Using Imputation Methods to Predict Listing Housing Unit Counts for Small Geographies[1]

Courtney Hill*, T. Trang Nguyen*, Timothy Kennel*
*U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-1912

Abstract
In designing the 2020 Post-Enumeration Survey sample, we conduct a simulation study by selecting repeated samples from a research frame of basic collection units. The post-enumeration survey uses two different housing unit counts – initial counts from the U.S. Census Bureau Master Address File and updated counts from listing – to construct measure of sizes for stratification and sample allocation. The initial housing unit count is available for every basic collection unit on the research frame, but not the listing count. It is too costly to list every basic collection unit on the research frame. In this paper, we compare two imputation methods to simulate the housing unit counts from listing. In the first method, we use sample and census data to model the listing housing unit counts. In the second method, we use sample data from the 2010 post-enumeration survey as donors to impute listing housing unit counts. We compare the results of the two imputation methods, and decide to use the imputation method that uses donors from the 2010 post-enumeration survey to create listing housing unit counts for the 2020 Post-Enumeration Survey research frame.

Key Words: Listing, Post-Enumeration Survey, Hot Deck Imputation, Prediction Model.

1. Introduction

The 2020 Post-Enumeration Survey (PES) is designed to assess the coverage of the 2020 Census by providing an independent estimate of the household population. The sample design for the 2020 PES is a three-phase stratified design that uses initial or updated housing unit counts to construct a measure of size for each phase for stratification. This design has a similar structure to that of the prior post-enumeration survey, the 2010 Census Coverage Measurement (CCM) (Moldoff, 2008). The first phase uses initial housing unit counts from the U.S. Census Bureau Master Address File to place each primary sampling unit into one of three size-strata: small, medium, and large. The first-phase primary sampling units are blocks or block-groups known as basic collection units in the 2020 PES and block clusters in the 2010 CCM. These two geographies do not have the same measures of size or geographical boundaries. The sample primary sampling units are then listed. The field staff canvass the entire primary sampling units to identify the location of all housing units and construct a list of housing units. Listing provides a frame of housing units and updated counts for subsampling in the later phases. The initial and listing counts could be

[1] Disclaimer: This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau. This paper meets all of the U.S. Census Bureau's Disclosure Review Board (DRB) standards and has been assigned DRB approval number DRB-B0003-DSSD-20180919.

very different, which can cause some primary sampling units to move from one stratum in the first phase to another in a later phase. These are known as stratum jumpers.

Stratum jumpers are a concern when designing the 2020 PES sample because the change in housing unit counts could potentially affect the housing unit sample sizes. If a basic collection unit is expected to have a small number of housing units, but it is found to have a large number, then the housing unit sample size would increase. This would affect workloads and cost estimates. We know that stratum jumpers existed in the 2010 CCM and the sample design took them into account. This needs to be done for the 2020 PES as well.

In designing the 2020 PES sample, we conducted a simulation study by selecting repeated samples from a research frame of basic collection units for the United States. Thus, we needed to first simulate the housing unit counts from listing for every basic collection unit in the nation. We wanted to simulate a research frame that will accurately reflect the attributes of the 2020 PES frame to allow us to better plan and design the 2020 PES sample. We aimed for the research frame to have some similar attributes as the 2010 CCM sampling frame such as the percentage of stratum jumpers and the relative size of strata in terms of primary sampling unit counts and listing housing unit counts. We had the initial housing unit counts from the Master Address File for every basic collection unit on the research frame. We investigated ways to simulate listing counts for every basic collection unit on the research frame using past data from the 2010 Census and listing housing unit counts of sample block clusters from the 2010 CCM.

We considered two commonly used methods of imputation, regression and hot deck to simulate the housing unit counts from listing for every basic collection unit in the United States. Our application of these imputation methods was different than how they are traditionally used. Every record on our research frame needed to be imputed, not just a small percentage of records. The regression method fit a regression model on the 2010 CCM sample data to predict listing counts for basic collection units on the 2020 PES research frame. The hot deck method randomly assigned a basic collection unit a listing housing unit count from a 2010 CCM block cluster of similar characteristics.

Durrant (2005) described regression imputation as fitting a regression model on the variable of interest using auxiliary variables that have values for the missing and non-missing records. In this paper, we investigated using regression imputation with historical data to predict the listing housing unit count for each frame record. Durrant (2005) also discussed random regression imputation, which added variability from a random draw to the predictions to produce imputed values. The random draw could be determined overall or in a subgroup from any distribution. The regression imputation discussed in this paper included a random draw from a normal distribution. We needed to add randomness to the predicted values to reflect the right level of differences between the initial and listing housing unit counts observed in the 2010 CCM to produce enough stratum jumpers.

For our research, we called the hot deck imputation method the cell-based method to reflect the narrow width of each imputation cell, which consisted of a single measure of size value or a small group of measure of size values. Durrant (2005) and Andridge and Little (2010) defined hot deck imputation as assigning the value of an observed record to a missing record. The observed record was called the donor and the missing record was called the recipient. The donors and recipients are grouped into imputation classes or cells using variables of interest that are available to both types. We used initial housing unit counts from the 2008 Master Address File and 2016 Master Address File to group donors and

recipients into cells of similar measure of size, respectively. The donor was randomly selected from the donor pool for each missing record within a cell. With this method we imputed a listing housing unit count for each frame record with past survey data.

We only had listing data similar to what we expect for the 2020 PES from the 2010 CCM sample. Andridge and Little (2010) discussed how having a limited number of donors can lead to over-usage, which can lead to a loss in precision and increased bias. This was the case with the cell-based method where we over-used the donors from a survey sample to impute for all records on the research frame. Also, we only had a limited number of variables that we could use in the regression imputation method because of the definitional differences in listing units of the past and current survey cycles.

Section 2 discusses the methodology and resulting models for regression imputation. Section 3 discusses the methodology for cell-based imputation. Section 4 provides the comparison of the two imputation methods and Section 5 provides the conclusions of both methods and a decision on the method for the simulation study.

2. Regression Imputation Methodology

Regression imputation fits a model on a variable of interest using auxiliary variables to produce imputed values. The variable of interest for this research was the listing housing unit count. The auxiliary variables studied included those that are known from our past research to be correlated with the response variable. The 2010 CCM block cluster sample had complete coverage for both response and independent variables for each sample block cluster. We used this data source to create model coefficients then applied them to the basic collection unit research frame to impute the listing housing unit count for each basic collection unit.

2.1 Listing Housing Unit Count Model

The response variable for the listing housing unit count model was the listing housing unit count for each block cluster i, y_i, in the 2010 CCM sample.

There were three independent variables considered for model fitting:

- 2010 Census housing unit count for each sample block cluster i, $census_i$.
- Size category for each sample block cluster i, $size_i$. The initial housing unit count from the 2008 Master Address File for each sample block cluster, MOS_i,was converted to a three-level categorical variable by applying the same definition used in the 2010 CCM sample design. The block clusters were grouped by size into three mutually exclusive categories: small (0 to 2 housing units), medium (3 to 79 housing units), and large (80 or more housing units).
- Tenure status for each sample block cluster i, $tenure_i$. Each block cluster was assigned a status of owner or non-owner based on the percent of non-owner population in the block cluster. A non-owner block cluster contains 40 percent or more of non-owner population based on 2000 Census data and an owner block cluster contains less than 40 percent of the non-owner population.

These were the same size and tenure definitions used in the 2010 CCM to stratify and select the block cluster listing sample. Size and tenure status are important design and estimation characteristics for both 2020 PES and 2010 CCM. The size stratification accounted for the known differential coverage between small block clusters that have few housing units and

large block clusters that have hundreds of housing units. Based on this stratification, it was assumed for model fitting that block clusters within the same size category were similar and had a linear relationship based on the initial housing unit counts. Block clusters that were in different size categories were assumed to be not similar and did not have a linear relationship. Tenure status stratification accounted for differential coverage between block clusters with high and low proportion of non-owner population.

A regression model was selected and fitted to predict y_i from the independent variables using the GLMSELECT procedure with the stepwise selection option and GLM procedure with the weight statement of the SAS®[2] software. The weights used in the regression model were the first-phase weight determined from selecting the block cluster sample multiplied by the second-phase weight obtained from subsampling small block clusters in the 2010 CCM.

The final listing housing unit count model with no intercept takes the form of

$$y_i = \beta_1 size_i + \beta_2(size_i \times census_i) + \beta_3(tenure_i \times census_i). \quad (1)$$

The estimated coefficients, standard errors, and p-values for the model parameters from the GLM procedure are listed in Table 1.

All estimated regression coefficients for the $size_i$ main effect were non-zero because no explicit intercept term was included in the model in Equation 1. The $size_i$ main effect was considered an implicit intercept with non-zero values (SAS, 2013, Parameter Estimates and Associated Statistics). Note that because the $census_i$ main effect was not included in the model in Equation 1, the estimated regression coefficients for the interaction of $size_i$ and $census_i$ are non-zero. Pasta (2011) discussed omitting a continuous main effect if there is an interaction between the continuous and categorical variable to simplify the interpretation of the coefficients. The coefficient of the interaction is the slope for each level of the categorical variable instead of the deviation from the slope when the continuous main effect is included. The example in Pasta (2011) showed that an interaction between a continuous and categorical variable could have only non-zero coefficients when omitting the continuous main effect. However, the coefficients of the interaction of $tenure_i$ and $census_i$ were not both non-zero. This was because the interaction of non-owner and $census_i$ was a linear combination of the other variables (SAS, 2013, Parameterization of PROC GLM Models).

The model had a high R^2 value, accounting for 95.7 percent of the variation in y_i.

All non-zero parameters are significant in Table 1 except for the small and medium size category, which have p-values of 0.1758 and 0.1502. A parameter was considered significant if the p-value was less than 0.10.

[2] Copyright © 2013 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Table 1: Block Cluster Listing Housing Unit Count Model Parameters

Parameter		Estimated Regression Coefficient	Estimated Standard Error	P-Value
Size	Small	0.54	0.40	0.1758
	Medium	0.49	0.34	0.1502
	Large	1.84	0.89	0.0392
Size*Census	Small	0.74	0.02	<0.0001
	Medium	0.92	0.01	<0.0001
	Large	0.93	<0.01	<0.0001
Tenure*Census	Owner	0.03	<0.01	<0.0001
	Non-Owner	0.00	.	.

Source: U.S. Census Bureau, 2010 Census Coverage Measurement.

The estimated coefficients from Table 1 were applied to the corresponding independent variables of the basic collection units to predict a listing housing unit count for each basic collection unit j, $\hat{y}_j^*$, on the research frame. The * notation indicates variables associated with basic collection units. The research frame was a list of basic collection units with initial housing unit counts from the 2016 Master Address File. It did not include basic collection units that were entirely water, in remote areas of Alaska, or in Puerto Rico. For this application to occur, the independent variables were constructed for each research frame basic collection unit. Since we did not have a current census housing unit count for the basic collection unit j, we estimated it from a model. We denoted this estimate by $\widehat{census}_j^*$ (see Section 2.2). Each basic collection unit j was categorized into one of three mutually exclusive size categories, $size_j^*$, by applying the same 2010 CCM size definition used for the block clusters to the initial housing unit counts from the 2016 Master Address File, MOS_j^*. Lastly, each basic collection unit j was assigned a tenure status (i.e., owner or non-owner), $tenure_j^*$, based on the 2010 Census data using the same tenure status definitions used for the block clusters. For example, a basic collection unit that has $\widehat{census}_1^*$ = 50, $size_1^*$= medium, and $tenure_1^*$ = owner would have a predicted listing count of $\hat{y}_1^* = 0.49 + 0.92 \times 50 + 0.03 \times 50 = 48$.

The 2010 CCM listing housing unit count y_i was sometimes very different than the MOS_i of the same block cluster i. This resulted in approximately 3 percent of stratum jumpers. However, the predicted $\hat{y}_j^*$ was more similar to the MOS_j^* for basic collection unit j, which lead to a smaller percent of stratum jumpers from the model predictions than what was observed in 2010. The standard deviation of $(y_i - MOS_i)$ was much larger than the standard deviation of $(\hat{y}_j^* - MOS_j^*)$, 52 and 7, respectively. This was because the 2010 CCM had some block clusters that had large differences between the initial and listing housing unit counts and it was difficult for the model to replicate these large differences. A standard deviation that is closer to zero indicates the predicted listing housing unit counts from regression imputation do not vary much from the initial housing unit counts.

We wanted the research frame to have a similar distribution of stratum jumpers as observed in the 2010 CCM data. For this reason, we added randomness to the predicted listing housing unit counts to allow for more stratum jumpers. The random component was

generated by drawing a random value for each basic collection unit j of group g from a normal distribution with mean of zero and standard deviation s_g as follows:

$$e_{g,j} \sim N(0, s_g).$$

The s_g is the standard deviation of the 2010 CCM sample predictions for each group g. There were 58 groups created using the initial housing unit counts. The first group contained block clusters with 0 initial housing units. The second group contained block clusters with 1 or 2 initial housing units. After the first and second groups, the range of the initial housing unit counts in each group changed depending on the number of block clusters. We tried to maintain at least 30 block clusters in each group.

The values from the random draw were added to $\hat{y}_j^*$ (i.e., the predicted listing housing unit count for each basic collection unit j) to produce the imputed listing housing unit count for each basic collection unit j, $\hat{y}_j^{imp*}$. Listing housing unit counts can not be less than zero so any $\hat{y}_j^{imp*}$ that was less than zero was forced to equal zero. This random draw approach was similar to what Durrant (2005) did for the random regression imputation. Durrant (2005) added a residual term to the predicted value from the regression. The residual term can be determined by drawing from a normal distribution.

The standard deviation of $(\hat{y}_j^{imp*} - MOS_j^*)$ for the random draw regression imputation was 21. This standard deviation is larger than the standard deviation of the regression imputation without adding the residuals. However, the random draw regression imputation standard deviation was still smaller than the 2010 CCM standard deviation value of 52.

2.2 Census Housing Unit Count Model

We considered using census housing unit counts from the 2010 Census and assigning them to each basic collection unit. However, the 2010 Census counts were from a different time period than the initial housing unit counts assigned to the basic collection units, which were from the 2016 Master Address File. For this reason, we did not use the 2010 Census housing unit counts directly; instead we used modeled census housing unit counts.

We used the 2010 CCM sample block cluster data to fit regression models. We fitted three separate models, one for each size category c (small, medium, and large) using the same definition in Section 2.1, using the GLM procedure with the weight statement in the SAS®[2] software. The continuous response variable was the 2010 Census housing unit count for each block cluster i of category c, $census_{c,i}$. The independent variable was the initial housing unit count for each block cluster i of category c, $MOS_{c,i}$, from the 2008 Master Address File. The multiplied first and second phase sampling weights from the CCM sample block clusters were used.

The three models with the size category subscript c and no intercept take the form of

$$census_{c,i} = \beta_c MOS_{c,i}. \quad (2)$$

The estimated coefficients, standard errors, p-values, and R^2 values for each model are shown in Table 2. The R^2 values were not corrected for the mean because the models did not contain an intercept. The parameter of each model is significant at the 10 percent level.

The small size category model has an extremely small R^2 value (see Table 2). The model accounts for only 1 percent of the variation in $census_{small,i}$ and therefore produces less accurate predicted census housing unit counts for the basic collection units. The differences between the initial and census housing unit counts for the small block clusters varied ranging from 0 to 443.

The models for medium and large block clusters have relatively high R^2 values, which account for most of the variation in $census_{c,i}$. In hindsight, we could have used $MOS_{c,i}$ instead of $census_{c,i}$ in the model in Equation 1 for the medium and large basic collection units because the estimated coefficients in Table 2 are close to one.

Table 2: Block Cluster Census Housing Unit Count Model Parameters

Model	Estimated Regression Coefficient	Estimated Standard Error	P-Value	R^2
Small	2.43	0.81	0.0027	0.01
Medium	0.99	0.01	<0.0001	0.76
Large	0.93	0.01	<0.0001	0.91

Source: U.S. Census Bureau, 2010 Census Coverage Measurement.

Within each size category c, the respective estimated coefficient from Table 2 was applied to the initial housing unit count of the basic collection unit j from the 2016 Master Address File, $MOS^*_{c,j}$, to produce a predicted census housing unit count for each basic collection unit j, $\widehat{census}^*_{c,j}$.

3. Cell-Based Imputation Methodology

Cell-based imputation used donor values to impute missing values in each imputation cell. In this research, the donor values were the listing housing unit counts from the 2010 CCM sample block clusters. The basic collection units on the research frame were the recipients of the donor values. The imputation cells were created by cross-classifying these characteristics:

- Small indicator based on initial housing unit counts (two categories: small and not small).
- Tenure status (two categories: owner and non-owner).
- Single or multiple measure of size based on the initial housing unit counts.

We chose initial housing unit count and tenure status because our past research showed these variables to be correlated with the listing housing unit count. They also can be constructed for the donor and recipient.

To form imputation cells for the donors, the block clusters were first grouped into two cells, small (0 to 2 initial housing units) and not small (3 or more initial housing units). The initial housing unit counts were from the 2008 Master Address File. Then, the block clusters in the not small cell were split into owner and non-owner based on 2000 Census data. This tenure status definition is the same definition used in the regression imputation and discussed in Section 2.1. Tenure status did not apply to the small cell to keep it consistent with the stratification used in the 2010 CCM sample design. There are also too few housing units in these block clusters for us to further subdivide the small cell.

Lastly, the imputation cells were formed using a single value of the measure of size for each cell, where possible. That is, block clusters in small, not small owner, or not small non-owner cell were subdivided into smaller cells of the same initial housing unit count. This allowed the pool of donors within a cell to have the same measure of size and the donor and recipient to have the same measure of size. This was desirable because the imputation counts needed to reflect the distribution of the observed 2010 CCM data, which occurred when the imputation cells were narrowly defined.

We were not able to create single measure of size imputation cells for all block clusters. The distribution of the block clusters in the sample was heavily skewed with the majority of the block clusters having smaller measures of size. We created single measure of size imputation cells for not small owner block clusters with 150 initial housing units or less and not small non-owner block clusters with 160 housing units or less. For block clusters that had more than this many initial housing units, we grouped every five measures of size together to form imputation cells. By grouping block clusters that had higher measures of size together, we tried to ensure enough donors in the imputation cell. However, we did not have a strict minimum number of donors in each imputation cell. The measures of size did not have to be consecutive when forming cells with multiple measures of size. For example, an imputation cell could contain five block clusters with non-consecutive measures of size of 151, 152, 155, 160, and 161.

Table 3 summarizes the three-level imputation cells formation. The imputation cells are sometimes referred to by the third-level name. We created 449 imputation cells. The average number of donors in each cell was 26 block clusters.

Table 3: Levels of Cell-Based Imputation Cells

First Level (Small Indicator)	Second Level (Tenure Status)	Third Level (Single or Multiple MOS)
Small (0 to 2 housing units)	Not Applicable	Single MOS: 0 to 2 housing units
Not Small (3+ housing units)	Owner	Single MOS: 3 to 150 housing units
Not Small (3+ housing units)	Owner	Multiple MOS: 151+ housing units
Not Small (3+ housing units)	Non-Owner	Single MOS: 3 to 160 housing units
Not Small (3+ housing units)	Non-Owner	Multiple MOS: 161+ housing units

MOS is measure of size.
Source: U.S. Census Bureau, 2010 Census Coverage Measurement.

The recipients (i.e., basic collection units) were grouped separately into the same number of imputation cells as the donors using the categorization definition from Table 3. The number of recipients in each cell ranged from 54 to 509,200 basic collection units. The basic collection unit initial housing unit counts were from the 2016 Master Address File and tenure status was created using 2010 Census data. The single measure of size imputation cells were formed similarly to the donor cells of the block clusters. However, the grouping of basic collection units into cells containing multiple measures of size were handled differently than the block clusters. Because we created the imputation cells for the donors based on the sample data, we expanded the imputation cells to cover every measure of size on the basic collection unit research frame. We included a basic collection unit in the imputation cell if it had a measure of size in the range defined for the block clusters. If a measure of size was not included in any of the imputation cell ranges, we included it in

the imputation cell of the closest preceding range. For example, if an imputation cell contained measure of size values of 200, 201, 203, 204, and 206 then the imputation cell of the basic collection unit would contain measure of size values of 200 through 206. This resulted in the imputation cells for the basic collection unit to contain more than five values of the measure of size.

Within each respective imputation cell, we selected a simple random sample with replacement of the block clusters. The sample size is the basic collection unit count of the same cell. For each block cluster selected, we used the listing housing unit count as the donor value. We randomly assigned each donor value to a recipient basic collection unit from the same imputation cell. These donor values became imputed listing housing unit counts for the basic collection units.

4. Comparison of Imputation Methods

We compared the results of each imputation method to the 2010 CCM because we wanted to choose the imputation method that produced similar attributes as the 2010 CCM. The two attributes we considered were the relative size of strata and the percent of stratum jumpers. First, we compared the distribution of basic collection units for each imputation method to the 2010 CCM weighted distribution of block clusters. Then, we compared the percent of stratum jumpers for each imputation method to the 2010 CCM weighted percent of stratum jumpers. Finally, the imputed listing counts were compared to the 2010 CCM weighted listing count. These comparisons were carried out separately for three size strata: small, medium, and large (block clusters and basic collection units).

We used the initial housing unit counts to group the basic collection units and sample block clusters into three size strata based on the first sampling phase stratification definitions as shown in Table 4. The first-phase size stratification definition for the 2020 PES using basic collection units is different than the definition for the 2010 CCM when using block clusters. The definition was changed to reflect the smaller size of the basic collection units compared to the block clusters. The 2020 PES definitions preserve the same proportion of frame housing units in these strata as the 2010 CCM.

Table 4: First Sampling Phase Size Stratification Definitions

Stratum	2010 Census Coverage Measurement Block Cluster Definition	2020 Post-Enumeration Survey Basic Collection Unit Definition
Small	0 to 2 housing units	0 to 2 housing units
Medium	3 to 79 housing units	3 to 57 housing units
Large	80 + housing units	58 + housing units

We examined the distribution of block clusters and basic collection units by the first-phase size stratification definitions using the initial housing unit counts as well as the listing housing unit counts. Table 5 shows the distribution of the block clusters for the 2010 CCM and the distribution of the basic collection units for the random draw regression imputation and cell-based imputation by the first-phase stratum.

Table 5: Percent of Block Clusters or Basic Collection Units by Stratum and Method

Stratum Based on Initial Housing Units	Stratum Based on Listing Housing Units	2010	RDR	Cell	RDR Absolute Difference	Cell Absolute Difference
Small	Small	19.26	7.05	11.76	12.21	7.50
	Medium	2.37	6.26	1.67	3.89	0.70
	Large	0.10	0.23	0.10	0.13	0.00
Medium	Small	2.13	5.11	3.48	2.98	1.35
	Medium	66.51	69.63	72.71	3.12	6.20
	Large	0.70	2.56	1.10	1.86	0.40
Large	Small	0.18	0.04	0.12	0.14	0.06
	Medium	1.42	1.67	1.43	0.25	0.01
	Large	7.32	7.46	7.61	0.14	0.29

2010 is 2010 Census Coverage Measurement weighted, RDR is random draw regression imputation, and Cell is cell-based imputation.
Source: U.S. Census Bureau, 2010 Census Coverage Measurement, Master Address File, and Geography Files.

The overall distribution of basic collection units for the cell-based imputation method was similar to the 2010 CCM weighted distribution of block clusters. Whereas, the distribution of the random draw regression imputation method was not as similar to the 2010 CCM weighted distribution as the cell-based imputation method. We wanted the distribution of the imputation method to be similar to the 2010 CCM weighted distribution because we assumed the 2020 PES distribution will be the same as the 2010 CCM. In particular, the distribution percentages for the small strata for the random draw regression imputation method were not close to the 2010 CCM compared to the cell-based imputation method. The absolute difference of the random draw regression imputation is larger than the absolute difference of the cell-based imputation, 24.72 and 16.51, respectively. This was an indication that the cell-based imputed values were closer to the 2010 CCM values.

We were also interested in the overall percent of stratum jumpers. A basic collection unit or block cluster was a stratum jumper if it was classified into a stratum based on the initial housing unit counts, but re-classified to a different stratum based on the listing housing unit counts. For example, a basic collection unit with 50 initial housing units and 60 imputed listing housing units is a stratum jumper from medium to large stratum. The 2010 CCM weighted overall percent of stratum jumpers was 6.91, which was calculated by subtracting the percentages of non-stratum jumpers from 100 (100 – [19.26+66.51+7.32] (see Table 5)). The overall percent of stratum jumpers for the cell-based and random draw regression imputation methods were 7.92 and 15.86, respectively. The overall percent of stratum jumpers for the cell-based imputation was closer the 2010 CCM weighted percent.

After reviewing the model parameters and distributions for the small basic collection units with zero to two initial housing units, we decided the random draw regression imputation did not yield acceptable imputed listing housing unit counts. The estimated listing housing unit count regression model coefficients for small basic collection units did not adjust $\widehat{census}_j^*$ (census housing unit count) by much (see Table 1 for the values of the coefficients). If a small basic collection unit had a $\widehat{census}_j^*$ that was similar to the initial housing unit count, the listing housing unit count would be similar as well and it would not

be considered a stratum jumper. This led to no stratum jumpers for small prior to the random draw. This is contrary to the medium and large categories. The random draw increased the percent of stratum jumpers for small. However, it increased the listing housing unit counts for non-stratum jumpers by too much. Because of this reasons, regression imputation for small basic collection units was not considered for further analysis in this section.

We also calculated the percentage of basic collection units or block clusters that were stratum jumpers by size (medium and large) and tenure status. The percent of stratum jumpers for block clusters was the sum of the sampling weights for block clusters that were re-classified to a different stratum divided by the sum of the sampling weights for the block clusters in the original stratum. This calculation assumed the block clusters that the sampling weights represent would also be stratum jumpers. For basic collection units, the percent of stratum jumpers was calculated by dividing the sum of the basic collection units that were re-classified to different stratum by the sum of the basic collection units in the original stratum on the research frame.

Figure 1 shows the percent of stratum jumpers by imputation method for medium and large strata by tenure status (defined in Section 2.1).

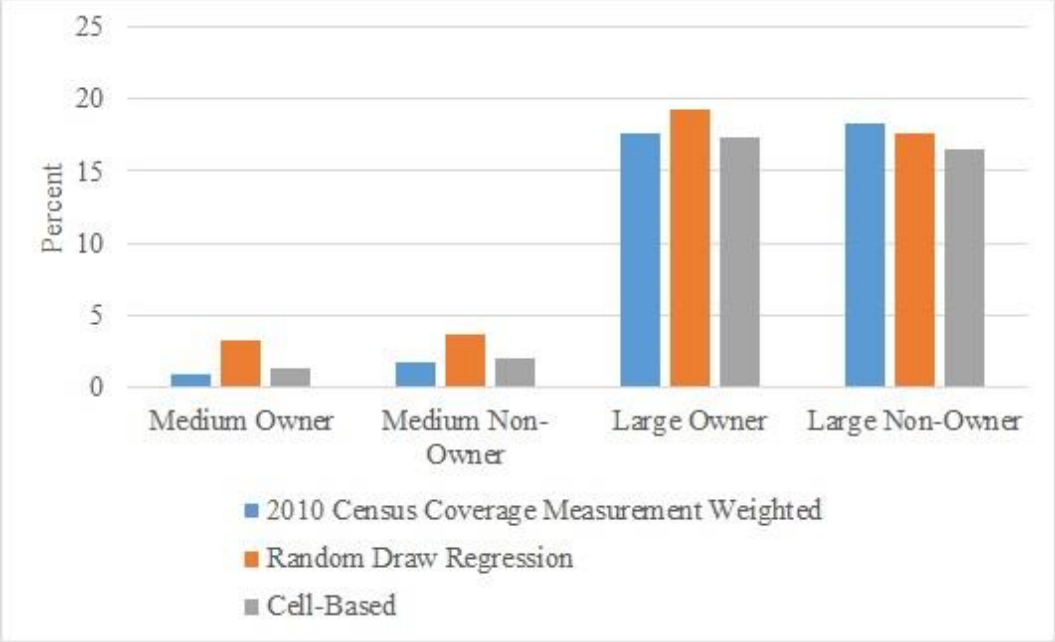

Figure 1: Bar Chart of the Percent of Stratum Jumpers

Source: U.S. Census Bureau, 2010 Census Coverage Measurement, Master Address File, and Geography Files.

In Figure 1, both medium owner and non-owner strata have a small percent of stratum jumpers for each imputation method, similar to the 2010 CCM weighted percent. We expected this level of stratum jumpers because the number of stratum jumpers in the medium stratum was small relative to the overall number of block clusters or basic collection units in the stratum, driving the percent of stratum jumpers down. The medium stratum comprised approximately 69 percent of block clusters and 77 percent of basic

collection units. However, the cell-based imputation percent of stratum jumpers were closer to the 2010 CCM weighted than the random draw regression imputation.

The percent of stratum jumpers was much higher for large strata than for medium strata. The number of stratum jumpers was high relative to the number of block clusters and basic collection units in the stratum, which was approximately 9 percent for both. Cell-based imputation and random draw regression imputation had a similar percent of stratum jumpers to the 2010 CCM weighted percent. The percent of stratum jumpers in the large owner stratum for cell-based imputation was closer to the 2010 CCM weighted than the random draw regression imputation. The opposite was true in the large non-owner stratum.

Overall, the distribution of stratum jumpers for the cell-based imputation was closer to the 2010 CCM weighted distribution than the random draw regression imputation. This was a reason for us to consider cell-based imputation for the listing housing unit counts.

Lastly, we used the listing housing unit counts to group the basic collection units and sample block clusters into two strata: subsampling and non-subsampling. The subsampling stratum contained

- basic collection units with 58 or more listing housing unit counts or
- block clusters with 80 or more listing housing unit counts.

The housing units in these basic collection units or block clusters were eligible for subsampling in the third sampling phase. All remaining basic collection units or block clusters were in the non-subsampling stratum and their housing units were retained in sample with certainty.

Both designs (2020 PES and 2010 CCM) oversampled large basic collection units or block clusters in the first sampling phase and subsampled their housing units in the third sampling phase to support a set target number of housing unit interviews.

For non-stratum jumpers, large basic collection units or block clusters were categorized into the subsampling stratum. For stratum jumpers, if the basic collection units or block clusters were categorized as large based on the initial housing unit counts, but had fewer than 58 or 80 listing housing units, respectively, then these were part of the non-subsampling stratum. Conversely, small or medium basic collection units or block clusters with 58 or more listing housing units or 80 or more listing housing units, respectively, were categorized into the subsampling stratum.

Table 6 shows the listing housing counts for each imputation method by stratum compared to the 2010 CCM weighted listing housing counts. Since the 2010 CCM weighted listing housing unit count was our baseline, we expected the better imputation method to produce imputed listing housing unit counts for the non-subsampling stratum and subsampling stratum that are similar to the 2010 CCM weighted counts. We calculated the percent difference between the listing housing unit counts of the imputation methods and the 2010 CCM weighted listing housing unit counts. We used this calculation to determine whether the imputation method was practically different. The threshold for being practically different was 15 percent.

Table 6: Listing Housing Unit Counts by Imputation Method and Stratum (in Thousands)

Stratum	2010 Census Coverage Measurement Weighted	Random Draw Regression	Cell-Based
Non-subsampling	**72,344**	**63,461**	**70,731**
Small	1,352	NA	1,205
Medium Owner	53,115	45,817	50,740
Medium Non-Owner	14,707	13,950	15,777
Large Owner	1,983	2,298	1,914
Large Non-Owner	1,187	1,396†	1,095
Subsampling	**60,435**	**65,109**	**63,820**
Small	746	NA	775
Medium Owner	2,238	7,477†	5,072†
Medium Non-Owner	1,675	2,666†	2,208†
Large Owner	33,074	29,638	30,488
Large Non-Owner	22,702	25,328	25,277

NA Not applicable.
† Practically different.
Source: U.S. Census Bureau, 2010 Census Coverage Measurement, Master Address File, and Geography Files.

For small strata, we only compared the imputed listing housing counts of the cell-based imputation to the 2010 CCM weighted listing housing unit counts, omitting the random draw regression method because the fitted model was not producing reasonable imputed listing housing unit counts. The small non-subsampling and subsampling cell-based imputation counts were not practically different than the 2010 CCM weighted listing housing counts.

For medium and large strata, we compared the imputed listing housing unit counts of both imputation methods to the 2010 CCM weighted listing housing unit counts. The cell-based imputation counts in Table 6 are mostly not practically different to the 2010 CCM weighted counts; only two strata are practically different. The random draw regression imputation counts were more often practically different to the 2010 CCM weighted counts than the cell-based imputation counts, though by only one stratum.

5. Conclusions

We imputed listing housing unit counts using two different imputation methods, random draw regression and cell-based imputation. The random draw regression imputation method used predicted housing unit counts from a regression model plus a random draw from a normal distribution to produce imputed listing housing unit counts. The cell-based imputation method used donor values from the 2010 CCM sample block clusters to impute the listing housing unit count values for the recipient basic collection units on the research frame.

We saw that the random draw regression imputation did not produce acceptable model parameters for the small size category with few housing units. For the medium and large size categories, both cell-based imputation and random draw regression imputation were similar to the 2010 CCM observed data in most comparisons. However, the cell-based

imputation method had an overall distribution of basic collection units closer to the baseline 2010 CCM distribution than the random draw regression imputation. Both imputation methods had similar percentages of stratum jumpers, though the cell-based imputation percentages were closer to the 2010 CCM overall. The random draw regression imputation compared to the 2010 CCM data was more often practically different than cell-based imputation to the 2010 CCM data.

Based on these comparison results, we decided to use the cell-based imputation method to create the listing housing unit counts for the 2020 PES research frame to use in our design simulation study. We accepted that there are limitations to using sample data to impute for every frame record. In future work, we could investigate using housing unit counts that are not from a sample. By using housing unit counts from a larger set of data, we could reduce the limitation of over-using donors in the cell-based imputation method. We could also make adjustments to the random draw regression imputation to produce more similar listing housing unit counts to the 2010 CCM. Two possible adjustments include exploring the addition of an error term in the listing housing unit count model and perhaps eliminating the census housing unit count models. By adding an error term, the random draw would not be needed. In hindsight, instead of modeling the census housing unit counts, we could use the initial housing unit counts as a covariate in the listing housing unit count model because the census housing unit count model coefficients for medium and large are approximately 1. We also saw that the random draw regression imputation failed for the small category. It might be useful to instead use an imputation method like cell-based for the small size category, but use a random draw regression imputation for the medium and large size categories.

Acknowledgements

Thanks to Krista Heim, Laura Davis, and Scott Konicki for their valuable input and comments on this document.

References

Andridge, R. and Little, R.J.A. (2010). "A Review of Hot Deck Imputation for Survey Non-response." Int Stat Rev. 78(1): 40–64. DOI: 10.1111/j.1751-5823.2010.00103.x.

Durrant, G. (2005). "Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review." ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton.

Moldoff, M. (2008). "The Design of the Coverage Measurement Program for the 2010 Census." DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-B-07, Retrieved from https://www.census.gov/coverage_measurement/pdfs/2010-B-07.pdf.

Pasta, D. (2011). "Those Confounded Interactions: Building and Interpreting a Model with Many Potential Cofounders and Interactions." SAS Global Forum 2011.

SAS Institute Inc. 2013. SAS/ACCESS® 9.4 Interface to ADABAS: Reference. Cary, NC: SAS Institute Inc.