The Development of Truth Decks for the 2010 Census Count Imputation Research

Todd R. Williams
United States Bureau of the Census, Washington DC 20233 [1]

Abstract

When the Census Bureau's decennial census is processed, housing units' can be missing their status (whether the units exist), occupancy (whether the units are vacant), and person counts within the units. For the 2010 Census, we are testing several methods for imputing these items. For comparing the imputation methods, we form truth decks using housing unit data from the 2000 Census. We fit logistic regression models to the data to find which variables (predictors) have the greatest impact on the propensity for each of the items to be missing. We use these predictors to create pseudo-strata within which some of the housing units on the truth decks are flagged as missing either their status, occupancy, or person counts based on the corresponding propensities to be missing. The various imputation methods can be tested using the truth decks by comparing the imputed values to the original values for the items flagged as missing.

Keywords: truth deck, logistic regression, predictor variables, missing item propensity

1. Introduction

In support of the 2010 Census, we are researching several count imputation methods. Count imputation involves filling in missing data for items that affect the final census population. These items include whether a housing unit exists, whether a unit is occupied, and the number of persons residing at a housing unit. To test the accuracy of the several methods of imputation, we need some way to compare the results. Since we do not know what the results should be, due to the fact that the values for the items are missing, we have to search for alternative benchmarks to use in making comparisons. One method is to set reported values for the items to missing so that we can apply the imputation methodologies to these items and compare the results with the true data. Because the true item values for all of the housing units remain on the data file with some of the housing unit items flagged to be treated as if they are missing, we refer to the file as a truth deck.

Because a housing unit's status (existing or not existing), occupancy (vacant or not vacant), and person count can be missing, we flag as missing some or all of these items for some of the housing units in the truth deck. The person count depends on the housing unit's status and occupancy. If the housing unit is found to not exist, there are no persons present and the count is zero. If the housing unit does exist, it can be vacant or not vacant. If the housing unit is vacant, then there are no persons present and the count is zero. If the housing unit is not vacant, we need to impute a nonzero person count.

The purpose of our work is to find an accurate and efficient way to create a truth deck that will benefit the 2010 Census count imputation research. In the following sections we will discuss the methodology we used to try to reach our goal, provide analysis of the results, and draw conclusions from what we accomplished.

2. Methodology

Since we retain the reported value for each item (housing unit status, occupancy and person count) on the truth deck, we create a flag value for each item that is to be treated as missing. Our flagging of items as missing is not done completely at random throughout the entire truth deck. Instead, it is done completely at random for housing units within defined subgroups called pseudo-strata. We randomly set the housing unit item flags within pseudo-strata because the propensity for the items to be missing can greatly vary from one pseudo-stratum to another and we want to reflect this when creating the truth deck.

[1] Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological issues are those of the author and not necessarily those of the U.S. Census Bureau.

Not only do the propensities for the items to be missing vary between pseudo-strata, but how the pseudo-strata are defined varies by state. As a result, we create a separate truth deck for each one of the fifty states and Washington, D.C. The data that we use is housing unit data taken from the 2000 Census.

When we are flagging the reported data items as missing, we have to be aware of the relationships between the items. Because a housing unit's occupancy is dependent on its status (occupancy is not applicable when the housing unit does not exist) and a housing unit's person count is dependent on its occupancy and status (person count is zero if either the housing unit is vacant or does not exist), we begin with the flagging of housing unit status. Based on the dependency, a housing unit's occupancy is automatically flagged as missing if its status is flagged as missing. As a result, we flag housing unit occupancy only on those units that do not have status flagged as missing. Likewise, a housing unit's person count is automatically flagged as missing if the unit's status or occupancy is flagged as missing. Our last step is to flag person counts for the housing units in which status and occupancy have not been flagged. We perform this whole procedure independently for each of the fifty states and Washington, D.C.

Due to the dependency between housing unit status, occupancy, and person count, we must form a separate set of pseudo-strata for each of the items. There are 153 sets of pseudo-strata; this is based on a separate set of pseudo-strata for each item (3 items) within each state and Washington, D.C. (50 states plus D.C.). Each item's pseudo-strata are defined by variables used to predict the probability that the item is missing for a housing unit.

We begin by fitting logistic regression models on all of the housing unit data within a state in order to estimate the probability that the status is missing for each housing unit. For the propensity for occupancy to be missing, we fit logistic regression models on the data from all housing units that have a reported status of existing. We do not use housing units which are listed as not existing because they do not have an applicable value for occupancy and will not be flagged regarding occupancy. We also do not use units which are missing a value for status. A missing value for status means that the housing unit has a 100% probability of missing a value for occupancy, and we are concerned with estimating the probabilities only for units with unknown probabilities. To estimate the probability of a missing person count for each housing unit, we fit logistic regression models on all housing units that have a reported status of existing and a reported occupancy designation of occupied. We are not interested in using housing units that are listed as not existing or vacant because they will automatically have person counts of zero.

Our procedure for creating a truth deck consists of two phases. The first phase finds the variables that define the pseudo-strata. The second phase calculates the propensity for each item to be missing within each pseudo-stratum. For every housing unit, flag values are set for each item based on the corresponding propensities to be missing.

2.1. Methodology Phase 1

In the first phase, we fit logistic regression models to 2000 Census data to find variables that best predict the probability of housing unit status, occupancy, and person counts to be missing. A separate model is fit for each item. The basic form of the model we use is the following:

$$\ln\left\{\frac{p}{1-p}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

where p is the probability (propensity) that the item is missing, x_1, x_2 and x_n are dichotomous first, second and n^{th} predictor variables and β_0, β_1, β_2 and β_n are the intercept, first, second, and n^{th} effect parameters being estimated by fitting the model to the data.

By finding a model that best predicts the probability that an item is missing, we can use the predictor variables to define the pseudo-strata in which the housing unit items are flagged. By using the variables that are found to be the best predictors, we can be sure of having pseudo-strata that exhibit the greatest differences in terms of the propensity for an item to be missing.

In fitting a logistic regression model to predict the probability for an item to be missing, we use a forward selection method (Stokes, 1995) where each possible predictor variable is modeled independently. When we fit the model, we look for the variable with the largest value for the Wald chi-square statistic. A larger value of this statistic indicates a predictor variable that is more statistically significant in estimating the probability. (We found almost all variables to be statistically significant at the $\alpha = 0.05$ level.) The variable that is found to be the best predictor based on the largest value for the Wald chi-square statistic is placed permanently in the model as the first predictor variable. Each of the remaining possible predictor variables are then individually tested with the first predictor by fitting the models using only main effects (no interaction between the two predictors). The variable that adds the most to predicting the probability for an item to be missing (based on having the largest value for the Wald chi-square statistic) is permanently added to the model as the second predictor variable. This procedure continues until we have seven predictors of the probability for an item to be missing.

Once we have seven predictor variables, we proceed to the second stage where we narrow the set of variables to four. We accomplish this by fitting four different logistic models to predict the probability for an item to be missing. In each model, we have three fixed predictor variables. These variables are determined to be the best predictors based on their Wald chi-square statistics taken from the model fitting performed in the first stage. The fourth predictor variable for each model is one of the remaining four out of seven that is not fixed in the models. In fitting these models, we not only include the main effects of each predictor variable, but also the two-variable interactions between all four variables. We use a stepwise selection procedure instead of a forward procedure, which allows for any predictor variable that becomes insignificant with the addition of other variables to drop out of the model. (It is possible for one or more of the variables that are fixed in the four models to drop out of one or more of the models.) We choose up to four predictor variables from the model which best fits the data based on the Hosmer-Lemeshaw goodness-of-fit test (Hosmer, 1989).

We wanted to fit logistic regression models that used a stepwise selection method to test not only the main effects, but also the two-variable interactions of all of the possible predictor variables in one stage. However, because of the large amount of computer run time and computer resources (such as disk space) that this process consumed for the largest states, we needed an alternative method. Consequently, we developed the two-stage model fitting procedure that we have described. By using the forward selection method in the first stage of the model fitting, we could stop the fitting procedure when seven predictor variables were found. This reduced processing time because in some states many more than seven predictor variables were found to be statistically significant at the $\alpha = 0.05$ level even though they did not affect the estimated probability as much as the first seven variables. Since we were going to look at two-variable interactions in the second stage of the model fitting and to choose only four final predictors, we decided to use only the main effects when fitting the logistic regression models in the first stage. In the second stage, the reduction of the number of predictor variables to four allowed us to feasibly fit, for even the largest states, a set of logistic regression models using the stepwise selection method while testing both main effects and two-variable interactions.

We were able to test the results of this two-stage model against the results obtained from the single-stage model derived from stepwise selection on all main effects and two-way interactions. We performed the testing on some of the smaller and medium sized states. The results were almost identical to the extent that we were comfortable about using the two-stage way to fit the models.

2.2. Methodology Phase 2

At this point, we have a set of four predictor variables for the probability of housing unit status, occupancy, and person count to be missing. Since these predictor variables are dichotomous, the number of pseudo-strata will be sixteen (2^4). In some cases, two predictor variables are mutually exclusive which means that it is impossible for a housing unit to be in a pseudo-stratum created by a particular cross-classification of the two predictor variables. For example, we have instances where both being in a small multi-unit structure (2 to 9 units) and being in a large multi-unit structure (10 or more units) are variables used to create the pseudo-strata. Obviously, a housing unit can not be part of both, so we remove any pseudo-strata that have this cross-classification. By doing this, we reduce our number of pseudo-strata from sixteen to twelve. In other cases, we find that only three important predictor variables are found, so the number of pseudo-strata is eight (2^3).

Using the 2000 Census data for each item (housing unit status, occupancy, and person count), we calculate the total number of housing units within each pseudo-stratum and the total number of units within each pseudo-stratum that are missing the item before the 2000 Census imputation is done. We calculate the propensity for the item to be missing, within each pseudo-stratum, by dividing the number of units missing the item by the total number of units.

The next step is to create the truth deck by setting flag values for each housing unit. Before this can be done, we remove all housing units that have one or more of the items (housing unit status, occupancy or number of persons) listed as having been imputed for the 2000 Census. (We only want housing units which have reported data for the truth deck because only the reported data should be used as benchmarks in testing the imputation methods.)

We begin the flagging procedure first by deciding if status is needed to be flagged as missing for the housing units. Once this is performed, we decide if occupancy needs to be flagged as missing for those housing units that are not flagged as missing status. (If a housing unit is flagged as missing status, then by default it is also missing occupancy and person count.) Finally, we decide if the housing units need to have person counts flagged as missing for those units not flagged as missing status or occupancy. For any one of the items, we determine that a housing unit should be flagged as missing that item if a random number selected from a uniform distribution between zero and one is less than or equal to the unit's propensity for that item to be missing. All housing units within the same pseudo-stratum will have the same propensity for the item to be missing. Since the housing units' propensities for the items to be missing depend on the units' pseudo-strata, the items are not flagged completely at random over the entire truth deck, but they are flagged completely at random within each pseudo-stratum.

We replicate the flagged data on the truth deck one hundred times. Since the flags are set at random, each replication will provide a different set of flagged data. With the replicated data, we have the ability to estimate such statistics as means and variation due to imputation.

3. Results

For fitting the logistic regression models, we have come up with two sets of possible predictor variables. The first set includes variables that are available for all housing units even if they are missing status, occupancy, or person counts. We refer to these variables as operational variables and they are listed in Table 1. They are dichotomous with a possible value of either 1 (Yes) or 0 (No).

Table 1: Operational Variables

Variable	*Description: Housing unit…*
FT_FUP	…received a follow-up form.
IN_NRU	…is in a non-response follow-up universe.
IN_CIU	…is in a coverage improvement follow-up universe.
CIU_LATE	…is a late addition in a coverage improvement follow-up universe.
TEA_GRP2	…is in a non-mailback area.
SM_MULT	…is in a small multi-unit structure (2 – 9 units).
LG_MULT	…is in a large multi-unit structure (10 or more units).
NO_ADDR	…has no address location.

Our second set of variables is designed to provide demographic information that is not directly available for housing units missing status, occupancy, or person counts, but can be obtained from the unit's surrounding area. The area that we select is the block group which is the smallest practical area since some blocks may have only one or two housing units. We define each block group variable based on an operational or demographic characteristic and look at the proportion of housing units having the characteristic within the block groups. If a housing unit is in a block group that contains a high proportion for the characteristic (the block group is in the 90th percentile for the state), it has a value of 1; otherwise, it has a value of 0. Table 2 gives a description of the block group variables.

Table 2: Block Group Variables

Variable	*Description: Housing unit is in a block group with a high proportion of…*
B_MSTAT	…housing units with imputed status.
B_MOCC	…housing units with imputed occupancy.
B_MCNT	…housing units with imputed person counts.
B_VAC	…vacant housing units.
B_SMULT	…small multi-unit structures (2-9 units).
B_LMULT	…large multi-unit structures (10 or more units).
B_RENT	…rented housing units.
B_SINGLE	…not married householders.
B_NCHILD	…householders with no children.
B_MCHILD	…householders with many (4 or more) children.
B_FEMALE	…female householders.
B_YOUNG	…young householders (age less than 26).
B_OLD	…old householders (age greater than 64).
B_HISP	…Hispanic housing units.
B_WHITE	...Caucasian housing units.
B_BLACK	…African-American housing units.
B_ASIAN	…Asian housing units.

After fitting the logistic models and creating the pseudo-strata, we are interested in seeing which predictor variables are used the most often. (Only the best predictors of the probability for an item to be missing are used in defining the pseudo-strata for that item.) Since the same variable can be used for each item in each of the fifty states and the District of Columbia when creating the separate truth decks, the maximum number of times it can be used for defining the pseudo-strata is 153 (3 items x 51 states). In Figure 1, we show the total number of times each possible predictor variable is used (for variables used 5 or more times) and the number of times it is used per item in creating pseudo-strata for calculating the propensities for the items to be missing.

We would expect to see the variables B_MSTAT, B_MOCC, and B_MCNT to be used to define the pseudo-strata pertaining to calculating the propensities for housing unit status, occupancy, and person counts to be missing, respectively. From Figure 1, we see that this is the case. However, none of the other block group variables appeared to be influential on a regular basis in predicting the probabilities that the items are missing.

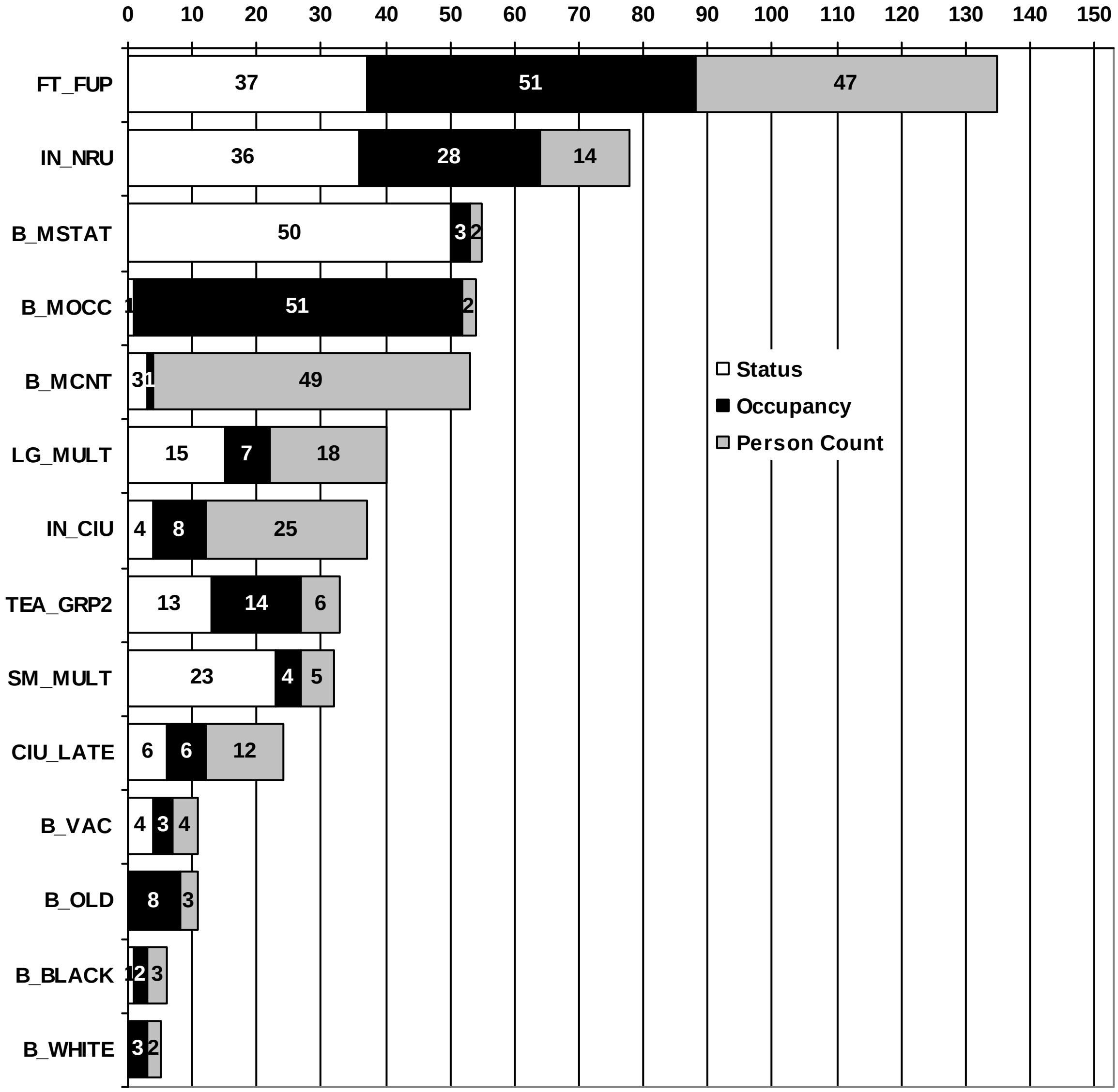
The variable that indicates that a housing unit is receiving a follow-up form (FT_FUP) and the variable that indicates that a housing unit is in a non-response follow-up universe (IN_NRU) appear to be good and consistent predictors for each of the items. We see, to a lesser extent, the variable that indicates if a housing unit is in a non-mailback area (TEA_GRP2) and the variables that indicate the number of units in the housing unit's building (SM_MULT and LG_MULT) contribute to predicting the probabilities that the items are missing.

We are surprised that the variable indicating that a housing unit does not have an address (NO_ADDR) is not used once in creating the pseudo-strata for any of the items. More work is needed to discover the reason for this.

Besides looking at which variables are used the most often in defining the pseudo-strata discussed above, we also calculate the propensities for housing unit status, occupancy and person counts to be missing within each pseudo-stratum based on the flagged data from the truth decks. We can do this by using data from one replicate or, more accurately, by calculating the propensities for each of the one hundred replicates and averaging them. Our resulting propensities should be very close to those calculated using the 2000 Census data. For each one of the states and Washington, D.C., we see that nearly all the propensities are very close for all three items.

However, we see that there are a few instances that the above does not hold true. These are a few pseudo-strata that contain a very small number of housing units in which every or almost every unit has at least one of the three items (status, occupancy, person counts) imputed for the 2000 Census. According to our procedure, all or almost all of these units will be removed from the truth deck (since units with non-imputed values are needed) which will leave either no housing units or just a few housing units remaining in the pseudo-strata. In the extreme case where every housing unit has an item imputed within a pseudo-stratum, the probability for the item to be missing is 100%, but a propensity for the item to be missing does not exist on the truth deck because all of the housing units within that pseudo-stratum have been removed. As a result, a comparison of propensities can not be made for the pseudo-stratum. Even though we perform the flagging procedure on these cases (when there are housing units in the pseudo-strata), we are not concerned about them because in most imputation processes that group units, the small pseudo-strata will most likely be collapsed into larger pseudo-strata. We are more concerned at this point with the propensities within the more conventional pseudo-strata. Within our data, the average sized pseudo-strata are considerably larger than the aforementioned small pseudo-strata. If we collapsed what remains in the small pseudo-strata with regular sized pseudo-strata, the resulting change in the propensity for the item to be missing is negligible for the regular sized pseudo-strata.

Figure 1. Number of Times a Predictor is used in Defining a Set of Pseudo-strata



4. Summary and Conclusions

We have created what we refer to as truth decks for each of the fifty states and Washington, D.C. using data from the 2000 Census. The truth decks consist of housing units in which the status, occupancy, and person counts for the units are all reported. In order to create benchmarks to test various methods of imputation for these items, some of the housing units will have one of these items flagged as missing so that values can be imputed and the imputed values then compared to the reported values. We randomly flag the items as missing within pseudo-strata that are created for this purpose. The random flagging of an item is based on the propensity for the item to be missing within the given pseudo-stratum. The variables that define the pseudo-strata are predictors of the probability of an item to be missing and are determined by fitting logistic regression models to the 2000 Census data. Once the pseudo-strata are created, we replicate the setting of the flags one hundred times.

When creating the truth decks, we were faced with a limited amount of time in which to produce them. Based on this, we chose to fit logistic regression models to predict the probabilities of each item being missing and let the models select the variables that best predict these probabilities. Starting with an intuitive set of potential predictor variable, we use programs and selection algorithms to find the best set of predictors. We were also able to cut down the processing time by fitting more simplified models in two stages, without sacrificing the reliability of the models to predict the probabilities.

We find several aspects of this work that can be researched more thoroughly. One question is whether or not a common set of predictor variables can be used to define the pseudo-strata for all of the states. We assumed that because of the differences in demographic characteristics among the states (mostly rural states versus predominately urban states, etc.), that there would be a variety in the sets of variables used in defining the pseudo-strata. However, we see that most of the possible predictor variables are seldom selected. With a smaller set of possible predictor variables from which to choose, we might be able to find a pair of variables that are highly correlated in such a way that if we were to replace one with the other as a predictor variable, the fit and predictive ability of the model changes very little. This could lead us to develop a common set of predictor variables.

Finally, we performed one way of providing benchmarks for testing imputation results. Our truth decks appear to provide what we require, but there are other ways to approach the problem. Our method of creating the truth deck is also not the only way of flagging data as missing. Further research can involve comparing different methodologies.

References

Hosmer, D.W., Jr. and Lemeshaw, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.

Stokes, M. E., Davis, C.S., and Koch, G.G. (1995), *Categorical Data Analysis Using the SAS System*, Cary NC: SAS Institute Inc.