# Applying Cluster Analysis to Improve the American Housing Survey Hot Deck December 2021

Brian Shaffer[1], Kathy Zha[1], Stephen Ash[2]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2]Bureau of Labor Statistics, 2 Massachusetts Avenue N.E., Washington, DC 20212

**Abstract**
The American Housing Survey (AHS) currently uses a hot-deck method to impute missing values for approximately 120 variables. Hot deck methodology for AHS involves imputing values for nonrespondents with values of respondents. The imputation process is completed within disjoint subsets of the universe, which we refer to as donor pools. We define the donor pools with auxiliary variables that are available for both the respondents and nonrespondents. In our paper, we introduce new auxiliary variables and apply cluster analysis to produce improved donor pools that minimize within-pool variation across all variables that use each set of donor pools. We also generate donor pools for imputing a single variable.

We describe the clustering methods used to define the donor pools; the methods include classification and regression trees (CART), hierarchical agglomerative clustering, and k-means clustering. We compare the donor pools by measuring the within-pool variation of the imputed variables using the current donor pools and the alternative donor pools. We also will compare our results with imputed values generated with multivariate multiple imputation methods.

**Key Words:** imputation, cluster analysis

## 1. Introduction

Survey respondents occasionally neglect to answer every question. This presents a challenge to the data analyst, who is attempting to produce an estimate with incomplete data. Since the analyst is estimating a population parameter, she/he is more concerned with the distribution of the sample rather than the individual values and therefore uses imputation to complete the dataset. Imputation is an accepted practice by which the analyst uses a plausible value to fill in for the one that failed to respond (Andridge and Little 2010).

Andridge and Little (2010) summarize the three mechanisms for missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR). If Y is MCAR, the probability of its missingness is equal for all sample units. We would expect the distribution of the unobserved values of Y to match the distribution of the observed values. We can use the entire sample of respondents as one donor pool. If Y is MAR, an auxiliary variable is providing the missingess mechanism. For instance, a renter would be less likely to know when their residence was built than an owner because construction date is given when the owner is buying the house. We must condition on the auxiliary variables by stratifying the sample into donor pools such that within a donor pool

Y is essentially MCAR. If Y is NMAR, the mechanism is directly related to Y. For example, if real estate recordkeeping were less reliable for older houses, we would expect more missingness to the 'year built' question for older houses.

The American Housing Survey uses hot deck imputation to impute values for most of its item-nonrespondents. In hot deck imputation, we obtain the 'plausible' value from one of the other respondents in the survey. We group respondents and non-respondents alike into mutually exclusive 'donor pools,' meaning that a respondent donates its value to the non-respondent within the same pool. We build the donor pools using auxiliary information that is known for all survey respondents. The auxiliary variables are most effective if (a) they are associated with Y, the imputed variable and (b) they are associated with the respondent's propensity to respond to Y. Andridge and Little (2010) note that both the variance in Y and the nonresponse bias are reduced if both conditions are met.

We obtain this auxiliary information from the frame, from other surveys, or from the survey responses. Auxiliary information from the survey responses should be complete, and we must either impute any missing values in those variables prior to using that auxiliary variable to build the donor pool or group those cases into a "don't know" donor pool stratum. This "don't know" stratum is purely pragmatic, as it then assumes a MCAR process.

## 2. Current Methods

The AHS edits assign sample housing units to donor pools, which are also called matrices. Each matrix is a collection of disjoint donor pools, which we define with a set of auxiliary variables that are known for all observations. The donor pool contains both observations that did respond to the question (donors) and observations that did not respond (recipient). To impute a response to a recipient, we first identify the variable's universe of interest. If an eligible observation responded, it becomes the donor for the next recipient. If an eligible observation did not respond, it becomes the recipient. The AHS hot deck is deterministic, as opposed to random (Andridge and Little 2010), in that it applies a nearest-neighbor method of sorting cases within a donor pool based on a set of geographic variables. Three matrices received our attention in this paper: Matrix A, Matrix B, and Matrix E.

Matrix A produces the imputation cells for three variables: structure containing the housing unit, housing unit type, and number of units. The cell definitions use four auxiliary variables: interview status, tenure, type of vacancy, and number of floors in the building.

Matrix B produces the imputation cells for twelve variables corresponding with the numbers of different types of rooms within the housing unit; for example, the number of bedrooms or the number of bathrooms. The hot deck produces 29 cells, using the variables interview status, tenure, type of vacancy, structure containing the housing unit, and persons in the unit as auxiliary variables.

Matrix E produces the imputation cells for a wide array of variables. Nine modules use Matrix E's donor pool definitions. Among these are the equipment module – which includes variables like heating equipment, cooking fuel type, source for water; the breakdown module – which includes variables like exposed wire and evidence of roaches; and a module called out-of-sequence households – which includes variables that become auxiliary variables for other imputed variables, like rent and housing unit value. This matrix

uses auxiliary variables interview status, tenure, structure containing the housing unit, demographic information about the householder, number of bedrooms, and value/rent. We impute value and rent using Matrix E before the other variables, using the remaining auxiliary variables to produce their donor pools.

## 3. Cluster Analysis

Cluster analysis refers to a set of techniques used to divide a set of observations into mutually exclusive, meaningful groups. Our goal is to produce groups that are homogeneous within and heterogeneous between groups, with respect to one or more variables. We may not know the underlying mechanism that produces these relationships, but we do assume that the observed relationship generalizes to unobserved observations.

### 3.1. Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering refers to a set of techniques in which each observation starts in a cluster by itself. Using a distance measure, we combine all of the clusters such that each cluster is paired with its nearest neighbor. This continues until we have combined all observations into one cluster. Finally, we employ a dendrogram, or tree diagram, to determine the clustering scheme that satisfies the number of clusters we specify.

Ward's method (Ward 1963) defines the distance between two clusters as the amount of increase in the sums of squares when two clusters merge. This distance measure between cluster K and cluster L, $D_{KL}$, is defined as:

$$D_{KL} = \frac{\|\bar{\mathbf{x}}_{\mathbf{K}} - \bar{\mathbf{x}}_{\mathbf{L}}\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

Where each element of the $\bar{\mathbf{x}}_{\mathbf{i}}$ vector is the variable's mean across all observations within the $i^{th}$ cluster and $N_i$ is the number of observations in the $i^{th}$ cluster.

When the cluster contains one observation, $\bar{\mathbf{x}}_{\mathbf{i}}$ is the vector of observed values for that observation. As we group observations into subsequent clusters, we recalculate $\bar{\mathbf{x}}_{\mathbf{K}}$ and $\bar{\mathbf{x}}_{\mathbf{L}}$ using the original observations.

After determining the order of cluster pairings from *n* to one, we select a number of clusters. Each decision minimize the number of clusters results in an increase in the between-cluster variation. This increase is represented in a dendrogram, or tree diagram. Figure 1 provides an example of a dendrogram. To interpret it, imagine a vertical line moving from right to left. Any values to the left of an intersection are part of the cluster. As the vertical line moves left, there are more intersections and hence, more clusters.

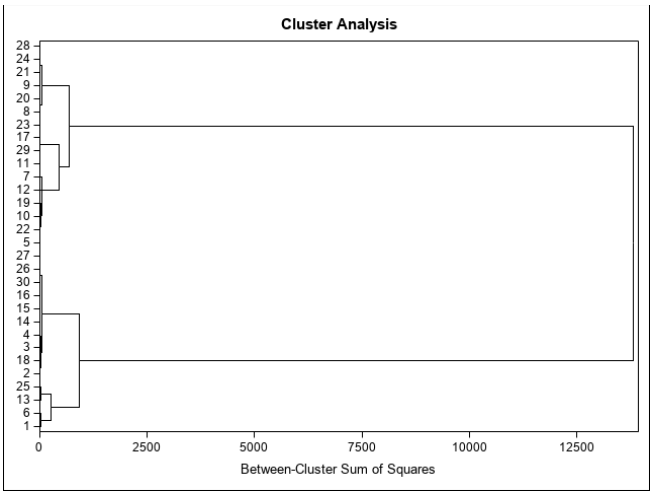


**Figure 1:** Dendrogram example
Source: U.S. Census Bureau, 2017 American Housing Survey

### 3.2. *K*-means clustering (MacQueen 1967; Anderberg 1973)

With *K*-means clustering, we set *K* to equal the number of clusters we want and group the observations into those *K* clusters.

We begin by selecting our initial seeds. The seeds represent the cluster centroids. The distance between the $i^{th}$ unit's reported values and the *K* cluster centroids is defined as the Euclidean distance between the $K^{th}$ centroid and the $i^{th}$ observation:

$$D_{i,K} = \|\mathbf{x}_i - \bar{\mathbf{x}}_K\|$$

Where $\bar{\mathbf{x}}_K$ is the vector of centroid means across the *n* variables. At the end of each iteration, we calculate $D_{i,K}$ for each observation across the *K* clusters. We assign each observation to the cluster where its value of $D_{i,K}$ is the smallest for the *K* clusters. After assigning each observation to a cluster, we recalculate each variable's centroid. We repeat this until the centroids stop changing.

### 3.3. Classification and Regression Trees

Decision trees provide a way to conceptualize prediction of a variable's levels based on a number of auxiliary variables. The response can be categorical or numeric. Each node of a decision tree typically represents a binary decision point; for instance, to predict whether a housing unit uses natural gas we may ask 'do most housing units in the block use natural gas?'

A classification tree models categorical responses while a regression tree models continuous responses. However, both methods follow a similar algorithm. First, we begin with the root node, which contains all observations.

Second, we split the observations based on independent partitions on the levels of each auxiliary variable, such that we seek to minimize impurity, or the variation of the response within each child node. If the auxiliary variable is continuous, the algorithm finds the cutpoint that best divides the observations. This second step is known as 'growing' the tree.

In the SAS HPSPLIT procedure (SAS Institute 2015), the reduction in impurity is defined as

$$\Delta i(s,\tau) = i(\tau) - \sum_{b=1}^{B} p(\tau_b|\tau)\, i(\tau_b),$$

where $i(\tau)$ is the impurity of the parent node, $i(\tau_b)$ is the impurity of child node $b$, and $p(\tau_b|\tau)$ represents the weighted proportion of the number of units in the sample that are in child node $b$. We did not use weights in our application of CART; therefore, each weight equaled one.

The definition of our impurity $i(\tau)$ varies based on the type of data we are clustering. For numeric responses, the SAS HPSPLIT RSS grow criterion defines $i(\tau)$ as the residual sum of squares,

$$i(\tau) = \frac{1}{N_\omega(\tau)} \sum_{i=1}^{N(\tau)} (Y_i - \bar{Y}_\omega)^2,$$

where $N(\tau)$ is the number of observations, $N_\omega(\tau)$ is the weighted sum of observations, $Y_i$ is the value of the response, and $\bar{Y}_\omega$ is the weighted mean of the response variable.

The within-node sum of squares displayed in the SAS output as 'RSS' is calculated as

$$SS_{within} = \sum_{b=1}^{B} \sum_{i=1}^{N(\tau_b)} \omega_i \left(Y_{bi} - \bar{Y}_\omega(\tau_b)\right)^2.$$

For categorical responses, we define $i(\tau)$ as the entropy impurity,

$$i(\tau) = -\sum_{j=1}^{J} p_j log_2 p_j,$$

where $p_j$ is the weighted proportion of the sample that have the $j^{th}$ response value.

The HPSPLIT procedure selects the best splitting variable and the best cutoff value to produce the highest reduction in impurity. We continue splitting the nodes based on our variance-minimizing criterion, which eventually could contain so many nodes that they could not be generalized back to new data.

To avoid overfitting the model to the data, the last step is to prune the tree. We employed the Cost-Complexity criterion proposed by Breiman et al. (1984), which essentially is a function that combines an error rate with a penalty function that increases as the number of leaves increases. For categorical responses, the error rate is equal to the proportion of cases misclassified. For numeric responses, the error rate is equal to $SS_{within}$.

## 4. Multiple Imputation

The Census Bureau is researching an alternative imputation method called Multiple Imputation (Dalby et al. 2019). Multiple Imputation (MI) is a model-based imputation method that estimates a distribution of the imputation variable and draws multiple values from the estimated distribution.

The current MI research evaluates the Fully Conditional Specification modeling approach (Dalby et al. 2019), which iteratively estimates distributions for imputation variables, one at a time, such that newly imputed variables can be included in the model to impute subsequent variable distributions. After a short number of iterations, the distributions converge. The evaluation method in Dalby et al incorporates randomness to the regression parameters as well as the predicted values.

## 5. Results

We applied cluster analysis to build hot-deck donor pools. Instead of grouping individual sample units, we combined groups of sample units. We calculated the group mean of each variable, and calculated our distance measures with these means. Either directly obtained from the respondent or produced through coding, the hot deck uses categorical auxiliary variables. We created the auxiliary variable groups by identifying all of their possible combinations; for example, one auxiliary variable with five levels and another with four levels can produce up to 20 groups. These 20 groups would serve as our initial donor pools. The cluster analysis determines how to combine the 20 groups into our final donor pools.

Both hierarchical and *k*-means clustering use a distance measure, which requires numeric values. However, a given hot deck matrix can contain different types of variables. For instance, Matrix E imputes categorical variables such as 'type of heating fuel used.' We expressed those variables as sets of binary variables. For example, HEAT1=1 if fuel type=A, HEAT1=0 otherwise; HEAT2=1 if fuel type=B, HEAT2=0 otherwise; and so forth. Additionally, when we worked with a mix of variable types, we grouped numeric variables into binned categorical variables and converted them to binary variables. This allowed us to keep all imputation variables in the same scale so that the numeric variables did not dominate the distance calculations.

The results in this paper represent a few case studies to evaluate the effect of applying cluster analysis to the hot deck. Except where noted, the intent of this research was to keep auxiliary variables constant between methods and evaluate the impact of changing the way we group the auxiliary variables.

We evaluated Matrices A, B, and E. For each matrix, we identified all AHS-National observations that (a) were completed interviews and (b) provided responses to all of the imputation variables in the matrix, either directly or through a consistency edit.

### 5.1. Matrix A

Matrix A produces the donor pools to impute three variables corresponding with the structure type of the housing unit. Two variables are categorical and one is numeric.

We recoded our auxiliary variables of occupancy status, tenure, and vacancy type into one categorical variable with five levels: owner-occupied, renter-occupied, vacant-sold/for sale, vacant-rented/for rent, and vacant-other. We also recoded number of floors into a categorical variable with six levels: missing, one through four floors, and 5-or-more floors. The five levels of our tenure recode and the six levels of our floors recode gave us 30 mutually exclusive pools with which we combined with the clustering algorithms. Of our imputation variables, structure containing the housing unit and housing unit type are categorical, while number of units, called NUNITS, is numeric. To standardize all three variables, we first recoded NUNITS into a categorical variable representing the published

ranges for multiunit buildings; 2-4, 5-9, 10-19, 20-49, and 50+. We then recoded our three variables into a series of binary variables. Standardizing our imputation variables put them all into the same scale when we applied the distance calculations in our cluster analysis.

We evaluate the variable NUNITS here. We focus on the number of units only in multiunit buildings because the value equals one for single-family attached, detached, and mobile homes. There were 67,000 respondents in the 2017 AHS. Of those, 64,500 housing units provided responses to all three variables in Matrix A – structure containing the housing unit, housing unit type, and NUNITS. We only used respondents to all three variables in the cluster analysis. Of the units we used, there were 17,000 housing units in multifamily buildings. Overall, 2,200 units in multifamily buildings did not respond to the NUNITS question.

We applied Ward's method and K-means clustering to produce donor pools that take into account all of the respondents, while we applied CART to produce donor pools for only those units in multiunit buildings. For each method we iterated cluster size from c=3 to 30, the maximum number of available clusters.

We applied Ward's method to group our 30 initial donor pool clusters hierarchically. For each recoded version of the three imputation variables, we calculated our 30 initial donor pool cell proportions. We used the initial donor-pool cell proportions as the basis for our distance measures. We used the relationships between initial clusters represented in this dendrogram to group the 30 initial donor pool cells into final clusters, from c=three clusters to c=30.

Next, we applied K-means clustering to our 30 initial donor pools. Similar to our approach in hierarchical clustering, we calculated distances between our *c* centroids and our pools' values, defined as the standardized variable proportions. We specified from *c*=three centroids to *c*=30.

Lastly, we applied CART to NUNITS, using the subset of the sample that only includes multiunit buildings. We wanted to compare the effectiveness of clustering one variable at a time to clustering all variables simultaneously. To keep our imputation variables consistent across methods, we used our recoded NUNITS, which we recoded as categorical. We used the entropy criterion to grow the tree and a cost complexity criterion to prune the tree.

After developing the clusters described above, we calculated Mean Square Error (MSE) of the observed value NUNITS with an Analysis of Variance for each method / number of clusters. Figure 2 displays the change in MSE by method as the number of clusters increase. We also provide the MSE associated with the current method, denoted as a horizontal black line for comparative purposes.

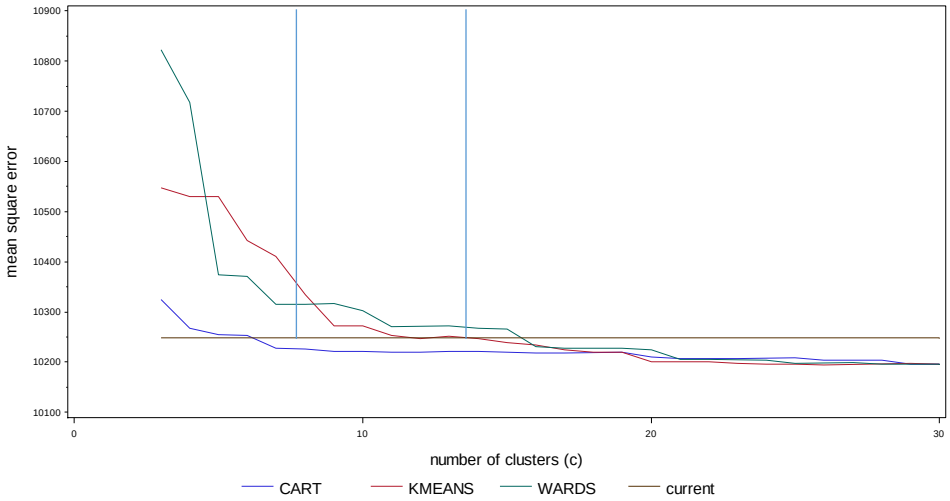


**Figure 2:** Mean Square Error, NUNITS within cluster.
Source: U.S. Census Bureau, 2017 American Housing Survey

We see in Figure 2 that the current method does an adequate job accounting for the variation in NUNITS captured with the auxiliary variables, when compared its Mean Square Error with those from the 30-cluster scenarios. With respect to improvements - we note the current method produces nine donor pools. Figure 2 suggests that for nine donor pools, CART performs slightly better than the current method, while K-means and Ward's Hierarchical clustering performs slightly worse. However, as the number of clusters exceeds 15, all methods perform slightly better than the current method.

We applied our donor pool definitions back to the sample to impute the 2,200 missing values for NUNITS. We specified nine clusters for all cluster-based methods, the same number of donor pools specified in the current method. We sorted our sample by donor pool, then by the series of geographic variables used in the current methodology. If the first value within a pool was missing, we selected from a random uniform distribution within each donor pool. In Figure 3, we compare the distributions of the currently imputed values with the three candidate methods, no cells – only a geographic sort, and the 17,000 values from those units that did provide responses.

Figure 3 suggests that all methods produce similar distributions of imputed values. We also note some slight but clear separation in the distributions starting near the 80$^{th}$ percentile. The reported values at this point are smaller than the imputed values. We surmise that the auxiliary variables are not completely capturing the mechanism that drives the missingness in NUNITS. We explore this at a future point.

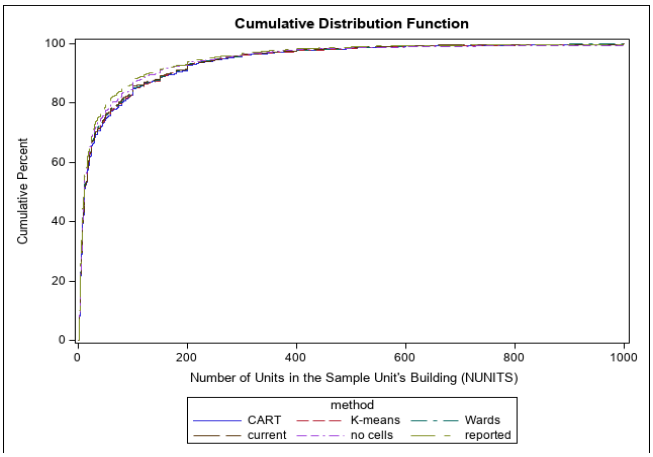


**Figure 3:** Cumulative distributions of imputed NUNITS values by method, with reported values for comparison.
Source: U.S. Census Bureau, 2017 American Housing Survey

We now compare values imputed with the hot deck to those imputed with multiple imputation. As our three methods performed relatively consistently, we selected one method, Ward's Hierarchical, as our base of comparison to MI. We simulated a MAR process in which the missingness varied based on the number of floors in the building. Using our 17,000 cases that (a) were in multifamily buildings (b) responded to the NUNITS interview question, and (c) responded to the 'number of floors in building' interview question. We produced three simulations with different random seeds to determine which observations to remove. For each simulation, we removed 2,700 values. We used the same auxiliary variables to apply our evaluation methods. We then combined the three simulations for evaluation. Figure 4 displays the cumulative distributions of the imputed values from the clustering and MI methods, as well as those imputed with the current hot deck methodology, a pure geographic sort (i.e., no cells), and the combined three sets of 2,700 actual responses we removed.

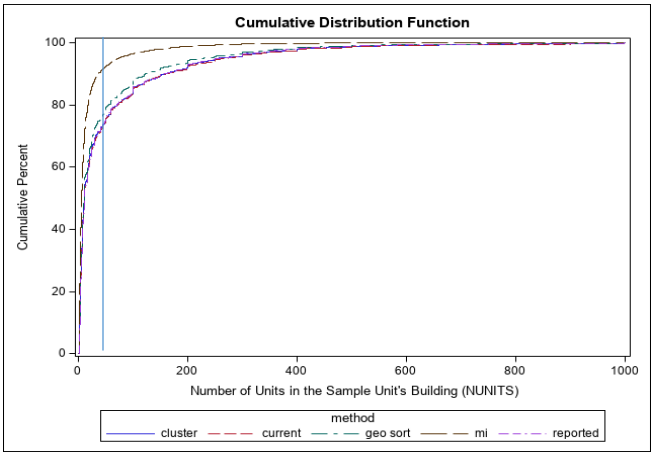


**Figure 4:** Cumulative distributions of imputed and known values from simulations.
Source: U.S. Census Bureau, 2017 American Housing Survey

The distributions of our imputed values using the current hot deck, cluster analysis, and a pure geographic sort all appeared to mimic the distribution of the reported values, while

the distribution of the MI-based imputed values suggests that method imputed a larger proportion of the smaller-sized 50+ unit buildings and fewer larger-sized buildings.

**Table 1:** Summary of differences between observed and imputed by building size, from simulations

| Current | | | | | | |
|---|---|---|---|---|---|---|
| Building Size | Min | Q1 | Median | Q3 | Max | Sum of Squared Differences |
| 2-4 | -996 | -14 | -4 | 0 | 2 | 11,390,000 |
| 5-9 | -992 | -16 | -2 | 2 | 7 | 10,250,000 |
| 10-19 | -988 | -11 | 2 | 8 | 17 | 11,830,000 |
| 20-24 | -978 | -26 | 11.5 | 21 | 46 | 22,050,000 |
| 50+ | -898 | -7 | 62 | 146 | 996 | 99,280,000 |
| Current Total | | | | | | 154,800,000 |
| Cluster Analysis | | | | | | |
| Building Size | Min | Q1 | Median | Q3 | Max | Sum of Squared Differences |
| 2-4 | -996 | -14 | -4 | 0 | 2 | 11,060,000 |
| 5-9 | -992 | -16 | -2 | 2 | 7 | 11,720,000 |
| 10-19 | -988 | -14 | 2 | 8 | 17 | 13,860,000 |
| 20-24 | -978 | -20.5 | 12 | 22 | 47 | 17,090,000 |
| 50+ | -898 | 0 | 65 | 147 | 996 | 98,670,000 |
| | | | | | | 152,400,000 |
| Multiple Imputation | | | | | | |
| Building Size | Min | Q1 | Median | Q3 | Max | Sum of Squared Differences |
| 2-4 | -180 | -7 | -2 | 0 | 2 | 253,700 |
| 5-9 | -731 | -2 | 2 | 4 | 7 | 680,600 |
| 10-19 | -615 | 2 | 8 | 10 | 17 | 1,110,000 |
| 20-24 | -481 | 12 | 20 | 28 | 47 | 2,246,000 |
| 50+ | -849 | 52 | 95 | 194 | 996 | 101,300,000 |
| Multiple Imputation Total | | | | | | 105,600,000 |
| Geo Sort | | | | | | |
| Building Size | Min | Q1 | Median | Q3 | Max | Sum of Squared Differences |
| 2-4 | -995 | -26 | -6 | 0 | 2 | 13,550,000 |
| 5-9 | -992 | -19 | -2 | 2 | 7 | 16,190,000 |
| 10-19 | -988 | -18 | 2 | 8 | 16 | 15,360,000 |
| 20-24 | -958 | -18.5 | 12 | 23 | 46 | 15,030,000 |
| 50+ | -938 | 42 | 93 | 191.5 | 996 | 103,300,000 |
| Geo Sort Total | | | | | | 163,400,000 |

Source: U.S. Census Bureau, 2017 American Housing Survey

To gain insight about the differences suggested in Figure 4, we produced summaries of the differences between the actual and imputed values of NUNITS. Table 1 provides the interquartile ranges of differences, defined as "actual – imputed," by building size and imputation method. We see that MI outperformed all methods for building sizes ranging from 2-unit buildings through 24-unit buildings. We also note that for the middle 50 percent

of the 50+ group, Q1 through Q3, MI consistently produced smaller imputed values than actual, as indicated by the positive differences.

### 5.2. Matrix B

Matrix B produces the donor pools to impute twelve variables corresponding with the types of rooms in the housing unit. Each variable is a count variable. There were 52,500 observations with responses to all 12 interview variables.

We developed our initial cells using the same auxiliary variables used in the current method: occupancy status, tenure, vacancy status, type of housing unit, and persons in the housing unit. We capped persons in the housing unit to six, as sample was sparse beyond six persons. We also recoded occupancy status, tenure, and vacancy into a variable with five levels: owner-occupied, renter-occupied, vacant-sold/for sale, vacant-rented/for rent, and vacant-other. As only occupied housing units contain persons, the variable was somewhat confounded with our occupancy status recode and was only applicable for the owners and renters. Crossing these three variables produced 60 initial donor pool clusters. Similar to Matrix A we converted each imputation variable to a series of indicator variables, calculated cell-level proportions for each indicator variable, and used these proportions to calculate the cluster algorithms' distance measures.

We applied Ward's method to group our 60 initial donor pool clusters hierarchically. For each of the 12 variables we calculated the 60 initial donor pool cell means. We used the initial donor pool cell means as the basis for our distance measures. We also standardized our variables' means prior to calculating distances. We grouped the 60 initial donor pool cells into final clusters, from c=three clusters to c=60.

Next, we applied K-means clustering to Matrix B. Similarly to our approach in hierarchical clustering, we calculated distances between our *c* centroids and our pools' values, defined as the standardized variable means. We specified from *c*=three centroids to *c*=60.

Lastly, we applied CART to two of the 12 variables in Matrix B. We evaluated bedrooms and family rooms, named BEDRMS and FAMRM, respectively. The distribution of BEDRMS is among the most variable of the 12 in the matrix, while the variability in responses decreases for FAMRM. We wanted to compare the effectiveness of clustering one variable at a time as opposed to clustering all variables simultaneously. We used a residual sum of squares criterion to grow the tree and a cost complexity criterion to prune the tree.

After developing the clusters described above, we calculated Mean Square Error (MSE) with an Analysis of Variance for each method / number of clusters. Figures 5 and 6 display the change in MSE by method as the number of clusters increase. We also provide the MSE associated with the current method, given as a horizontal line for comparative purposes.

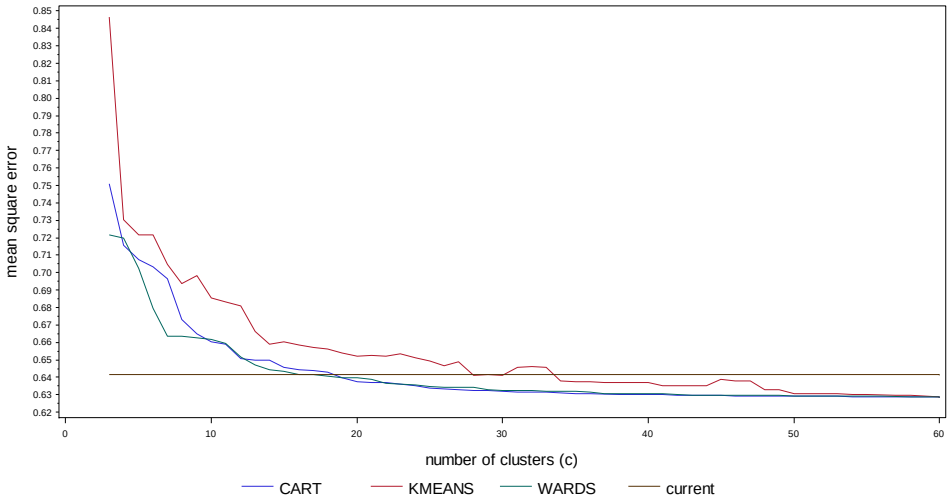


**Figure 5:** Mean Square Error for BEDRMS for the number of clusters (c), by method. Source: U.S. Census Bureau, 2017 American Housing Survey

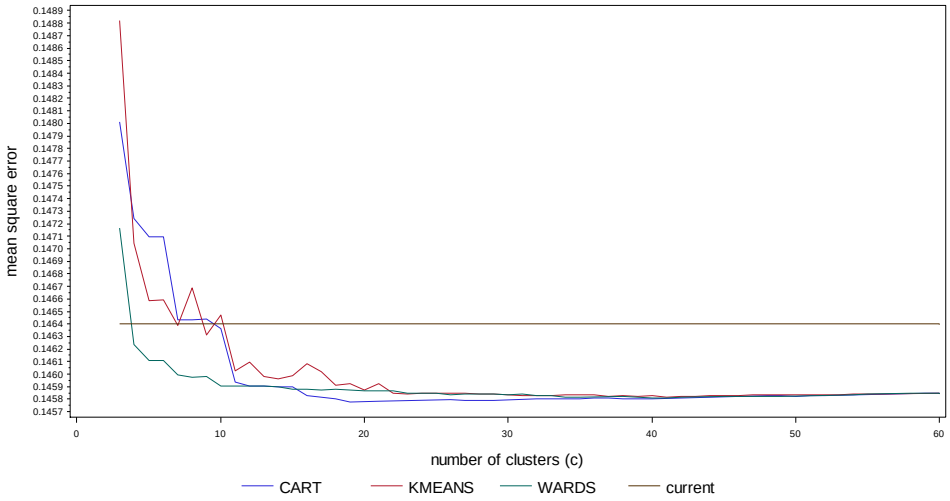


**Figure 6:** Mean Square Error for FAMRM for the number of clusters (c), by method. Source: U.S. Census Bureau, 2017 American Housing Survey

In both figures above, we observe that our within-cell variability hits an inflection point at about ten clusters across all methods. We also notice that with CART and Ward's hierarchical clustering, we can achieve a similar level of within-cluster variability to the current method with fewer than 29 clusters. Overall, CART and Ward's appear to outperform K-means clustering, particularly in the number of bedrooms. Lastly, we note that the compromises we make in accounting for all 12 variables in Ward's hierarchical clustering are not hindering the method's ability to reduce variation in all variables, when compared to clustering individual variables with CART.

We focus on BEDRMS here. In Figure 7, we compare the distributions of the currently imputed values with the three candidate methods, no cells – only a geographic sort, and the values from those units that did provide responses for the number of bedrooms. The purely

geographic sort appeared closest the distribution of the responses, while CART and Ward's Hierarchical clustering appeared to produce consistent distributions with the current method. K-means produced visually inconsistent comparisons between two and four bedrooms.

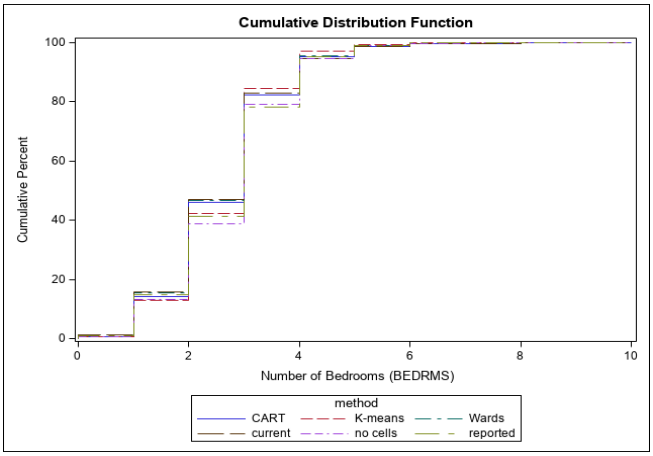


**Figure 7:** Cumulative distributions of imputed and reported Number of Bedrooms.
Source: U.S. Census Bureau, 2017 American Housing Survey

We now compare values imputed with the hot deck to those imputed with multiple imputation. To simplify our visual analysis, we selected one method, Ward's Hierarchical, as our base of comparison to MI. We simulated a MAR process in which the missingness varied based on the tenure and vacancy status of the unit. Using our 52,500 cases that responded to the BEDRMS question, we produced three simulations with different random seeds to determine which observations to remove. For each simulation, we removed on average 650 values. We used the same auxiliary variables to apply our evaluation methods. We then combined the three simulations for evaluation. Figure 8 displays the cumulative distributions of the imputed values from these two methods, as well as those imputed with the current hot deck methodology, a pure geographic sort (i.e., no cells), and the actual responses we removed.

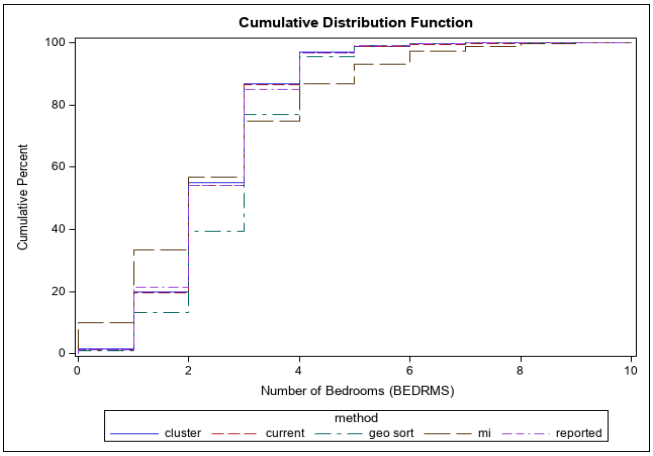


**Figure 8:** Cumulative Distribution of Values with Forced Missingness for Number of Bedrooms, by Method
Source: U.S. Census Bureau, 2017 American Housing Survey

Figure 8 suggests that for the simulated missingness mechanism, both cluster analysis and the current method produce hot deck cells that mimic the distribution of the actual responses. We also note that the auxiliary variables added information that gave us improvements over a pure geographic sort. We finally note that the auxiliary variables, when used by the MI model, produced a similar distribution given in Figure 2d in Dalby et al (2019); such that the MI model drew higher proportions of 0, 1, and 5-or-more-bedroom responses from the estimated distribution than what was available in the distribution of actual responses. We provide a summary of the differences between the reported and imputed values in Table 2. We see that the current and cluster-based methods produce symmetric distributions. We also note the reduced Sum of Squared differences after applying a cell-based hot deck method, when compared to a pure geographic sort. We also note that MI produced the greatest amount of total variability between reported and imputed values in our simulations.

**Table 2:** Summary of differences between observed and imputed, from simulations

| Method | Min | Q1 | Median | Q3 | Max | Sum of Squared Differences |
|---|---|---|---|---|---|---|
| Current | -7 | -1 | 0 | 1 | 8 | 3,050 |
| Cluster | -5 | -1 | 0 | 1 | 8 | 2,824 |
| MI | -7 | -1 | 0 | 1 | 9 | 7,114 |
| Geo Sort | -7 | -1 | 0 | 1 | 7 | 4,900 |

Source: U.S. Census Bureau, 2017 American Housing Survey

### 5.3. Matrix E

Matrix E produces the donor pools to impute sets of many variables from nine modules. Eight of these modules are housing unit-level and one is person-level. In this section we evaluate the impact of clustering with the imputation variables from the eight housing unit-level modules on the quality of the donor pools with respect to one of those modules – equipment. The equipment module contains 26 variables related to kitchen and laundry appliances, bathroom equipment, types of fuel used, and the types of heating and cooling equipment. We first created clusters using the variables from the eight housing unit-level modules and observations with responses to all the imputation variables, which consisted of approximately 13,000 observations. Then we created clusters using only the equipment module's variables from full-respondents, which consisted of about 62,500 observations. After developing clusters, we conducted simulations by mapping the clusters back to the 62,500 observations that provided responses to all equipment variables.

We developed our initial cells using the same auxiliary variables used in the current method: occupancy status, tenure, vacancy status, type of housing unit, number of bedrooms, demographic information about householder, and rent/value. In this research, we are assuming that value and rent are already reported & imputed. We recoded occupancy status, tenure, and vacancy into a variable with five levels: owner-occupied, renter-occupied, vacant-sold/for sale, vacant-rented/for rent, and vacant-other. We recoded number of bedrooms into two levels to mimic Matrix E; one represents two or fewer bedrooms, while the other represents three-or-more. We coded the demographic information to mimic Matrix E: a variable with three levels representing whether (a) the unit is vacant or a mobile home, (b) the householder is present, under 65, and white/non-Hispanic, or (c) the compliment of (b). Altogether, these variables produced 80 initial

donor pool clusters. Similar to Matrix A we converted each imputation variable to a series of indicator variables, calculated cell-level proportions for each indicator variable, and used these proportions to calculate the cluster algorithms' distance measures.

For Matrix E, we produced clusters with Hierarchical clustering and CART. We specified 29 clusters for each method, as the current hot deck contains 29 donor pools. Figure 9 below provides a side-by-side comparison of the within-donor pool distributions for the current method's pools and those constructed with Hierarchical clustering, using the 13,000 observations that provided responses to all imputation variables.

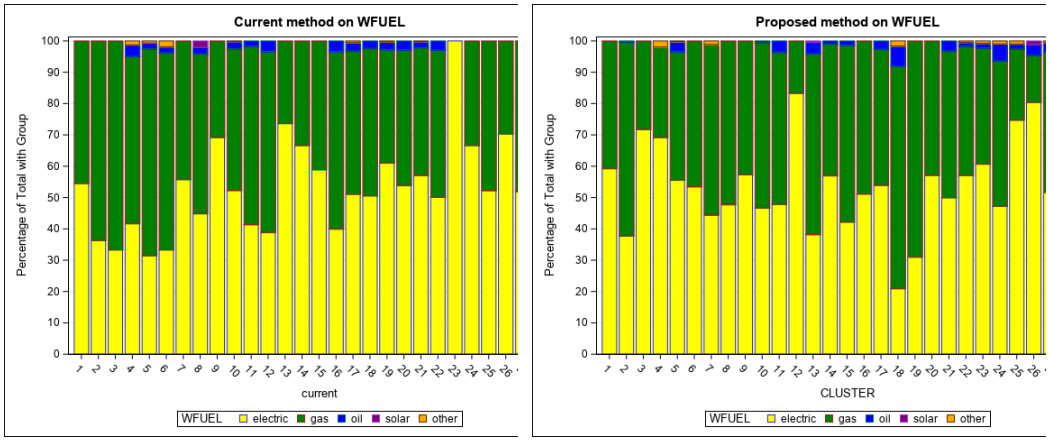


**Figure 9:** Distributions of Water Fuel (WFUEL) donor pools using current method and hierarchical clustering, respectively.
Source: U.S. Census Bureau, 2017 American Housing Survey

We see in Figure 9 that the current method contains one pool, the 23rd pool, consisting of only values of WFUEL=electricity. With a few exceptions, both sets of distributions suggest all donor pools contain approximately 50 percent of cases using electricity and 50 percent using gas.

Of the eight modules we incorporated into our cluster analysis, we evaluate the methods' impacts on the equipment module. We calculated the chi-square statistic for each imputation variable in that module, for the three methods: current, hierarchical clustering, and CART. We excluded K-means to reduce the number of evaluation methods.

We calculated chi-square statistics for all variables/methods to compare their relative amounts of between-cluster variability. We did this for our two aforementioned scenarios: first, we included all eight modules and their observations with responses to all questions; second, we included only the equipment module and their observations with responses to all of those questions. These corresponded with 13,000 and 62,500 observations, respectively. Their chi-square statistics are given in Table 3.

We did not compare chi-square statistics across scenarios, as sample sizes were drastically different. Within each scenario, we see that our two evaluation methods produced donor pool cells with similar levels of between-cluster variability as the current method, suggesting that clustering eight modules' variables produced donor pools of similar quality as those we produced with only the equipment module. We bolded WFUEL, the variable

representing water fuel. These values correspond with the distributions from Figure 9. We reference this variable in a future section.

**Table 3:** Chi-Square statistics calculated for each variable, by evaluation method and input-data scenario.

| | 13,000 observations | | | 62,500 observations | | |
|---|---|---|---|---|---|---|
| Variable | current | cluster | CART | current | cluster | CART |
| COOK | 189 | 269 | 289 | 3,625 | 3,991 | 3,976 |
| BURNER | 49 | 54 | 63 | 277 | 274 | 292 |
| OVEN | 45 | 54 | 56 | 445 | 451 | 460 |
| CFUEL | 455 | 506 | 487 | 2,043 | 2,272 | 2,037 |
| REFR | 318 | 429 | 499 | 7,439 | 7,723 | 7,750 |
| SINK | 112 | 182 | 249 | 2,060 | 2,342 | 2,406 |
| DISH | 1,896 | 1,986 | 1,995 | 10,740 | 11,400 | 11,420 |
| WASH | 1,906 | 2,312 | 2,297 | 19,530 | 22,950 | 22,880 |
| DFUEL | 270 | 315 | 321 | 1,226 | 1,356 | 1,369 |
| DRY | 1,987 | 2,398 | 2,390 | 19,390 | 22,600 | 22,420 |
| HOTPIP | 106 | 223 | 250 | 3,816 | 4,869 | 4,792 |
| TUB | 25 | 53 | 86 | 364 | 854 | 936 |
| TOILET | 27 | 58 | 100 | 379 | 959 | 1,027 |
| BSINK | 26 | 56 | 93 | 379 | 897 | 984 |
| **WFUEL** | **398** | **533** | **481** | **3,705** | **3,901** | **3,814** |
| WATER | 1,241 | 2,086 | 2,504 | 4,127 | 5,112 | 4,946 |
| HEQUIP | 1,077 | 1,614 | 1,512 | 6,683 | 8,190 | 8,322 |
| HFUEL | 1,142 | 1,870 | 1,805 | 6,098 | 6,753 | 6,483 |
| OAFUEL | 54 | 82 | 85 | 34 | 45 | 41 |
| NUMAIR | 1,295 | 1,057 | 1,198 | 1,526 | 1,460 | 1,899 |
| AIR | 380 | 550 | 548 | 2,311 | 2,658 | 2,520 |
| OARSYS | 70 | 63 | 104 | 938 | 519 | 527 |
| FPLWK | 1,471 | 1,814 | 1,857 | 14,860 | 14,850 | 14,860 |
| AFUEL | 104 | 134 | 135 | 193 | 145 | 194 |
| AIRSYS | 369 | 629 | 620 | 3,114 | 3,848 | 3,774 |
| BATHEXCLU | 19 | 19 | 18 | 56 | 62 | 61 |

Source: U.S. Census Bureau, 2017 American Housing Survey

Next, we simulate missingness for each variable and evaluate how closely our imputed values match actual values. For each of our two sample-size-based clustering scenarios, we mapped our clusters back to the sample of 62,500 observations. We note that when mapping the clusters produced with 13,000 observations to the larger sample, some initial donor pool groups were not represented in the final pools; we therefore created a "don't know" stratum to capture them. Next, we conducted 30 simulations where we randomly excluded 10 percent of our responses. We assumed a MCAR process to give all donor pool cells an overall equal rate of missingness.

In our 30 simulations, we imputed values and identified those imputed values that matched the actual value. Next, we calculated the proportion of imputed values that match the actual. We calculated the average of those proportions, and the standard deviation of the difference between the current and proposed methods. Table 4 provides results from our simulations.

If a standardized difference between the current method and evaluation method was greater than 1.645, we flagged it as significant with bold font.

**Table 4:** Results of imputation simulations for each variable, by evaluation method and input-data scenario: Proportion of imputed values that match the actual value. Significant differences in bold.

| | 13,000 observations | | | 62,500 observations | | |
|---|---|---|---|---|---|---|
| Variable | current | cluster | CART | current | cluster | CART |
| COOK | 96.1% | 96.2% | 96.2% | 96.1% | 96.2% | 96.2% |
| BURNER | 73.8% | 75.5% | 73.7% | 73.8% | 75.9% | 76.9% |
| OVEN | 73.3% | 73.1% | 65.7% | 73.3% | 73.2% | 73.8% |
| CFUEL | 65.3% | 65.7% | 65.1% | 65.3% | 65.3% | 65.4% |
| REFR | 97.3% | 97.3% | 97.4% | 97.3% | 97.4% | 97.4% |
| SINK | 98.9% | 98.9% | 98.9% | 98.9% | 98.9% | 98.9% |
| DISH | 69.9% | 69.2% | 69.5% | 69.9% | 70.0% | 70.1% |
| WASH | 80.4% | **81.6%** | 81.2% | 80.4% | **81.7%** | **81.7%** |
| DFUEL | 76.0% | 76.3% | 76.3% | 76.0% | 76.1% | 76.4% |
| DRY | 79.5% | **80.5%** | 80.3% | 79.5% | **80.6%** | **80.7%** |
| HOTPIP | 98.5% | 98.5% | 98.5% | 98.5% | 98.5% | 98.5% |
| TUB | 99.7% | 99.7% | 99.7% | 99.7% | 99.7% | 99.7% |
| TOILET | 99.8% | 99.8% | 99.8% | 99.8% | 99.8% | 99.8% |
| BSINK | 99.8% | 99.7% | 99.8% | 99.8% | 99.8% | 99.8% |
| WFUEL | 64.9% | 65.6% | 65.4% | 64.9% | 65.4% | 65.5% |
| WATER | 88.5% | 89.0% | 88.8% | 88.5% | 88.8% | 88.9% |
| HEQUIP | 56.9% | 57.2% | 57.3% | 56.9% | 57.4% | 57.3% |
| HFUEL | 60.6% | 61.3% | 61.3% | 60.6% | 61.1% | 61.3% |
| OAFUEL | 92.0% | 92.2% | 92.4% | 92.0% | 92.5% | 92.3% |
| NUMAIR | 40.4% | 40.2% | 40.7% | 40.4% | 40.8% | 41.0% |
| AIR | 71.9% | 72.0% | 72.1% | 71.9% | 72.2% | 72.0% |
| OARSYS | 83.3% | 83.1% | 83.2% | 83.3% | 83.2% | 83.2% |
| FPLWK | 69.7% | 69.0% | 69.7% | 69.7% | 69.8% | 69.7% |
| AFUEL | 93.7% | 93.6% | 93.6% | 93.7% | 93.6% | 93.7% |
| AIRSYS | 74.3% | 74.1% | 74.1% | 74.3% | 74.6% | 74.7% |
| BATHEXCLU | 84.6% | 86.7% | 86.3% | 84.6% | 87.5% | 90.0% |

Source: U.S. Census Bureau, 2017 American Housing Survey

From Table 4, we see that our evaluation methods overall performed consistently with the current method in imputing the actual value. The WASH and DRY variables, representing whether the unit has a working washing machine and clothes dryer, respectively, were the only variables to show a significant increase in matches from the current method.

## 6. Improvements using Auxiliary Variables

This section contains potential improvements we found for Matrices A and E. While we saw in this paper that cluster-based methods could produce hot deck donor pools, the methods alone have not produced substantial improvements with respect to increasing the homogeneity within donor pools. We also need to consider alternative auxiliary variables. We explored the American Community Survey (ACS) as a source for auxiliary variables.

### 6.1. Matrix A - NUNITS

We now discuss the differences between the distributions of the imputed and reported values of NUNITS. Recall we assumed a MAR process when we conditioned on vacancy status, tenure, and floors when developing hot deck donor pools. We surmise that the missingness mechanism may actually be related to the variable itself, as a respondent may have a more difficult time responding to the question 'how many units are in this building' if the building contains many units.

This time we introduced a new variable to our models: block-level housing-type distributions calculated from the ACS. Using the ACS we produced block-level proportions of single-family detached units, single-family attached units, mobile units, units in 2-4 unit buildings, units in 5-9 unit buildings, units in 10-19 unit buildings, units in 20-49 units buildings and units in buildings with 50 or more units. Then we recoded these proportions into one categorical variable representing which type of unit constitutes the majority of the units in the block. This variable, called BLKUNIT, contains ten levels; one for each of the types given above, the ninth representing no clear majority, and the tenth represents a "don't know" stratum. We need this "don't know" stratum because we were not able to map the entire AHS sample to the ACS blocks.

We also evaluated response propensity of NUNITS as a function of BLKUNIT. Table 5 provides odd ratios comparing the eight levels of BLKUNIT that represent a majority in the block to level representing level I – no clear majority, modeling the probability of a response as a function of the levels of BLKUNIT. We used that subset that contained a block-level match to the ACS.

**Table 5:** Response propensity by block housing-type composition, units in multiunit buildings

| BLKUNIT level, based on majority | Respondents | Non-Respondents | Estimate | Lower 95 percent Wald | Upper 95 percent Wald |
|---|---|---|---|---|---|
| A – Single-Detached | 1600 | 100 | 1.878 | 1.539 | 2.291 |
| B – Single-Attached | 400 | 40 | 1.399 | 0.998 | 1.960 |
| C – Mobile | 60 | N < 15 | 2.531 | 0.790 | 8.109 |
| D – 2-4 units | 1900 | 100 | 2.641 | 2.134 | 3.268 |
| E – 5-9 units | 1400 | 80 | 2.284 | 1.806 | 2.888 |
| F – 10-19 units | 1400 | 150 | 1.219 | 1.019 | 1.458 |
| G – 20-49 units | 1000 | 150 | 0.918 | 0.763 | 1.105 |
| **H – 50+ units** | **2800** | **700** | **0.558** | **0.499** | **0.623** |
| I – no majority | 5800 | 800 | N/A | N/A | N/A |

Source: U.S. Census Bureau, 2017 American Housing Survey

Table 5 supports that units have a lower response propensity if they are in blocks that contain a majority of units in buildings with 50 or more units.

We calculated MSE for our scenarios ranging from c=3 to 30 clusters. Figure 10 shows that for Ward's method and CART, the BLKUNIT variable captured substantial between-cluster variation, when compared with the current method, still denoted by the horizontal black line in the figure. We also note that K-means did not stabilize until c=10 clusters,

and some of those clusters only contained one initial donor pool that represented a single observation.

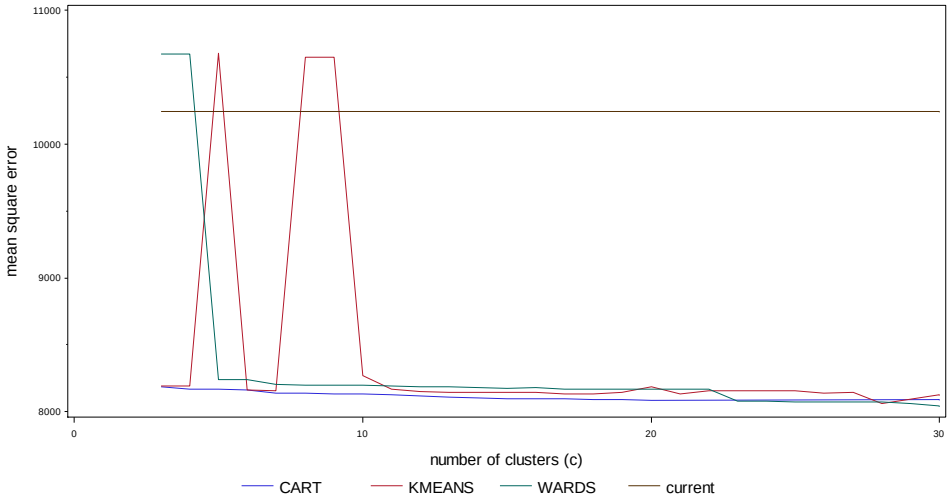


**Figure 10:** Mean Square Error for NUNITS after adding BLKUNIT auxiliary variable for cluster=c.
Source: U.S. Census Bureau, 2017 American Housing Survey

Next, we imputed values using our test methods, specifying nine clusters for all methods. The cumulative distributions are given in Figure 11. We note that CART and Ward's both produced distributions that represented a higher proportion of imputed values containing larger values of NUNITS than from the distribution of reported values.

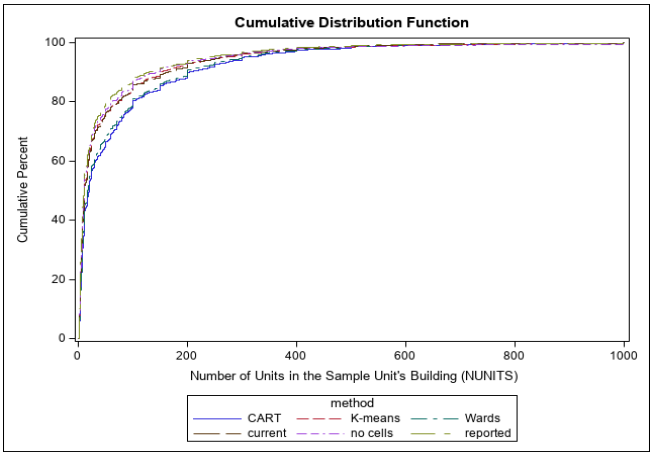


**Figure 11:** Cumulative distributions of imputed NUNITS values by method after including ACS block-level structure-type auxiliary information to cluster analysis, with reported values for comparison.
Source: U.S. Census Bureau, 2017 American Housing Survey

To test our surmise we simulated a NMAR mechanism in which we introduced missingness to our observed data at a variable rate depending on the number of units in the building. We estimated nonresponse rates using Table 5. For buildings with 2-4 units and 5-9 units, the rate was five percent; for buildings with 10-19 units, the rate was 10%; for buildings

with 20-49 units, the rate was 15%; and for buildings with 50+ units, the rate was 20%. We produced three simulations with different random seeds. Figure 12 provides the cumulative distributions from these simulations. We see from such simulations that while no method's imputed values overlapped the cumulative distribution for the respondents, the distribution produced with the cluster-based imputations, using Ward's method, was the closest to the distribution of the actual responses.

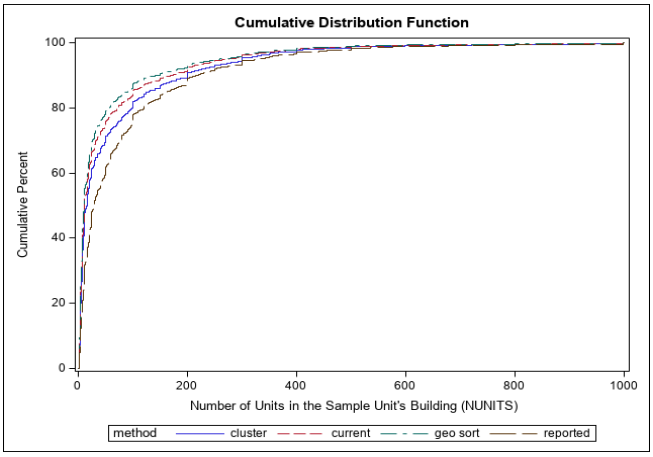


**Figure 12:** Cumulative distributions of imputed and known values of NUNITS from simulations, after including ACS block-level structure-type auxiliary information to cluster analysis.
Source: U.S. Census Bureau, 2017 American Housing Survey

We also summarize the distributions of the differences between observed and imputed values from our three simulations in Table 6. We see that both the middle 50 percent of the distribution of those differences and the total sum of squared differences was smallest for the cluster-based hot deck.

**Table 6:** Summary of differences between observed and imputed, from simulations

| Method | Min | Q1 | Median | Q3 | Max | Sum of Squared Differences |
|---|---|---|---|---|---|---|
| Current | -996 | -8 | 7 | 53 | 996 | 144,000,000 |
| Cluster | -996 | -8 | 4 | 38 | 994 | 131,700,000 |
| Geo Sort | -996 | -6 | 8 | 62 | 996 | 150,300,000 |

Source: U.S. Census Bureau, 2017 American Housing Survey

### 6.2. Matrix E - Fuels

We know from life experience that some localities' infrastructures provide natural gas to their housing units and some do not. We surmised that this could help us to improve imputation of fuel variables, as most housing units use electricity and/or natural gas/propane to heat their homes, hot water, stovetops, etc. We calculated block-level proportions of housing units that use natural gas/propane as their heating fuel and mapped those proportions to the AHS. We produced a "don't know" stratum where we did not have a block-level match. Then we created four more levels corresponding with less than ten percent using gas to heat the home, ten to 50 percent, 50 to 90 percent, and greater than 90 percent. We call this variable BLKGAS.

Figure 13 below shows the distributions of the WFUEL variable produced with the current method, and with hierarchical clustering using BLKGAS, respectively. We produced both sets with the 13,000 observations with responses to all the Matrix E variables we discussed earlier. The increased between-cluster variability is visible after adding the BLKGAS auxiliary variable.

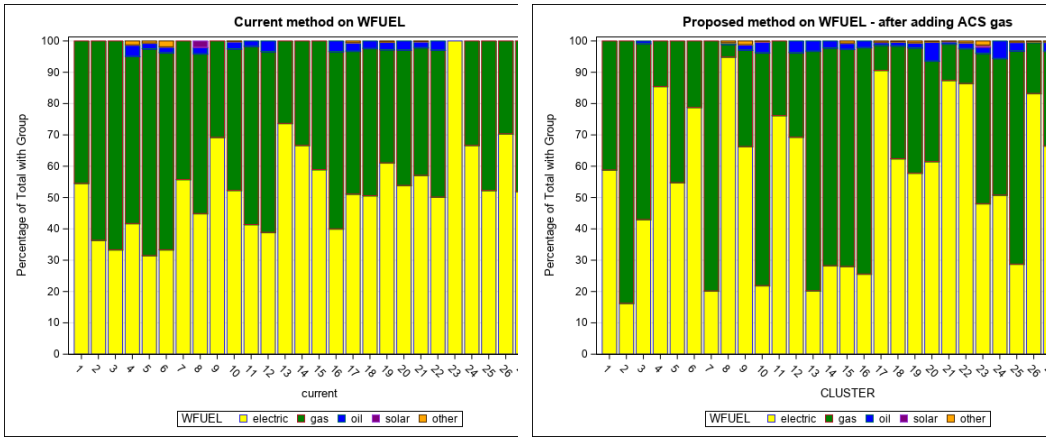


Figure 13. Distributions of Water Fuel donor pools using current method and hierarchical clustering – including BLKGAS variable, respectively.
Source: U.S. Census Bureau, 2017 American Housing Survey

We recalculated chi-square statistics from section 5.3, after introducing the BLKGAS auxiliary variable to hierarchical clustering and CART. We compare their relative amounts of between-cluster variability to the current method in Table 7.

**Table 7:** Chi-Square statistics calculated for each variable after including BLKGAS auxiliary variable, by evaluation method and input-data scenario.

| | 13,000 observations | | | 62,500 observations | | |
|---|---|---|---|---|---|---|
| Variable | current | cluster | CART | current | cluster | CART |
| COOK | 189 | 372 | 474 | 3,625 | 3,994 | 4,373 |
| BURNER | 49 | 52 | 82 | 277 | 260 | 327 |
| OVEN | 45 | 55 | 74 | 445 | 403 | 480 |
| **CFUEL** | **455** | **2,709** | **2,775** | **2,043** | **9,453** | **9,279** |
| REFR | 318 | 665 | 847 | 7,439 | 7,746 | 8,606 |
| SINK | 112 | 319 | 538 | 2,060 | 2,248 | 2,673 |
| DISH | 1,896 | 2,206 | 2,329 | 10,740 | 11,360 | 11,650 |
| WASH | 1,906 | 2,378 | 2,428 | 19,530 | 23,040 | 22,780 |
| **DFUEL** | **270** | **1,020** | **1,004** | **1,226** | **5,102** | **4,568** |
| DRY | 1,987 | 2,457 | 2,494 | 19,390 | 22,750 | 22,200 |
| HOTPIP | 106 | 329 | 428 | 3,816 | 4,977 | 5,696 |
| TUB | 25 | 69 | 122 | 364 | 674 | 1,081 |
| TOILET | 27 | 57 | 125 | 379 | 769 | 1,174 |
| BSINK | 26 | 58 | 130 | 379 | 727 | 1,114 |
| **WFUEL** | **398** | **3,607** | **3,801** | **3,705** | **16,610** | **16,130** |
| WATER | 1,241 | 1,800 | 3,341 | 4,127 | 5,390 | 6,113 |
| HEQUIP | 1,077 | 2,850 | 2,767 | 6,683 | 14,150 | 11,410 |

| | 13,000 observations | | | 62,500 observations | | |
|---|---|---|---|---|---|---|
| **HFUEL** | **1,142** | **5,230** | **4,784** | **6,098** | **18,400** | **17,160** |
| OAFUEL | 54 | 67 | 114 | 34 | 111 | 190 |
| NUMAIR | 1,295 | 574 | 1,256 | 1,526 | 1,217 | 1,703 |
| AIR | 380 | 1,000 | 982 | 2,311 | 3,449 | 2,007 |
| OARSYS | 70 | 62 | 178 | 938 | 537 | 611 |
| FPLWK | 1,471 | 1,998 | 2,121 | 14,860 | 15,390 | 15,260 |
| AFUEL | 104 | 217 | 333 | 193 | 514 | 626 |
| AIRSYS | 369 | 1,290 | 1,220 | 3,114 | 5,211 | 4,844 |
| BATHEXCLU | 19 | 18 | 18 | 56 | 61 | 62 |

Source: U.S. Census Bureau, 2017 American Housing Survey

We see in Table 7 that by adding auxiliary information related to home heating fuel, we increased the between-cluster variability for most of our fuel variables. This increase in between-cluster variability is demonstrated for water fuel in Figure 13 above.

**Table 8:** Results of imputation simulations for each variable after adding BLKGAS variable, by evaluation method and input-data scenario: Proportion of imputed values that match the actual value. Significant differences in bold.

| | 13,000 observations | | | 62,500 observations | | |
|---|---|---|---|---|---|---|
| Variable | current | cluster | CART | current | cluster | CART |
| COOK | 96.1% | 96.2% | 96.1% | 96.1% | 96.2% | 96.3% |
| BURNER | 73.8% | 75.9% | 72.0% | 73.8% | 77.0% | 77.1% |
| OVEN | 73.3% | 73.1% | **64.4%** | 73.3% | 72.0% | 73.7% |
| CFUEL | 65.3% | **66.5%** | 65.9% | 65.3% | **66.8%** | **66.2%** |
| REFR | 97.3% | 97.3% | 97.2% | 97.3% | 97.3% | 97.4% |
| SINK | 98.9% | 98.9% | 98.9% | 98.9% | 98.9% | 98.9% |
| DISH | 69.9% | **68.0%** | 69.1% | 69.9% | 69.5% | 69.9% |
| WASH | 80.4% | 80.9% | 80.5% | 80.4% | **81.3%** | **81.5%** |
| DFUEL | 76.0% | 76.7% | 76.1% | 76.0% | 76.5% | 76.0% |
| DRY | 79.5% | 79.7% | 79.4% | 79.5% | **80.3%** | 80.1% |
| HOTPIP | 98.5% | 98.5% | 98.5% | 98.5% | 98.5% | 98.5% |
| TUB | 99.7% | 99.7% | 99.7% | 99.7% | 99.7% | 99.8% |
| TOILET | 99.8% | 99.8% | 99.8% | 99.8% | 99.8% | 99.8% |
| BSINK | 99.8% | 99.8% | 99.7% | 99.8% | 99.8% | 99.8% |
| WFUEL | 64.9% | **67.0%** | **67.3%** | 64.9% | **67.9%** | **67.4%** |
| WATER | 88.5% | 88.8% | 88.8% | 88.5% | 89.0% | 89.0% |
| HEQUIP | 56.9% | 57.3% | 57.5% | 56.9% | 57.5% | 57.5% |
| HFUEL | 60.6% | **62.6%** | **62.9%** | 60.6% | **62.9%** | **62.6%** |
| OAFUEL | 92.0% | 92.6% | 92.3% | 92.0% | 92.7% | 92.4% |
| NUMAIR | 40.4% | 40.6% | 40.0% | 40.4% | 40.5% | 40.6% |
| AIR | 71.9% | 71.5% | 71.3% | 71.9% | 72.0% | 71.1% |
| OARSYS | 83.3% | 83.0% | 83.0% | 83.3% | 83.0% | 83.2% |
| FPLWK | 69.7% | **68.0%** | 69.4% | 69.7% | 69.5% | 69.8% |
| AFUEL | 93.7% | 93.6% | 93.5% | 93.7% | 93.7% | 93.6% |
| AIRSYS | 74.3% | 73.8% | **73.3%** | 74.3% | 74.3% | 74.3% |
| BATHEXCLU | 84.6% | 81.3% | 82.1% | 84.6% | 85.4% | 88.8% |

Source: U.S. Census Bureau, 2017 American Housing Survey

Next, we reran our simulations from section 5, this time including the BLKGAS auxiliary variable. We highlighted in bold where the standardized difference between the proportions of correct matches produced with the current method and the evaluation method was greater than 1.645. Table 8 provides the updated results.

From Table 8 we see that our evaluation methods overall improved the imputations in the fuel variables CFUEL, WASH, DRY, WFUEL, and HFUEL; or, cooking fuel, washing machine, clothes dryer, hot water fuel, and heating fuel, respectively. We saw some decreases in the quality of the imputation for OVEN, DISH, FPLWK, and AIRSYS; microwave oven, dishwasher, working fireplace and central air conditioner, respectively, when we produced clusters with the 13,000 responses to all module questions and mapped those clusters back to the 62,500 respondents to the equipment module questions. Those differences did not exist in the latter dataset. We note that the 13,000 cases did not represent all possible combinations of our auxiliary variables from the data set of 62,500; there were therefore cases dumped into a "don't know" stratum.

## 7. Conclusions

Cluster analysis can help us develop hot deck cells, though the reductions in within-pool variation we observed from clustering are similar to the current methods. All methods' clusters converge to the variability within the lowest level in our clusters, so improvements arise when we find auxiliary variables that reduce the within-cluster variability. The American Community Survey provided us with block-level estimates that we could modify into useful auxiliary variables. We found compelling evidence to suggest the imputation of number of units in multifamily buildings and fuels used in the housing unit can be improved, and consider it as an opportunity for future research to confirm these findings.

Whether clustering one variable at a time with CART, one module at a time, or several modules together, cluster analysis produced donor pools that are similar in usefulness as our current method. We had slightly more success iteratively grouping many clusters into one with hierarchical clustering than assigning cases to the nearest cluster with K-means. We did identify a potential pitfall of clustering several modules together, as our pool of eligible cases for clustering is smaller due to our need for 100 percent response to all variables. For the same auxiliary variables, multiple imputation overall imputed more reasonable values of the number of units in multifamily buildings than what cluster analysis produced. We observed differences in the number of bedrooms distributions of imputed values obtained by the clustering and multiple imputation methods for which we could not account.

Clustering methods can produce clusters with only one element. This can be a problem when we use the clusters to produce donor pools, because we would produce a pool that had only one donor. Recall we clustered on summaries, and applied the cluster assignments to all units. We need to ensure the final donor pools contain an adequate number of observations.

*Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or the Bureau of Labor Statistics.*

**Acknowledgements**

We thank Sean Dalby for providing us with the Multiple Imputation simulation results to allow us to compare methodologies.

**References**

Anderberg, M.R. (1973), *Cluster Analysis for Applications,* New York: Academic Press, Inc.

Andridge, R.H. & Little, R.J. (2010), A Review of Hot Deck Imputation for Survey Nonresponse. *International Statistical Review*, 78, 1, 40-64.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth

Dalby, S., Ash, S., Zha, K., Mulley, G. (2019), "Imputation in the American Housing Survey: Comparing Multiple Imputation with the Current Hot Deck Methods," *Proceedings of the 2019 Joint Statistical Meetings*.

MacQueen, J.B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkley Symposium on Mathematical Statistics and Probability*, 1, 281-297.

SAS Institute Inc. (2015), *SAS/STAT® 14.1 User's Guide.* Cary, NC: SAS Institute, Inc.

U.S. Census Bureau; American Community Survey (ACS), 2017 Five-Year Public Use Microdata Sample (PUMS).

Ward, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236-244.