Hot-Deck Imputation Cells for the American Housing Survey

Chrystine Tadler[1], Richard Griffiths[1]
[1]Insight Policy Research, 1901 N. Moore Street, Suite 1100, Arlington, VA 22209

Abstract
One technique the American Housing Survey uses to address missing data is hot-deck imputation. This methodology has been largely unchanged since 1998, with limited documentation surrounding the formation of imputation cells. A review of the hot-deck imputation process and assessment of various methods of creating imputation cells has led to an alternate hot-deck imputation methodology, with an improvement in preserving associations between dependent and independent variables. Techniques including stepwise logistic regression and classification and regression trees identify variables highly correlated with those being imputed, forming the basis for improved imputation cells. Classification and regression trees are also used to determine improved sort order and grouping during production.

Key Words: American Housing Survey, hot-deck imputation, stepwise logistic regression

1. Introduction

The American Housing Survey (AHS) is a longitudinal survey conducted over a nationally representative sample of housing units that provides valuable information about the nation's current housing stock. AHS experiences missing data for survey respondents as a result of item nonresponse, which is addressed by various imputation techniques. Several variables across topic-specific modules rely on the hot-deck imputation method, a methodology that has not changed since 1998. Documentation describing the construction of imputation cells is limited with little justification for variable selection for imputation matrices. After reviewing the current imputation methodology for AHS, this paper evaluates and proposes an alternative to the current methodology. It identifies new variables to form imputation cells and revised imputation sort order and grouping.

The research for this paper extended only to survey items using hot-deck imputation and did not include variables imputed using cold-deck, model-based, or deductive imputation. Prior to imputation, all survey respondent records are sorted by a geographical indicator. This step ensures donor and recipient records are geographically near one another to control for variation related to physical location.

For each imputed variable under the current methodology, donor and recipient records are sorted into imputation cells based on values for variables that have an expected, but not verified, correlation with the imputed variable. After sorting within imputation cell, recipient records are filled with data from the most recent donor record as the imputation procedure sequentially cycles through the data. Table 1 provides an example of imputation cell construction used to impute variables within the housing unit module. See Barger (2016) for more detail on the current imputation processes used in the AHS.

Table 1: Imputation Cells for Housing Unit Module

Number of floors	Tenure status	Imputation cell number
One	Owner occupied or vacant for sale or off market	1
	Renter occupied or vacant for rent	2
Two to three	Owner occupied or vacant for sale or off market	3
	Renter occupied or vacant for rent	4
Four	Owner occupied	5
	Rented or vacant or usual residence elsewhere	6
Five or more	All housing units	7
Unknown	Owner occupied or vacant for sale or off market	8
	Renter occupied or vacant for rent	9

Source: Barger (2016)

2. Proposed Alternate Imputation Cells

This research maintains the fundamental methodology for hot-deck imputation used in the AHS while focusing on the selection of variables used to create imputation cells. Imputation cells are still formed using correlated independent variables, and imputed variables are grouped into modules based on similar imputation cell structure for processing. The proposed alternate methodology focuses on selecting more highly correlated variables to create imputation cells for each imputed variable along with a corresponding alternate modular grouping for processing.

To determine an alternate imputation cell construction, a measure of association was calculated between each imputed variable and potential variables to form imputation cells. Here, association is defined as a measure of the strength of the relationship between two variables, and correlation is a specific type of association. Correlation is defined as the measure of linear association between two variables.

2.1 Imputed and Independent Variables

Imputed variables included in this research extended only to those imputed through hot-deck imputation. Person-level imputed variables and continuous variables were not evaluated because of time constraints.

Potential independent variables for imputation cells came from a version of the household-level and person-level 2017 AHS datafiles prior to any imputation or edits. This file contained more than 600 variables that would be considered as potential variables for imputation cell construction. The intention was to consider all AHS variables equally for optimal imputation cell construction. However, for efficiency and successful execution of analyses, this list of variables was limited based on guidance from Census Bureau staff. For example, administrative variables and variables with large rates of missing information were eliminated. After elimination and additions based on the above criteria, the list of potential independent variables included slightly more than 200 variables.

2.2 Analysis

Three research questions framed this research. The first related to the identification of three to five variables most highly associated with each imputed variable. It was hypothesized that using more highly associated variables to construct imputation cells would result in more precise imputations. Three separate analyses were run for each imputed variable, and results were compared to assess similar findings for the top five most associated variables. The analyses included a preliminary logistic regression, stepwise logistic regression, and classification and regression trees (CART).

2.2.1 Preliminary logistic regression

A preliminary logistic regression was run for each unique pairing of dependent variables with independent variables. Values of "don't know" and "refused" were not included in analyses for the dependent and independent variables. The generalized logit function was used for dependent nominal variables, while the cumulative logit function was used for dependent ordinal variables. This research did not extend to continuous dependent variables.

Using SAS 9.4 and PROC LOGISTIC, the R^2 deviance (DevRsq) was calculated after adding the predictor variable to the model:

$$DevRsq = \frac{L_i - L_c}{L_i}$$

where L_i is the log likelihood value of the model with intercept only, and L_c is the log likelihood value of the model with intercept and covariates (after the predictor variable is entered into the model).

This measure was calculated for each unique pairing of imputed and independent variables. As a result, independent variables could be sorted according to their R^2 deviance, considered as the measure of association with the imputed variable. A list of the most highly correlated variables to each imputed variable was obtained. As an example, table 2 displays an ordered list of the top 15 most associated variables with the imputed variable indicating the presence of a basement within a unit from the preliminary logistic regression analyses. The five most associated variables to the imputed variable were the number of floors in the unit, the type of fuel used most for heating the unit, the year the unit was built, the unit's entrance requiring stairs, and the main type of heating equipment.

Table 2: Ordered Association for Imputed Variable Indicating the Presence of a Basement in Unit Using Preliminary Logistic Regression

Independent variable	L_i	L_c	$DevRsq = \frac{L_i - L_c}{L_i}$
Number of floors in unit	154,588.8008	124,109.0143	0.1972
Fuel used most to heat unit	153,194.9390	145,105.8891	0.0528
Year unit was built	151,866.6937	145,192.5824	0.0439
Unit's entrance requires stairs	154,560.9949	148,050.4215	0.0421
Main type of heating equipment	154,033.1733	149,062.4023	0.0323
Country of birth of householder	108,962.3915	106,881.0019	0.0191
Evidence of roaches in unit	153,371.1296	150,470.5022	0.0189
Fuel used most to heat water	151,963.5688	149,092.4822	0.0189
Spanish origin of householder	108,962.3915	107,269.2593	0.0155
Year householder moved in	108,962.3915	107,332.0977	0.0150

Unit square footage	126,240.9499	124,434.9841	0.0143
Number of half bathrooms in unit	154,385.0639	152,227.0584	0.0140
Household uses gas	152,754.8699	150,648.5820	0.0138
Presence of central air conditioner	154,131.5745	152,135.7111	0.0129
Number of full bathrooms in unit	154,449.9714	152,518.9378	0.0125

2.2.2 Stepwise logistic regression

In the second analyses, PROC LOGISTIC was run with the stepwise option for each imputed variable. The top 15 most associated variables based on the $DevRsq$ values from the preliminary logistic regression were entered as possible predictor variables. Again, the generalized logit function was used for dependent nominal variables, while the cumulative logit function was used for dependent ordinal variables. Continuous dependent variables were not run through analyses.

PROC LOGISTIC was terminated after five variables were entered,[1] and the R^2 deviance was calculated after each step to measure the improvement in the model after the addition of each predictor. The stepwise model built could have terminated earlier with fewer than five independent variables if

$$DevRsq_n = \frac{L_{n-1} - L_n}{L_{n-1}}$$

where L_n is the log likelihood value of the model after the n^{th}step.

The stepwise logistic regression iteratively tested combinations of the independent variables against entry and exit criteria to provide a list of the five best independent variables for prediction of the imputed variable. For purposes of addressing the research question, the independent variables were ranked by the step at which they were entered into the model, not by their $DevRsq_n$ values. Table 3 shows the associations for the top five variables associated with the imputed variable indicating the presence of a basement within a unit using stepwise logistic regression.

Table 3: Ordered Association for Imputed Variable Indicating the Presence of a Basement in Unit Using Stepwise Logistic Regression

Step	L_n	$DevRsq_n = \frac{L_{n-1} - L_n}{L_{n-1}}$
Intercept	86,385.582	--
Number of floors in unit	70,655.125	0.1821
Number of half bathrooms in unit	70,409.107	0.0035
Unit's entrance requires stairs	67,361.376	0.0433
Country of birth of householder	66,656.307	0.0105
Year unit was built	64,181.841	0.0371

2.2.3 CART

PROC HPSPLIT was run for each imputed variable using the same top 15 possible predictors from the preliminary logistic regression analysis. The SAS output provided the importance and relative importance of each predictor used in the final model. Variable

[1] In PROC LOGISTIC the selection method was set to STEPWISE with an entry and exit significance level for the chi-square score as 0.05. It was possible that model building was terminated prior to identifying five independent variables if the entry criteria were not met or if a variable was removed from the model based on results of the Wald chi-square test.

importance was used as an indication of which independent variables are most useful in predicting outcomes for the imputed variable.

Variable importance is based on the residual sum of squares (RSS) as—

$$Importance = \sqrt{\sum_{d=1}^{D}\left(RSS_d - \sum_i RSS_i^d\right)}$$

where *d* represents the node, *D* is the total number of nodes, and *i* is the index of the leaf for node *d*.

The five variables with the highest relative importance values were identified and ordered. Table 4 provides the relative importance for each of the top five variables for the imputed variable indicating the presence of a basement within a unit.

Table 4: Ordered Relative Importance for Imputed Variable CELLAR Using CART

Step	*Importance*	*Relative importance*
Number of floors in unit	63.6619	1.0000
Unit's entrance requires stairs	31.9633	0.5021
Year unit was built	31.2000	0.4901
Presence of central air conditioner	16.3363	0.2566
Fuel used most to heat unit	13.9177	0.2186

3. Grouping

The second research question focused on the organization and ordering of imputation during processing. Current methodology groups imputed variables into topic-specific modules and imputes variables in a specific order to allow for some variables to be used to form imputation cells during the imputation of other variables. Based on the three to five most associated variables with each imputed variable, this second research question aimed to create new modules by identifying clusters of imputed variables with similar highly associated variables that enhanced the imputation process.

A clustering analysis was performed using the ranks of independent variables related to R^2 deviance from the stepwise logistic regressions. This hierarchical clustering analysis used the average linkage distance measure,

$$m_{ik} = \sqrt{\sum_{j=1}^{P}\left(d_{ij} - d_{kj}\right)^2}$$

where d_{ij} was the R^2 deviance measure between Y_i and X_j (Y_i being the i^{th} imputed variable and X_j being the j^{th} independent variable).

For each imputation variable, ranks were assigned to the independent variables in descending order of R^2 deviance. In other words, the predictor variable with the largest R^2 deviance (most highly associated) for a given imputation variable was assigned a rank

of 1; the predictor variable with the second largest R^2 deviance was assigned a rank of 2; and so on, up to rank 5. All variables not ranked in the top 5 were assigned a rank of 6 for the analysis.

Using the clustering analysis, each of the imputation variables were assigned to 1 of 10 revised clusters. These clusterings were compared with the current AHS clusterings (modules) as provided in Barger (2016). Table 5 shows the interrelationships between the current AHS clustering and the revised clustering. Variables in the current AHS clustering were often split into several separate clusters in the revised methodology. As the largest of the revised clusters, Revised Cluster 1 (RC1) contained variables originally in each of the current AHS clusters.

Table 5: Interrelationship Between the Current AHS Clustering and the Revised Clustering

Current AHS cluster	*New revised cluster*
BC1	RC1
BC2	RC1
BC4	RC1, RC8, RC5
BC5	RC1, RC2, RC3, RC5, RC7
BC6	RC1, RC2, RC5, RC6, RC9, RC10
BC9	RC1
Cold deck/deductive Imputation	RC4

Table 6 shows the interrelationships between the revised clustering and the current AHS clustering. Variables currently spread across all the current AHS clusters were grouped together into one revised cluster (RC1). Many variables were found to have little in common as related to predictor variables and were therefore separated into new, smaller revised clusters.

Table 6: Interrelationship Between the Revised Clustering and the Current AHS Clustering

Revised cluster	*Original current AHS cluster*
RC1	BC1, BC2, BC4, BC5, BC6, BC9
RC2	BC5
RC3	BC5
RC4	Cold deck/deductive imputation
RC5	BC4, BC5, BC6
RC6	BC6
RC7	BC5
RC8	BC4
RC9	BC6
RC10	BC6

4. A Comparison of Methodologies

The third research question was to empirically evaluate the magnitude of association lost under the current methodology. To address this question, a comparison of imputations under the current methodology and alternate methodology was completed. Performance of each methodology was reviewed two ways.

With new hot-deck imputation cells based on the most highly associated independent variables, it is expected that associations between the imputed variable and those independent variables would increase. Alternatively, because the current methodology uses imputation cells based on independent variables assumed, but not measured to be associated, it is hypothesized that the level of association would not necessarily increase after imputation. Therefore, the first review of methodologies produced a comparison of the R^2 deviance before and after imputation. This comparison was completed using both the revised imputation methodology and the current hot-deck imputation methodology.

An R^2 deviance was calculated before imputation and again after imputation under each methodology ($DevRsq_{before}$ and $DevRsq_{after}$, respectively). The difference between these R^2 deviances and relative differences was calculated to assess individual measures of association:

$$Diff = DevRsq_{before} - DevRsq_{after}$$

$$RelDiff = \frac{Diff}{DevRsq_{before}}$$

Table 7 provides the relative differences for the R^2 deviances under each methodology for two imputed variables. Cells highlighted in grey show an improvement in association over the compared methodology. Over a comparison of 56 variable associations before and after imputation under both methodologies, 35.7 percent of comparisons showed a larger improvement under the current methodology, 51.8 percent of comparisons showed a larger improvement under the alternate methodology, and 12.5 percent showed the same level of association under both methodologies.

Table 7: A Comparison of Associations Under Both Methodologies

Imputed variable	*Associated variable*	*Relative difference of R^2: Current hot-deck methodology*	*Relative difference of R^2: Revised methodology*
Presence of a basement in unit	Number of bedrooms in unit	0.382	12.369
	Year unit was built	-0.175	-0.359
	Value or rent of unit compared to number of bedrooms	0.541	5.129
	Number of floors in unit	-0.056	-0.368
	Number of half bathrooms in unit	-0.437	1.011
	Age of householder	-0.436	5.001
	Household occupant composition	-0.401	10.5
	Country of birth of householder	-0.089	-0.011
	Race of householder	-0.399	2.056
	Unit's entrance requires stairs	-0.279	-0.691
	Type of building unit is part of	1.385	14.437
	Owner or renter status	-0.681	4.858

Imputed variable	*Associated variable*	*Relative difference of R^2: Current hot-deck methodology*	*Relative difference of R^2: Revised methodology*
Indicator if the unit is a condominium	Number of bedrooms in unit	0.003	-0.009
	Value or rent of unit compared to number of bedrooms	-0.02	0.003
	Presence of working dishwasher	-0.02	-0.024
	Number of floors in unit	-0.005	-0.011
	Age of householder	-0.008	-0.013
	Household occupant composition	-0.001	-0.011
	Race of householder	-0.002	0.015
	Unit is part of homeowners association	0.002	-0.003
	Type of building unit is part of	-0.003	-0.005
	Owner or renter status	-0.054	0.101
	Water and sewage are billed separately	0.002	-0.012

Note: The measures of association calculated prior to imputation include only units with no missing data. Therefore, comparisons should be considered with caution.

The second review of imputation methodologies compared the distribution of point estimates for imputed variables using the alternate imputation methodology and the current hot-deck imputation methodology. These point estimates were compared to identify differences in overall estimates between the two methodologies.

Tables 8 and 9 show distributions for two imputed variables, the indicator of the presence of a basement within a unit and the indicator of whether the unit is a condominium. They include distributions under both the current and alternate methodology. A larger shift in distribution is seen for the indicator of the presence of a basement than for the indicator of whether the unit is a condominium. This suggests an alternate imputation cell structure when imputing the indicator of the presence of a basement within the unit will have a larger effect on published estimates.

Table 8: Comparison of Point Estimates After Implementation of Both Methodologies: Indicator of the Presence of a Basement in Unit

Distribution of values	*Weighted frequency; current hot-deck methodology*	*Percent*	*Weighted frequency; revised methodology*	*Percent*
Missing	3,732	0.00	0	0.00
1	39,724,826	31.58	31,770,181	25.25
2	13,819,115	10.99	9,611,440	7.64
3	25,218,147	20.05	18,332,902	14.57
4	44,153,464	35.10	60,975,283	48.47
5	2,880,297	2.29	5,109,776	4.06
Total	**125,799,582**	**100**	**125,799,582**	**100**

Table 9: Comparison of Point Estimates After Implementation of Both Methodologies: Indicator of Whether Unit Is a Condominium

Distribution of values	*Weighted frequency; current hot-deck methodology*	*Percent*	*Weighted frequency; revised methodology*	*Percent*
1	795,133	0.63	795,820	0.63
2	6,744,664	5.36	6,707,152	5.33
3	118,259,784	94.01	118,296,610	94.04
Total	**125,799,582**	**100**	**125,799,582**	**100**

5. Conclusions and Future Research

This research explored alternative imputation cell construction using highly associated variables for each imputed variable while maintaining all other aspects of the current hot-deck method of imputation for the AHS. Three analyses were used to identify up to five most highly associated variables with any given imputed variable. A comparison of output from these three analyses provided a sense of consistency. However, ultimately, the stepwise logistic regression analyses was used to inform subsequent aspects of the research including a cluster analysis and a comparison of imputation under both methodologies. A further exploration of other methods such as CART or Chi-square Automatic Interaction Detection could also be conducted as alternative means for developing imputation cells.

While this research focused on improving the hot-deck method, and the scope of this project was limited to methodologies similar to the current AHS hot-deck method, there are alternative methods of imputation that could be considered. Multiple imputation is a popular method of imputation among large Federal surveys; it helps alleviate some complications with the hot-deck method, such as underestimation of variances. While Census Bureau surveys tend to rely on hot-deck imputation to fill in missing data for the AHS, there are examples of Federal Government surveys having switched to multiple imputation to avoid some of the issues with the traditional hot deck. See Kennickell (1998) and Schenker et al. (2006).

A fusion of the two methodologies, hot-deck multiple imputation, described by Reilly (1993), could also be explored. Andridge and Little (2010, section 7.3) discuss hot-deck analogies of multiple imputation. The Bayesian Bootstrap (BB) and Approximate Bayesian Bootstrap (ABB) have both been studied as methods of "proper" multiple imputation algorithms. The authors note that applications of BB and ABB to complex sample designs "remain largely unexplored" (Andridge & Little, p. 9).

Acknowledgements

The U.S. Census Bureau supported the research in this paper. Specifically, the authors thank Stephen Ash, Sean Dalby, and Ernie Lawley for their help in understanding the current AHS imputation methodology and the variables and modules.

References

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40–64.

Barger, J. (2016). *Imputation in the American Housing Survey*. Bureau of the Census. Internal report. June 13.

Kennickell, A. (1998). *Multiple imputation in the Survey of Consumer Finances*. Federal Reserve System. Retrieved from https://www.federalreserve.gov/econresdata/scf/files/impute98.pdf

Reilly, M. (1993). Data analysis using hot deck multiple imputation. *Journal of the Royal Statistical Society, 42*(3), 307–313.

Schenker, N., Raghunathan, T. E., Chiu, P. -L., Makuc, D. M., Zhang G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association, 101*, 924–933.