Does Sequence of Imputed Variables Matter in Hot Deck Imputation for Large-Scale Complex Survey Data?

Amang Sukasih, Jean Wang, Peter Frechtel, and Karol Krotki
RTI International

Abstract
Hot deck is an imputation method for complex survey data, especially popular when many survey items need to be imputed. Items are frequently correlated in surveys; one goal of imputation is to preserve these relationships. When imputing many variables and deciding which should be imputed first, one can decide on the sequence in which the variables are imputed—based on order of appearance in the questionnaire (a screener question is imputed first before its follow-up questions) or based on rate of missing data (items with lowest rate would be imputed first, followed by items with higher rates). Iteratively cycling the imputation may address association among variables (once all variables with missing values are imputed, imputation is rerun with previously imputed values in the covariates being treated as reported values). This paper discusses results from investigating the sensitivity of final estimates to the sequence of imputed variables. We also measure the impact of factors such as missing data rates and number of levels in categorical variables and imputation cycles. We use empirical simulation and focus on bias reduction and preservation of variable relationships.

Keywords: Missing data, single imputation, massive imputation, weighted sequential hot deck, classification and regression trees, cyclical imputation, Residential Energy Consumption Survey (RECS)

1. BACKGROUND

Nonresponse is common in sample surveys. Nonresponse creates missing data classified as unit missingness and item missingness. Unit missingness occurs when the sampled unit does not respond to the survey, so all surveyed items for the unit are missing; item missingness occurs when the sampled unit responds to the survey but a particular item remains missing. In this paper, we focus only on the statistical treatment of item missingness, specifically by imputing valid values using a hot deck imputation method that imputes (fills in) a valid value using data from an item respondent in the same survey (donor) that has similar characteristics to the recipient. This imputation takes place within imputation cells that are created using variables correlated with the imputed variable.

Complex survey data usually consists of many variables with some interdependencies or correlation among them—for example, “parent-child” structure of survey items, where child items are asked of the respondent if the response to parent items indicates that this respondent is eligible for child/follow-up items. When imputing for missing values in these variables, these interdependencies should be considered. In hot deck imputation, we consider the correlations among variables in constructing the imputation cells. For example, for parent-child variables, when a child variable is imputed, usually a parent variable becomes one of the variables used to construct the imputation cells.

Selection of variables used for creating imputation cells is usually done through some kind of regression-based and classification-tree technique. Among popular methods are CHAID (Ault et al. 2003; Creel & Krotki 2006; Kass 1980), regression tree (Loh et al. 2018), and regular regression-based models such as logistic or linear multiple regressions where the imputed variable or indicator of response/nonresponse of the imputed variable (dependent variable) is modeled using a set of predictors/covariate (independent variables). To run a model for variable selection, sample cases that are used are those observations without missing values in any of these dependent and independent variables. Therefore, the sample size in running a model would depend on missing rates in these variables; the higher the missing rates, the smaller the number of cases used in fitting the model. When imputing one variable after another, variables that have been imputed in earlier steps can be used as predictors in the next steps. So, given a set of predictors with no missing values, to gain more power using more sample size, imputation order would start with imputing variables with the least missing rate and continue until the last imputed variables with the most missing rate. Alternatively, one can perform imputation simply by following the order of items in the questionnaire. In this paper we address choosing order of imputation in hot deck between the two options of sequence of imputing variables: (1) in order of their appearance in the questionnaire, or (2) in order of their missing rates from the smallest to the largest. We investigated these two options, compared, and reviewed the results using a simulation study.

2. METHODOLOGY AND DATA

2.1. Hot Deck Imputation

In the hot deck imputation method, missing values (in a recipient record) are replaced by reported data from a similar unit (a donor) in the same survey. The recipient and potential donors are grouped into cells (called *imputation cells*) based on variables (called *class variables*) strongly correlated with the variable being imputed, such that the recipient and the donor are expected to have the same characteristics. Imputation then takes place within the cell. Continuous variables need to be recoded into categories before they are used as class variables. Recipients and donors within an imputation cell may be sorted based on additional variables correlated with the variable being imputed (called *sorting variables*) when forming classes based on these variables is impractical. Additionally, the imputed variable may be used to sort donors to ensure that the (weighted) distributions of imputed values are like donor values.

Several variations of hot deck techniques are available for imputation. The primary differences between methods may be attributed to variation in forming imputation cells and the process by which donors are chosen within cells. For example, imputation cells may be formed by a complete cross of several variables, as in the regular Unweighted Sequential Hot Deck (USHD) and the Weighted Sequential Hot Deck (WSHD) imputation methods (Cox 1980; Iannacchione 1982; Little & Rubin 1987), via Proxy Pattern-mixture Models (Andridge & Thompson 2015), or based on classification or regression trees (Bigss, Ville, & Suen 1991; Breiman & Friedman 1993; Creel & Krotki 2006; Kass 1980). In selecting donors, both weighted and unweighted sequential hot deck methods sort cases on a set of additional (class) auxiliary variables. The USHD picks an adjacent case with a reported value as the donor for the missing value. The WSHD sorts donors and recipients separately but does something similar. By contrast, a method like random nearest neighbor hot deck chooses a donor randomly from a donor set or neighborhood deemed “close to” the recipient, with respect to several additional covariates.

2.2. Cyclical Tree-Based Hot Deck (CTBHD)

RTI developed the CTBHD imputation system to provide a system that can handle massive imputation in large-scale survey data with many variables and complex interdependencies among the survey items. It was developed using Microsoft Excel for user interface, statistical software SAS for data management and processing, *R* for imputation cell construction, and SUDAAN for hot deck imputation. It is complex programing code, designed to minimize user intervention during the imputation process, freeing up the user to focus more on data preparation and results.

CTBHD consists of two steps: formation of imputation cells and donor selection and imputation within each cell. A special feature added in CTBHD imputation is to repeat (cycle) the whole imputation process several times and retain the imputed values from the final cycle as the final imputed data. The process is explained in more detail next.

Imputation Cell Formation

In constructing imputation cells, the CTBHD implements the "**`tree`**" package (Ripley 2016) in *R* to create mutually exclusive imputation cells for each item value requiring imputation (prediction) using records with no missing values for both the item needing imputation (the left-hand variable) and potential predictors (the right-hand variables). The program employs a standard classification-and-regression-tree methodology to partition data and create cells (homogeneous subsets) based on the values of the predictor items. For the item needing imputation (the left-hand variable), a set of potential covariates or predictors in the model (the right-hand variables) are determined by the user. Not all covariates from the right hand of the model are used to grow a tree; only variables that correlate strongly with the variable being imputed are used.

A tree starts with a root node and is then grown from top to bottom by binary recursive partitioning using the response in the specified model. The root is split first based on the most important significant variable. At each internal node in the tree, this method applies a test to split the data. For categorical variables, the levels of an unordered factor are divided into two non-empty groups. There could be many possible splits, and the split that produces homogeneous cases within the group and maximizes the reduction in misclassification rate is chosen, the data are split, and the process is repeated. Numeric variables were divided based on a cutoff value a: $X < a$ and $X \geq a$. Ordered categorical variables were treated like continuous variables.

The tree stops creating cells when the current node is smaller than the user-provided value "**`minsize`**" or when one of two nodes that would be created by a split of the current node is smaller than the user-provided value "**`mincut`**." Usually **`minsize`** = 2 × **`mincut`**. The terminal nodes become the imputation cell.

The CTBHD program requires that all items needing imputation and all predictor items be of numeric type. Moreover, each must be designated as either categorical or continuous. Ordinal variables can be treated either way. No categorical variable can have more than 32 levels. Because a classification tree involves a search over $\left(2^{k-1} - 1\right)$ groupings for k levels, tree growth is limited to a depth of 31. Moreover, a categorical item with too many levels can cause the program to fail. This will happen when a record needing imputation for an item has a categorical level for that item that no potential donor has.

The use of regression trees to form imputation cells has an advantage over the traditional complete cross-classification of covariates. In the regression tree, the nodes/cells are formed only when the variable levels/categories are statistically significant, whereas in the complete cross-classification, all combinations of levels are used to form potential cells, some of which may have a small number of cases. In such a case, cells may be collapsed together, either in an ad hoc or subjective manner or by prespecified rules. In a regression tree, the splits are done based on formal statistical tests. Note, however, that the "**`tree`**" program does not use weights when forming imputation cells.

Donor Selection

After imputation cells are constructed, the CTBHD implements the WSHD to select donors. The WSHD uses weights to "match" chosen donors to recipients in a cell. Each group is first sorted in some manner, then a weighted sequential sampling routine matches donors to recipients that makes the weighted means of the two groups (with respect to any item value) equal in expectation. Within an imputation cell, the cases with missing values and the cases with reported values can be thought of as two separate files (one of nonrespondents and the other of respondents). Then, the weight of the case needing imputation, $w(j)$, is adjusted as follows:

$$v(j) = w(j) \times \frac{s(+)}{w(+)}$$

where

$$w(+) = \sum_{j=1}^{n} w(j)$$

$$s(+) = \sum_{i=1}^{r} s(i)$$

n = number of cases with missing value (recipients),
r = number of donors,
$w(j)$ = weight attached to the j-th recipient within the cell, $j = 1, \cdots, n$,
$s(i)$ = weight attached to the i-th donor within the cell, $i = 1, \cdots, r$.

Now the sum of adjusted weights across cases with missing values is equal to the sum of donor weights. When missing values and reported values (donors) are expanded by their weights, we view this as two matched files, and we partition the values into zones with width $v(j), j = 1, \cdots, n$. The WSHD imputation algorithm then finds a donor for the jth missing value from the reported value in the corresponding zone.

The WSHD is implemented using the SUDAAN procedure "IMPUTE" (Research Triangle Institute 2012). By ordering the imputation of items, items with previously imputed values can be used as predictors in subsequent imputations. A hot deck imputation process is finished when all variables with missing values have been imputed. In RTI's CTBHD systems, however, the complete imputation cell construction and donor selection process are repeated ("cycled") any number of times, with all items having imputed values now allowed to be used as predictors.

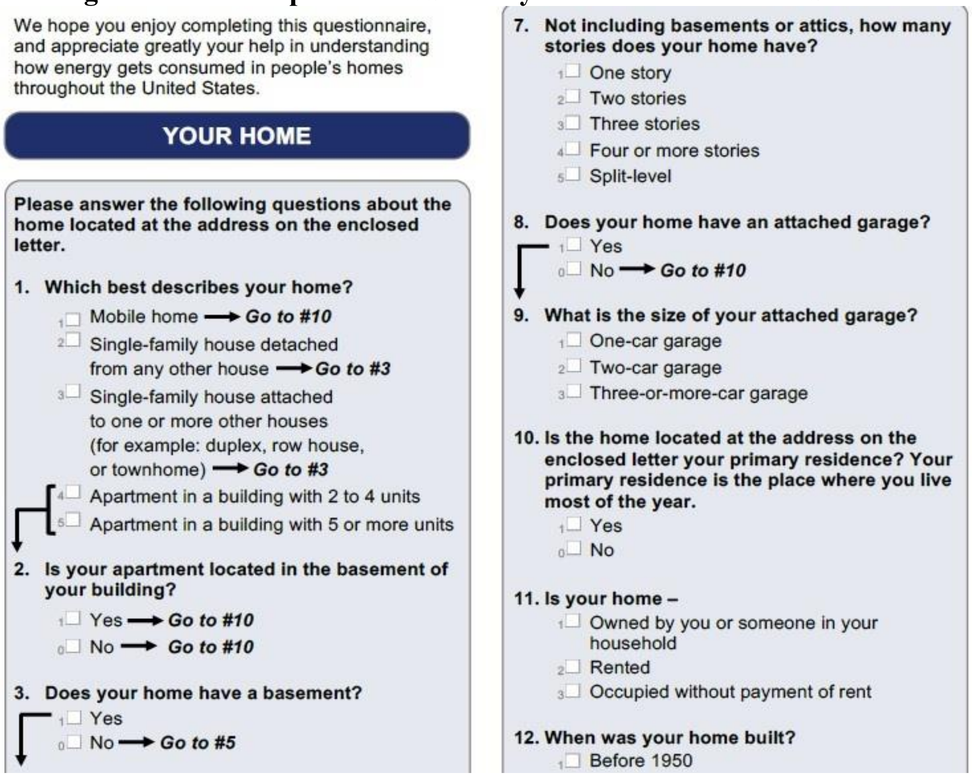
Cycling offers the advantage that the imputation cells (regression tree nodes) after the first cycle are developed based on all cases instead of just the subset of complete cases. In Martin et al. (2017), cycling showed a clear advantage for continuous and polytomous categorical variables relative to dichotomous variables. Using data from the Residential Energy Consumption Survey (RECS), Martin et al. (2017) showed the effects of cycling the imputation on 31 variables of varying data types. They conclude that simple binary variables do not benefit as much from cycling, but a small number of cycles (three or fewer) may be used to assure convergence. The paper recommends that cycling methodology with three to five cycles be used as a standard procedure in the CTBHD imputation process, and testing for convergence should be done after five cycles. In practice, depending on the data it may be enough to cycle three times. If no convergence occurs after five cycles, it is possible to run for further cycles. For results that do not converge after 10 cycles, however, it may be necessary for an analyst to investigate the characteristics of the data set under investigation.

2.3. RECS Data

In this paper we demonstrate our application of CTBHD to the 2015 RECS public use file (PUF). RECS is sponsored by the Energy Information Agency of the U.S. Department of Energy. It collects data on household energy usage, demographics, and home characteristics from a nationally representative sample of housing units. The data are combined with data from energy suppliers to these homes to estimate energy costs and usage for heating, cooling, appliances, and other home uses. The 2015 RECS is the 14th collection in the series since 1978. The data collected from the survey support evaluations of trends in home energy use and projections of future energy needs through estimates for the nine census divisions, four census regions, and the country as a whole.

The 2015 RECS is a mixed-mode data collection mostly using CAPI and web survey and in-person visits for some variables; for example, in addition to obtaining questionnaire data, field interviewers measure the square footage of the selected homes. As in many large federal sample surveys, RECS collects many variables that correlate with each other. An example of an obvious relationship is among the parent and child (or sometime called “gate and follow-up”) survey items. Figure 1 provides an example of a section (“Your Home”) in the RECS questionnaire that has parent-child survey items. For example, question #8 (“Does your home have an attached garage?”) is the parent survey item for question #9 (the child item: “What is the size of your attached garage?”). This relationship creates a logical skip pattern in the dataset. That is, if a respondent answers No to question #8, then question #9 will be skipped by this respondent.

Figure 1. An example of RECS survey items: “Your Home” Section[1]



We hope you enjoy completing this questionnaire, and appreciate greatly your help in understanding how energy gets consumed in people’s homes throughout the United States.

YOUR HOME

Please answer the following questions about the home located at the address on the enclosed letter.

1. Which best describes your home?

1☐ Mobile home → ***Go to #10***
2☐ Single-family house detached from any other house → ***Go to #3***
3☐ Single-family house attached to one or more other houses (for example: duplex, row house, or townhome) → ***Go to #3***
4☐ Apartment in a building with 2 to 4 units
5☐ Apartment in a building with 5 or more units

2. Is your apartment located in the basement of your building?

1☐ Yes → ***Go to #10***
0☐ No → ***Go to #10***

3. Does your home have a basement?

1☐ Yes
0☐ No → ***Go to #5***

7. Not including basements or attics, how many stories does your home have?

1☐ One story
2☐ Two stories
3☐ Three stories
4☐ Four or more stories
5☐ Split-level

8. Does your home have an attached garage?

1☐ Yes
0☐ No → ***Go to #10***

9. What is the size of your attached garage?

1☐ One-car garage
2☐ Two-car garage
3☐ Three-or-more-car garage

10. Is the home located at the address on the enclosed letter your primary residence? Your primary residence is the place where you live most of the year.

1☐ Yes
0☐ No

11. Is your home –

1☐ Owned by you or someone in your household
2☐ Rented
3☐ Occupied without payment of rent

12. When was your home built?

1☐ Before 1950

[1] Extracted from https://www.eia.gov/survey/form/eia_457/2015_EIA-475A_paper.pdf

In the imputation process, we distinguish between missing value because of logical skip (ineligible respondent) and because of nonresponse (eligible respondent). When the respondent's answer to a parent item indicates that he or she is eligible to answer the child item but the value response to the child item is blank/missing then the child item should be imputed. On the other hand, if the child item is reported (nonmissing) but the parent item is missing, then the parent item may be simply edited to a value that indicates the respondent is eligible for the child item. This relationship of survey items suggests that the parent item should be used as one of class variables to group the samples into imputation cells. Note that the sequence of imputation order may be done by imputing the parent item first (with child item being used as one of covariates) then imputing the child item (with the imputed parent item used as covariate). This order follows the order of items in the questionnaire. The imputation under this order should result in a consistent parent-child item response pattern. However, when the parent item has a higher missing rate, the regression tree will use fewer cases in selecting class variables. Depending on the missing rates in the imputed variable and covariates for imputation cells, less sample size in modeling may result in imputation cells that do not fully group/match recipients and potential donors with similar characteristics. If this is the case, hot deck imputation may not fully address nonresponse bias.

Alternatively, to increase sample size in modeling, the order of imputation may be based on item missing rates where the item with the least missing rate is imputed first. This is to gain power through a larger sample size in tree modeling. However, care must be taken to ensure that parent-child items are imputed consistently.

The work in this paper investigated two options of imputation order to answer the key research question of whether the two options lead to similar imputation results. We used a simulation study to answer this question.

2.4. Simulation Technique

To ensure that the simulation is timely feasible, we chose only two variables to be imputed from the "Your Home" section of the 2015 RECS data: attached garage (variable name: PRKGPLC1, values: two levels) and size of the attached garage (variable name: SIZEOFGARAGE, values: three levels—see Figure 1). However, we used an additional 46 variables from Your Home and Household Characteristics modules as candidate variables for imputation cells. These include type of housing unit (TYPEHUQ), number of stories (STORIES), has swimming pool (SWIMPOOL), etc.

The 2015 RECS data in PUF have no missing values. The file comes with an imputation flag for each variable so that we can reconstruct the original missing values in the survey. For our simulation, we reconstructed the missing values only in PRKGPLC1 and SIZEOFGARAGE variables, while all 46 candidates for class variables are kept as is with no missing values (the imputed values were treated as if they were true reported values).

The original missing rates in PRKGPLC1 and SIZEOFGARAGE are 0.59% and 0.78%. These original missing rates are considered as very small missing rates. Therefore, for our simulation, we simulated higher missing rates (four other missing rates) as given in Table 1.

Table 1. Five missing rates scenario for simulation

Missing Pattern	Missing rate in PRKGPLC1	Missing rate in SIZEOFGARAGE
Scenario 1 (orig. RECS)	0.59%	0.78%
Scenario 2 (simulated)	2.25%	0.78%
Scenario 3 (Simulated)	2.84%	0.78%
Scenario 4 (simulated)	0.59%	7.82%
Scenario 5 (simulated)	6.17%	1.82%

In addition to using missing patterns in the original RECS data (where missing rate in the parent variable is smaller than that in child variable—here Scenario 1), we simulated several additional missing at random (MAR) patterns/scenarios as follows:

a. missing rate in the parent variable is larger than that in child variable (Scenarios 2, 3, 5);
b. imputation of unedited data; that is, missing values that could be edited were not edited but instead were passed into imputation (Scenarios 3, 5); and
c. varying missing rates (smaller in Scenarios 1, 2, 3, larger in Scenarios 4, 5).

For each data scenario, CTBHD imputation went through three cycles, in which the imputed values from the third cycle are considered final. For each data scenario, we also altered the order of imputation, so that we have a total of 10 imputation scenarios (5 missing patterns by 2 imputation orders). Then each imputation scenario was replicated 100 times. The distribution of imputed values across 100 replicates is compared across 10 scenarios of imputations.

3. RESULTS

For each of 100 simulated imputed datasets under five scenarios and two imputation orders, the proportion of own garage (PRKGPLC1 = 1), proportion of 1-car size garage (SIZEOFGARAGE = 1), and proportion of 2-car size garage (SIZEOFGARAGE = 2) are calculated. Distributions of these proportions are plotted and compared against the proportions calculated from the PUF data without missing values (being treated as the true values). Figure 2 presents three pictures of distributions corresponding to the proportion of own garage, proportion of one-car size garage, and proportion of two-car size garage. For each picture, 10 distributions corresponding to 10 simulated imputed data are overlaid on the same axes, where each plot represents distribution of proportions across 100 simulation replicates.

Figure 2. Simulation results: final (simulated) imputed data

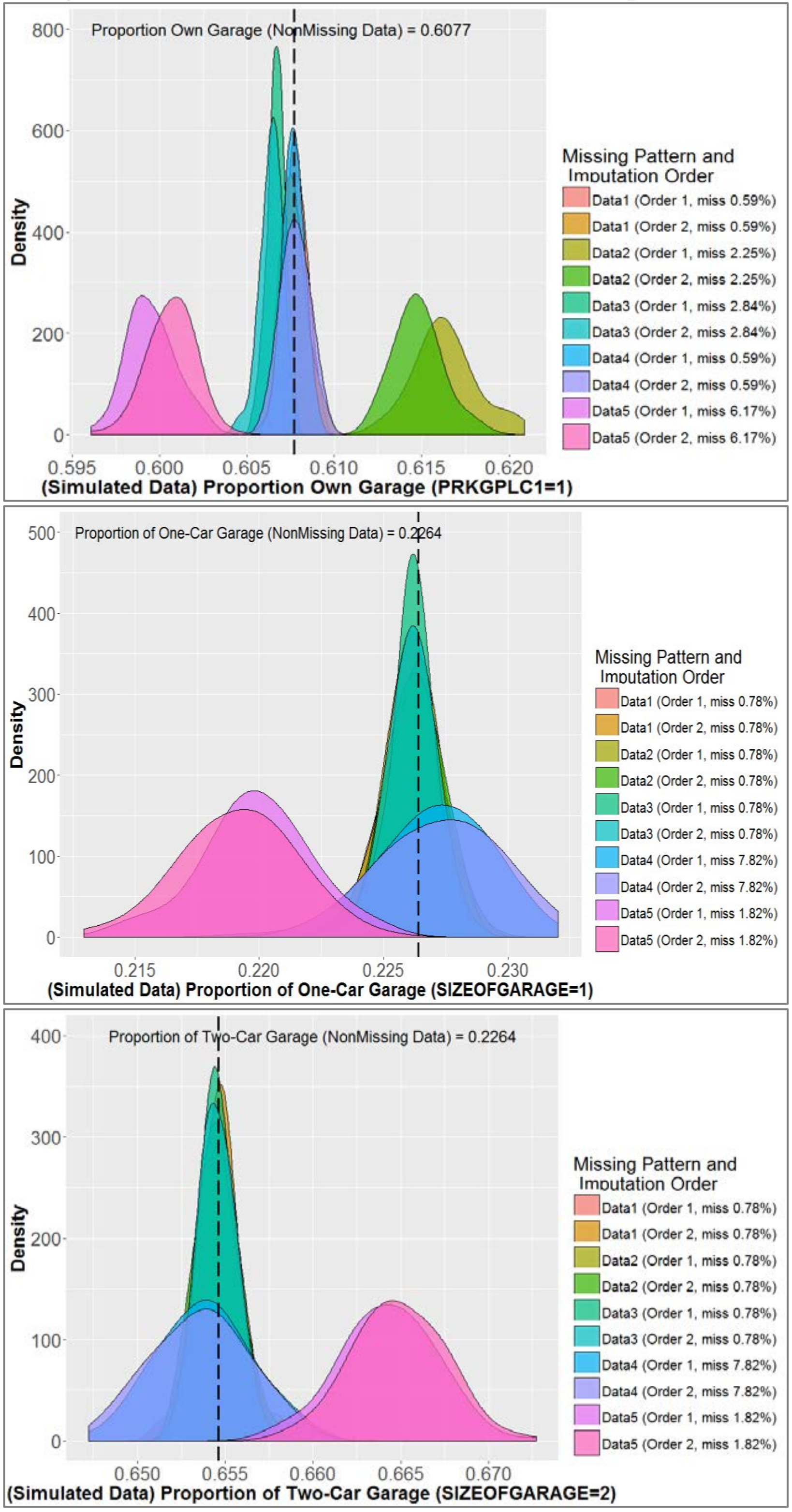


Table 2. Simulation results: means and standard deviations of proportions by data scenario and imputation order

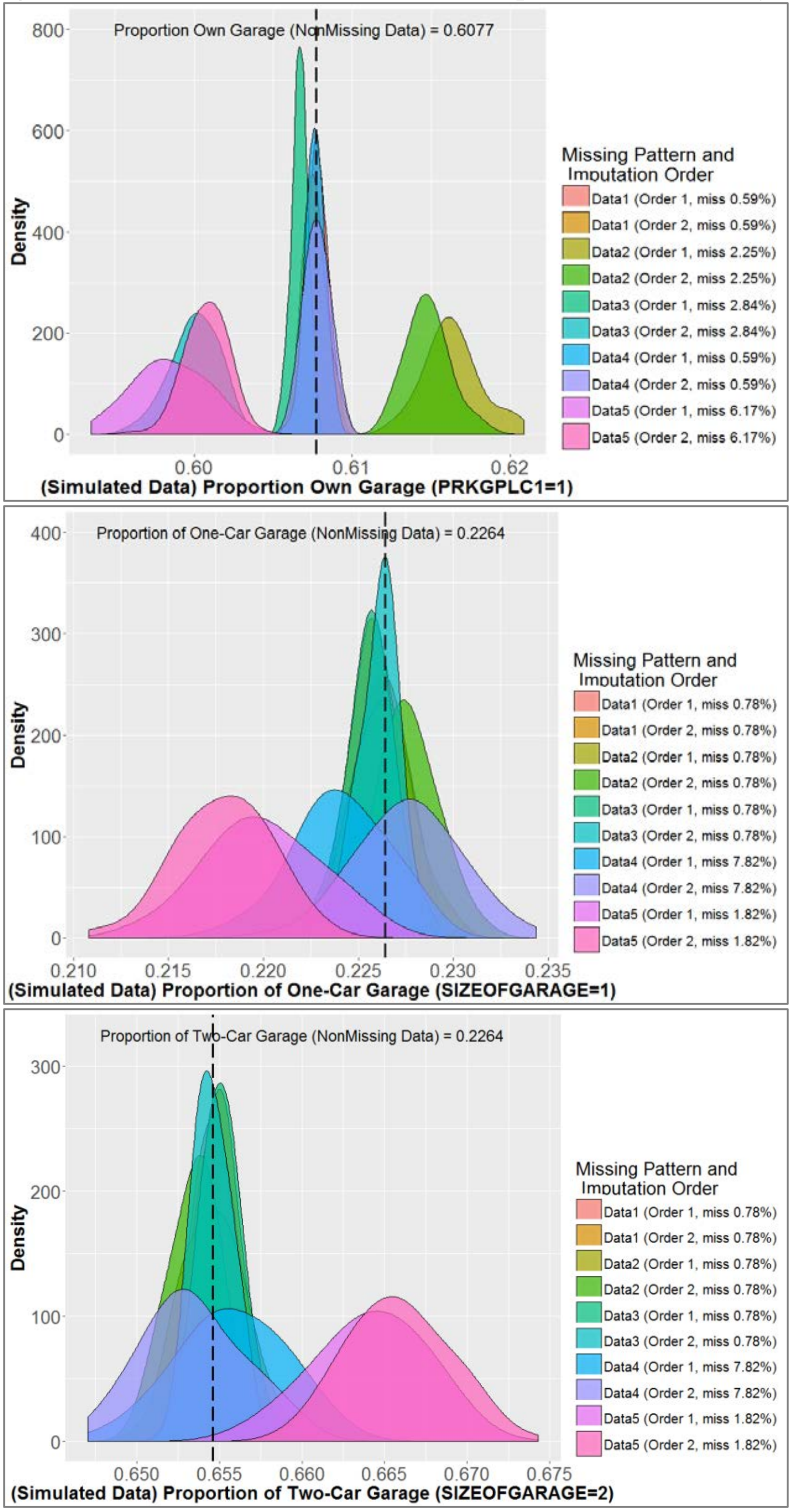
		Mean of proportions			Standard deviation of proportions		
Data scenario	**Order of imputation**	**Own garage**	**1-car size garage**	**2-car size garage**	**Own garage**	**1-car size garage**	**2-car size garage**
1	Response rate	0.6077	0.2261	0.6546	0.0007	0.0010	0.0010
	Questionnaire	0.6078	0.2262	0.6546	0.0007	0.0012	0.0013
2	Response rate	0.6148	0.2262	0.6545	0.0014	0.0011	0.0012
	Questionnaire	0.6164	0.2262	0.6546	0.0018	0.0011	0.0012
3	Response rate	0.6067	0.2261	0.6546	0.0005	0.0010	0.0010
	Questionnaire	0.6064	0.2261	0.6546	0.0006	0.0010	0.0010
4	Response rate	0.6077	0.2270	0.6538	0.0006	0.0022	0.0026
	Questionnaire	0.6078	0.2273	0.6536	0.0008	0.0023	0.0027
5	Response rate	0.6007	0.2191	0.6649	0.0013	0.0022	0.0025
	Questionnaire	0.5995	0.2199	0.6643	0.0014	0.0022	0.0027

Under different missing rates (different simulated data), CTBHD resulted in different imputed values reflected in differences in the values of proportions (see Table 2 and Figure 2). In each picture, we see distributions of proportions having different means and variances that can be grouped roughly into three groups. The “true” proportions of own garage, one-car size garage, and two-car size garage are, respectively, 0.6077, 0.2259, and 0.6546, indicated by a vertical line in each picture. The largest difference between these true values for proportions of own garage, one-car size garage, and two-car size garage and (any) estimate of proportion from 10 simulated imputed data are, respectively, 0.0132, 0.0130, and 0.0181. Therefore, we see that there is still a potential nonresponse bias in the imputed data; however, this bias is extremely small. We discuss this issue in the summary section.

Now our focus is on the comparison between imputed data under two imputation order scenarios; that is, to compare the results between (Data 1, Order 1) vs. (Data 1, Order 2), (Data 2, Order 1) vs. (Data 2, Order 2), …, and so on for all five data scenarios. Within each scenario, different order of imputation (Order 1 vs. Order 2) resulted in very minor differences in proportion estimates. When the missing rates are moderate (Data 2) to high (Data 5), the differences are more pronounced than in scenarios with smaller missing rates. Nevertheless, differences in the proportions of interest under the two imputation Orders 1 and 2 are very small and ignorable (less than 0.01 for PRKGPLC1 and less than 0.013 for SIZEOFGARAGE).

We also reviewed results of imputed data that were imputed only using base cycle; in other words, the imputation process was only done once without cycling the imputation. Figure 3 present the results of simulated imputed data after base cycle.

Figure 3. Simulation results: (simulated) imputed data at base cycle



We can see that the gaps of distributions of proportions between Order 1 vs. Order 2 are wider. The differences in the imputed values because of different orders of imputation are more pronounced in the data imputed only at base cycle than that in the imputed data after last cycle (imputing with three cycles). This result reinforces the conclusion that cycling the imputation has a stabilizing influence on the imputed values.

4. CONCLUSION AND DISCUSSION

In our simulation with only one pair of parent-child variables, the order of imputed variables in CTBHD has very little impact on the imputation result. We realize that this conclusion is based on a very limited number of imputed variables. In our next simulation study, we will include more intercorrelated variables, probably a full set of survey items from a section/module of the questionnaire.

In practice, it is preferable to order the imputed variables based on the order of their appearance in the questionnaire because the flow and logic of the data resemble the flow of an interview (or in a self-administered survey the flow of what the respondent saw in the questionnaire). The imputed values should also result in consistent values with regard to parent-children items relationship. In addition, calculation of response rates may not be needed prior to imputation. With this order of imputation that ignores the missing rates, we stressed the important of cycling the hot deck imputation to ensure that the imputation would reduce potential nonresponse bias, especially when the missing rates are high for survey items that appear earlier in the questionnaire.

We see potential small bias of imputed data when response rate is high. However, we have not investigated whether this small bias is the result of missing rate or other factors such as missing pattern that deviates from MAR. This investigation will be part of our continuing investigation.

REFERENCES

Andridge, R. R., & Thompson, J. K. (2015). Using the fraction of missing information to identify auxiliary variables for imputation procedures via proxy pattern-mixture models. *International Statistical Review*, *83*, 472-492.

Ault, K., Black, S., Chromy, J., Fahimi, M., Siegel, P., Trofimovich, L., Whitmore, R., & Berkner, L. (2003). *Imputation methodology for the National Postsecondary Student Aid Study: 2004*. NCES 2003–20. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Bigss, D., Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, *18*, 1, 49-62.

Breiman, L., & Friedman, J. H. (1993). *Classification and regression trees*. New York: Chapman & Hall.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 721–726.

Creel, D. V., & Krotki, K. (2006). Creating imputation classes using classification tree methodology. In *Proceedings of the Survey Research Methods Section, American Statistical Association, Joint Statistical Meeting 2006,* pp. 2884–2887.

Iannacchione, V.G. (1982). Weighted Sequential Hot Deck Imputation Macros. *Seventh Annual SAS User's Group International Conference*, San Francisco CA, February 1982.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *20*, 2, 119-127.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Loh, W. Y., Eltinge, J., Cho, M. J., & Li, Y. (2018). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*: Preprint doi:10.5705/ss.202017.0225. Available at http://www3.stat.sinica.edu.tw/preprint/SS-2017-0225_Preprint.pdf

Martin P., Wang, J., Frechtel, P., Sukasih, A., Lewis, K., Deng, G., & Kinyon, D. (2017). *Tree-based hot deck imputation cycling—Does cycling help?* Poster presentation, Joint Statistical Meeting 2017.

Research Triangle Institute. (2012). *SUDAAN language manual,* volumes 1 and 2, release 11. Research Triangle Park, NC: Research Triangle Institute.

Ripley, B. (2016, January 21). *Package 'tree': Classification and regression trees.* Available at https://cran.r-project.org/web/packages/tree/tree.pdf