found in other surveys. The finding that no method is superior is important because it implies that the decision on the imputation method can be based on non-statistical criteria such as operational efficiency and convenience.

When the same type of variables are examined separately for statistics, such as the bias of means, proportions, and totals for different periods (at the time of the interview or the previous year) and different population or domain parameters, the evaluation shows some differences in performance among the methods. Still, no clear pattern of one method being superior can be identified – for example, one method produces mean estimates with smaller biases, but the total estimates have larger biases than the other methods.

Some of these differences are due to how the methods are set up to handle the CE data structure. A key example is the imputation of zero-balance accounts. Method 1 is the only one that does not have a separate step to address the fact that some accounts have zero balances. This omission is because the income imputation model did not have such a step, and the goal was to be consistent with this method. A change in Method 1 could be considered for those accounts that change quickly and where a zero balance is sensible. Such a change would be straightforward. However, this suggestion does not imply implementing different ways to handle each type of A&L account. Instead, the observed data should determine when imputing zero-balance account values. In this investigation, the need for this step appears to be only for those accounts with a non-negligible proportion of zero balances.

We now consider the software-related issues for the implementation as criteria for recommending a method. Although Method 1 was implemented in SAS, it could be implemented in R. However, it is more difficult to implement Methods 2 and 3 in SAS. The reason is that R is a low-level specialized language with a big advantage for complex statistical computations, but it also has a limited capacity for data processing. R is an open-source language, and among the advantages are its availability without a license or fee, platform-independent, availability of user-contributed packages, and continuous growth. Among the disadvantages is data handling, where data sets are kept in memory, limiting their use for large data sets, a complicated syntax language with a steep learning curve, package maintenance that depends on R contributors who may no longer maintain their packages, so the methods become deprecated, and the need of strict coding standard as the language offers multiple functions to produce the same operations. The last disadvantage leads to code being difficult to maintain.

Based on our experience, the main challenges of using R as a production language are the lack of default, well-formatted output, difficulty in debugging code, and the variable quality of the packages. An example of the standard output is the complexity of native code needed for producing a simple frequency table with cumulative counts and percentages (one option is to rely on a package that may become obsolete). Since data processing is a large part of the imputation process, this makes production implementation more difficult. This limitation is evident when control output is such as a simple frequency table requested to determine if the common data problems such as duplicate values or incorrect data file merges occur. Furthermore, debugging R coded is difficult as most of the error messages are uninformative. Finally, the quality of the R packages is highly variable as there are no incentives for contributors to improve them. For example, we found errors in the package mice. Furthermore, the available packages were heavily modified to address the unique features of the imputation of the A&L imputation method and data (e.g., use of a mixture of imputation models such as linear regression and mean bracket imputation or the nested structure of the missing values of the A&L indicators and values).

On the other hand, SAS is a procedural language easier to learn, handles large data sets, is easy to debug, algorithms are fully tested by the SAS developers, the language complies with many data security requirements as SAS is a closed source language that can only be modified by the developers, and produces a well-formatted and easy to understand output. On the other hand, the main disadvantage is the cost – SAS is expensive. Moreover, the complex statistical methods not found in the available procedures are harder to implement in SAS as they need to be encapsulated with specific procedures.

Although Method 1 was implemented in SAS, an alternative language such as SPSS could be used. SPSS shares many of the same advantages and disadvantages as SAS except for the cost. Based on the software characteristics and the fact that the more complex methods do not produce significantly better inferences, SAS-like languages are better suited for producing the A&L multiple imputation since they are easier to maintain, mostly are bug-free, and have customer support.

A possible hybrid approach is to rely on SAS for the main data processing steps and use R for the specialized imputation procedures. Such an approach might be very attractive if either Method 2 or Method 3 were superior to Method 1 but little seems to be gained by the hybrid in the current situation. Moving between SAS and R is feasible, but it is not seamless and adds complexity and opportunities for error. Thus, we see no reason for recommending the hybrid approach.

Another criterion for recommending Method 1 is the timing and flexibility of producing the imputed values. The main advantage of Method 1 is its flexibility because the same type of A&L variable (e.g.., retirement account or credit cards) can be imputed independently in any order. The imputation results can be saved without reimputing all the A&L variables in case of failures in the imputation methods for specific variables. Since the method does not rely on the equilibrium of a Markov chain, there is no need to run multiple iterations as in Methods 2 and 3. Methods 2 and 3 use chained equations, where the sequence of the imputation matters and where all variables are imputed simultaneously in a single run. In case of an error in one variable, the imputation process needs to be re-run with the likelihood of not reproducing the same imputed values despite fixing the random seeds (the variables are predictors in the model). The process is also repeated to achieve equilibrium in the Markov chain. As a result, the implementation time of these methods is l0 times longer than Method 1. In this case, Method 1 is better suited for the imputation of the A&L variables.

We now focus on the advantages of the implementation of the methods based on the current CE production. Method 1 has some advantages in this regard because it is similar (in fact, it was built to be similar) to the imputation process for income currently implemented in production in the CE program. Unlike Method 2 and 3, the implementation of Method 1 can be done in SAS, and thus, does not introduce an additional burden from using a different programming language like R that was used for the other methods. In addition, using SAS for the A&L imputation facilitates the data flow in production because files do not need to be migrated from one platform (SAS) to another (R), increasing production time and introducing potential errors.

On the other hand, we suspect that even if Method 1 is to be implemented for imputing the A&L variables, it may not be wise to attempt to recycle the programs for the CE income imputation for this purpose. One consideration is that many changes have occurred in SAS since the income imputation was first implemented in the CE. The more up-to-date versions of SAS have more flexibility in programming, output generation, and execution speed. Another consideration is that income imputation and A&L imputation are both complex undertakings, and it may be best to begin anew on the A&L programming.

In addition, the implementation of any method might benefit by developing more documentation of the process with the relevant diagnostics to monitor the imputation process. The output documents the steps of the process, such as the variable being imputed, number of model cases, number of excluded cases, initial and final models, and the conditions for overfitting. Because of the large number of A&L variables to impute, the production implementation of the method should be flexible to impute in a stepwise fashion to save time as this method does not require imputing all variables simultaneously.

We also suggest undertaking additional research on domain estimates. In particular, methods for reducing the bias in key domain estimates might be achieved by retaining the important domain indicators in the predictors for the imputation models. One way of thinking about this is that the method currently focuses solely on identifying predictors and models for estimates for the total population, and this could be expanded to consider important domains. This topic has not been fully addressed in the literature.

We now address some issues related to the imputation that cannot be evaluated in our research. The first issue is the target population. As described in Chapter 1, the data collection protocol excludes those CUs who held the A&L account in the previous year but did not hold it at the interview time. Depending on the type of A&L account, this undercoverage can lead to biased estimates that may impact the validity of published estimates. Addressing this problem requires reviewing the data collection protocol to determine if this undercoverage can be mitigated by modifying the instrument. This is related to how the zero-balance accounts are collected. Currently, zero accounts are identified for the previous year when the CU provides an account value of zero, prompting a question to identify if the account is held or not. However, no such verification is implemented for the current year accounts. Furthermore, zero value accounts cannot be identified through the prompted range values after the CU refuses because none of the A&L account ranges include $0 as the lowest value. Again, these issues are beyond the scope of the feasibility of imputation of the A&L missing data but are ones that we encountered in our research.

Our simulation provided a form of internal validation of the imputation process in the sense that the simulation population and the sample design, and the missing data patterns for the simulation were derived from the CE data itself. A different external evaluation of the A&L estimates would be to compare the estimates to other sources. This may not be possible but is worth considering.

Finally, we discuss some aspects of the implementation of the imputation methods. Following the guidelines from income imputation, the process requires the use of 20 quarters. Although currently there is not enough A&L data to support this (the current data collection began in 2017), the imputation could be implemented and evaluated in production even though all 20 quarters are not yet available. For example, the data might be used for evaluation only or could be released depending on BLS statistical guidelines. Even if the data are not released, this implementation in a production environment would help by accumulating additional data for the donor models and fine-tuning any problems with the imputation during production.

References

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16*(1), 3-14. doi:10.1016/0304-4076(81)90071-3

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174.

Gilks, W. R. (1996). Full conditional distributions. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (pp. 75-88). London: Chapman & Hall.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 19*(22), 473-489. doi:10.2307/2291635

Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician, 58*(4), 298-302.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16*(3), 219-242.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press LLC.

van Buuren, S., & Groothuis-Oudshoorn, C. G. (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual*. Leiden: Technical Report PG/VGZ/00.038, TNO Prevention and Health.

van Buuren, S., & Groothuis-Oudshoorn, C. G. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1-67.

van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064.

Wolter, K. (2017). *Introduction to variance estimation* (2nd ed.). New York: Springer-Verlag.

Yang, S., & Kim, J. K. (2016). Fractional Imputation in Survey Sampling: A Comparative Review. *Statistical Science, 31*(3), 415-432.

[i] Numbers are rounded according to the US Census Bureau disclosure avoidance guidelines