IMPUTING PERSON AGE FOR THE 2000 CENSUS SHORT FORM: A MODEL-BASED APPROACH

Todd R. Williams, U.S. Bureau of the Census*
U.S. Bureau of the Census, Washington, D.C. 20233

Key words: Census Item Imputation, Multiple Regression, Variance Estimation

I. Introduction

The 2000 Census short form will collect demographic, household and person item information for each occupant of every household in the nation. Most of the data is collected by having the respondent of the household fill out and mail in the Census form. In some cases, an enumerator or interviewer will have to visit the household in order to obtain the information. Even if an enumerator has visited a household, one or more of the household or person items can be missing for an individual either from omission or failure of an item value to meet predetermined consistency checks. When a person's age is missing, the imputation method used for the 1990 Census short form involves a hot-deck procedure which imputes a value using data from the nearest household that has the same characteristics as the household containing the person with the missing age (Census, 1994). The purpose of our paper is to show possible improvements that can be observed when using a model-based approach for imputing missing person age for the 2000 Census short form. This paper will concentrate solely on the missing person age portion of the household and person item imputation system we are testing at the Census Bureau (Thibaudeau, et al., 1997). Using 1990 Census data, we will compare the imputations derived by using our modeling methodology to those created using the 1990 Census methodology. In the comparison, we will show that our method helps preserve some of the multi-variable characteristics found in the data. We will also demonstrate the ability to estimate variances associated with the imputed ages which is not currently available with the 1990 Census methodology.

Imputation is performed separately on data collected by each district office (DO). The United States contains 550 DOs with each DO representing approximately 300,000 to 700,000 individuals. The DOs are divided into tracts. A tract is a geographically contiguous region consisting of approximately 1,500 to 2,000 households. We perform imputation for each DO using tract level information when needed. For this study, we use three DOs from the 1990 Census. The first DO covers most of Bergen County, New Jersey. We are interested in this DO because it contains individuals from an urban area, but does not have a high percentage of minorities. The second DO covers Sacramento, California and is of interest to us because it is used in the 1998 test Census. The third DO covers parts of Los Angeles, which we are using because it contains a high percentage of minorities.

The remainder of this paper is divided into four sections. Section II describes the modeling approach we used for imputing missing age. We compare the results of the imputation using the 1990 Census method and our method in section III. In section IV, we discuss estimating imputation variances using our method. Section V provides our conclusion.

II. The Procedure

We use four multiple regression models to predict the values of missing person ages. The models are fitted to the complete data found in the DO. The complete data is comprised of households in which all household and person item responses are listed as not missing. We impute the age of the householder before we impute the age of anyone else in the household. This allows us to use the age of the householder to predict a value for any other missing age. The first two models predict a value for missing householder age. The third model predicts a value for the missing age of a child or stepchild of the householder. We use the fourth model to find a value for the missing age of all other persons in the household. The general form of our multiple regression models is

$$AGE = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad ,$$

where the betas are the parameter estimates, the x's are the predictor variable values and k is the number of predictor variables. The set of predictor variables we use for each model is fixed. This set of variables produced the best possible fit to the complete data when we tested data from five different DOs. These DOs include the three we use in our analysis, a Florida DO which has a higher average householder age, and a Kansas DO containing individuals from rural areas.

For predicting the age of the householder, the age of another person in the household is the strongest predictor variable. Only one person's age is used in our first model and this person is determined by the following order: 1. spouse, 2. oldest child, 3. youngest parent, 4. unmarried partner, 5. first listed roommate with a non-missing age and 6. oldest grandchild. The age of a person is used only if there does not exist a person higher in the

order or the ages of all persons higher in the order are missing. We determined this order by comparing the fit of the model using each predictor age independently and ranking the ages based on the best fit. The other predictor variables we use in our first model include the sex of the householder, the number of persons in the household and the tenure of the householder (owner or renter).

When none of the persons listed above are available or have an age value that can be used to predict the age of the householder, we use our second model. The most important predictor variable for this model is the average age of the complete data householders by tract. Within the tract, we further separate the average ages by tenure, number of persons in the household, sex of the householder, whether or not a spouse of the householder exists in the household, and whether or not an enumerator visited the household. Other predictor variables in our second model indicate if a grandchild, roommate, or parent of the householder is present in the household.

Our third model predicts the age of a child or stepchild of the householder. The strongest predictor variable in this model is the age of the householder which, if originally missing, has been imputed. To ensure that the ages differ when there is more than one child within the household with a missing age, we create a predictor variable that provides the order in which the child is listed in the household in relation to the other children. The number of persons in the household, the sex of the householder, and an indicator variable indicating the presence of a householder spouse are the other predictor variables in this model.

Our last model predicts the age of a person in the household who is not the householder or a child of the householder. As with our third model, the strongest predictor variable is the age of the householder. The other predictor variables include the number of persons in the household, an indicator variable indicating the presence of a householder spouse, and indicator variables that state the relationship to the householder of the person whose age is being predicted.

We avoid having the same imputed age for all persons with the same set of characteristics by adding random error to the predicted age. We accomplish this by randomly selecting a residual from the distribution of residuals obtained by fitting the model, where the residual is the observed age minus the predicted age. The randomly selected residual is added to the predicted age to produce the final replacement value for the missing age. To prevent imputing an outlying value for age, we select the residual from the middle eighty percent of the distribution.

III. The Comparison

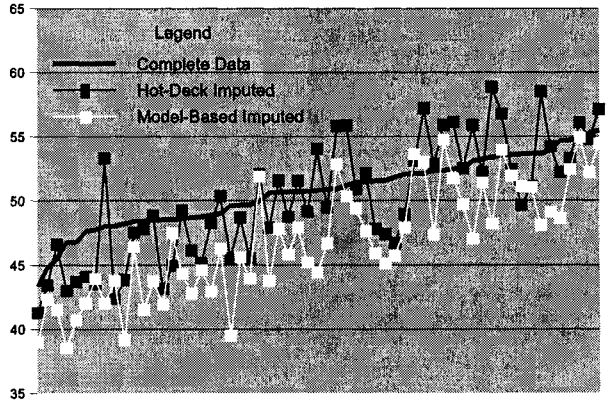
Our analysis begins by comparing the average imputed ages derived using our model-based imputation method to those derived from the 1990 Census hot-deck imputation method. In our procedure, householder age is used to predict the values of all other persons with a missing age; therefore, we will concentrate our following comparison only on missing householder age. We note that the same results are obtained for the spouse and oldest child of the householder ages. The comparison is shown using data from the Bergen County, New Jersey DO. Similar results can be found using data from the Sacramento and the Los Angeles DOs.

We give in Table 1 the average age of the householders found in the complete data households and the average age of the hot-deck and the model-based imputed householders for the entire Bergen County DO. We display in Figure 1 the same comparison charted by tract. The tracts are ordered from the lowest to the highest average age based on the complete data households. Only those tracts that have at least ten householders with an imputed age are displayed.

Table 1. Average Age of Householders

Complete Data	Hot-Deck Imputed	Model-Based Imputed
50.6	48.8	46.0

Figure 1. Average Age of Householders by Tract



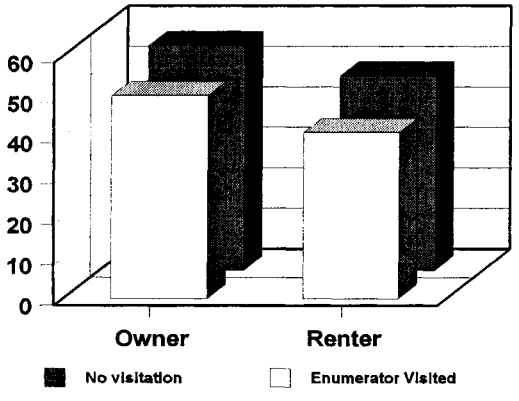
We see from Table 1 and Figure 1 that our model-based method provides the lowest average householder age. We are not concerned that our imputed ages appear to be lower than that of the complete data householders. Our suspicion is that the proportion of imputed householders that exhibit certain key characteristics are higher than the proportion of complete data householders. If the householders with this set of characteristics show a lower average age than the average

age for the complete data householders, the lower average for imputed age can be explained. What we find interesting is that we also exhibit lower ages for average householder age when compared to the results from the 1990 Census hot-deck method.

Our next step is to find the key characteristics that are influencing our lower imputed ages. The first important consideration is that we are using either one of two models to find a predicted householder age depending on the availability of another person's age in the household. For the Bergen Co. DO, ninety-three percent of the imputed householder ages are derived from the model we use when no other person age within the household is available as a predictor variable. This is because a large number of imputed householders either live by themselves or the ages are also missing for the other persons in the household. As a result, we focus our analysis on the set of householders whose age is imputed from this model.

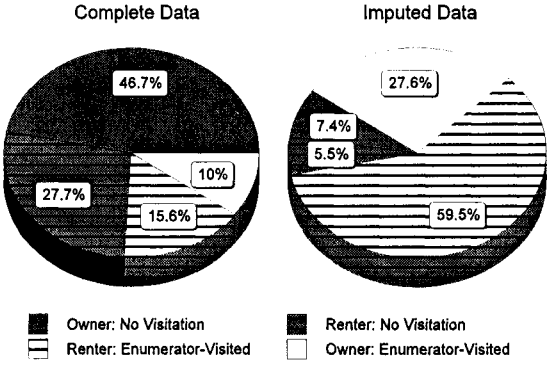
This model uses the average householder age by tract to predict a missing householder age. To improve the fit of the model, we provide the average householder age within each tract by several characteristics which include tenure and the possibility that the household is visited by an enumerator. By crossing tenure with enumerator visitation we develop four groups whose average householder age for the complete data householders are shown in Figure 2.

Figure 2. Average Age of Complete Data Householders for Bergen County



We see from Figure 2 that the average householder age is lower for renters and enumerator-visited householders. If the percentage of renters and the percentage of enumerator-visited householders are higher for the imputed data when compared to the complete data, we can expect a lower average imputed householder age. In Figure 3, we show that these percentages are higher for the imputed data.

Figure 3. Percentage of Householders



Since most of the householders with an imputed age are visited by an enumerator, we show the average age of enumerator-visited householders for the complete data households and for the imputed householders using both imputation methods in Table 2. This table displays total, owner and renter householder average ages for the entire Bergen Co. DO. Figures 4, 5 and 6 show the same information by tract for enumerator-visited total, owner and renter householders respectively. As in Figure 1, only those tracts that have at least ten householders with an imputed age are displayed.

Table 2. Average Age of Enumerator-Visited Householders

	Complete Data	Hot-Deck Imputed	Model-Based Imputed
All	44.5	47.3	44.3
Owner	50.1	51.3	50.5
Renter	40.9	45.4	41.4

Figure 4. Average Age of Enumerator-Visited Householders by Tract

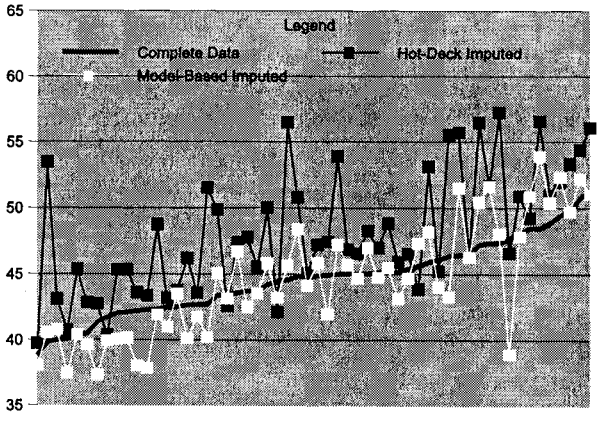


Figure 5. Average Age of Enumerator-Visited Owner Householders by Tract

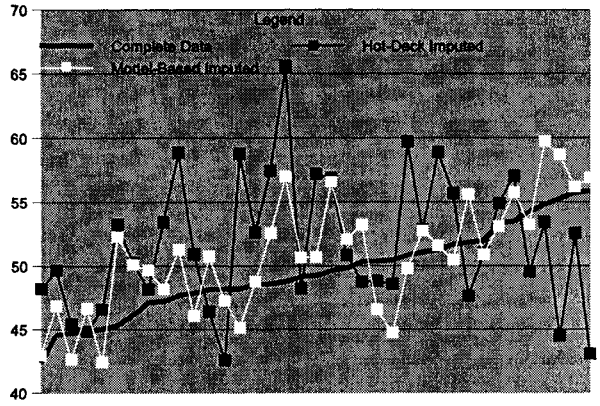
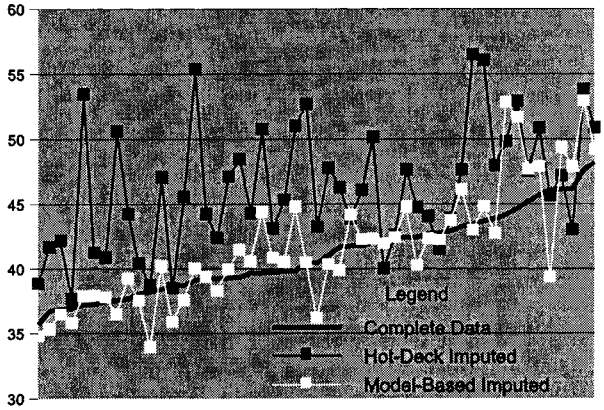


Figure 6. Average Age of Enumerator-Visited Renter Householders by Tract



We see from Table 2 and the figures that for the enumerator-visited households, our model-based imputation method provides average householder ages that are closer to the complete data than those of the hot-deck method. This is especially true for the householders who are renters. The hot-deck method did not include tenure and enumerator visitation as a characteristics for finding a donor value that would replace the missing householder age. Since enumerator-visited households with no available person age for predicting householder age contain 81% of all householders with missing age and 68% of these householders are renters, we feel that the lower average householder ages calculated after imputation by our model-based method are justified based on our findings.

Our next area of concern is to determine how the difference in the average householder ages from the two imputation methods affects the overall average household age. This average age includes both the complete and the imputed householders. Table 3 shows the average age of the householder after imputation by both methods for all householders in each of the three DOs that are used in our analysis. We can see that for each DO the average householder age does not change significantly by changing the imputation methods. This is mostly due to the small percentage of imputed householders.

Table 3. Average Householder Age after Imputation

District Office		Bergen Co.	Sacramento	Los Angeles
Total number of householders		106,307	215,335	135,548
Percent with imputed age		4.3%	2.8%	10.1%
Average age using:	Hot-deck method	50.6	46.0	47.4
	Model-based method	50.5	46.0	47.1

When we look at a set of householders with specific characteristics in an area where there is a higher percentage of householders with imputed age, we see a much greater difference between the overall average householder ages. We give an example in Table 4 which displays the average householder age after imputation by both methods for householders in enumerator-visited households that contain only the householders. The averages are displayed for tracts, one from each DO, which have higher rates of imputation than their corresponding DOs. Here the differences in the average householder ages is very noticeable. The higher average householder ages shown previously in Figure 4 for the hot-deck method seem to have an effect on the overall average ages for the tracts listed in Table 4.

Table 4. Average Householder Age after Imputation for Enumerator-Visited Householders Living Alone

Tract		Bergen Co. Tract 23401	Sacramento Tract 4201	Los Angeles Tract 234000
Total number of householders		205	81	178
Percent with imputed age		47.3%	17.3%	35.5%
Average age using:	Hot-deck method	45.5	44.2	49.0
	Model-based method	39.8	40.5	45.1

IV. The Variances

Our model-based imputation procedure allows us to estimate variation in average ages due to the imputation of missing age values at both the DO and tract level. We estimate the variances by calculating two components of variation. We add the two components together to derive the overall variance estimate and take the square root of this estimate to get the standard error.

The first component of variation we refer to as the model component. In this component, we estimate the variation in age averages due to using the predicted ages from our multiple regression models as imputed values. This component captures the variation associated with the parameter estimates and the variation between the predicted and the observed values when fitting the models to the set of complete data households. Our model component variance estimate $s^2_{model}(\bar{Y})$ is calculated as follows (Neter, et al.,1990):

$$s^2_{model}(\bar{Y}) = \left(\frac{MSE}{n^2}\right)\mathbf{1}'\cdot[\mathbf{I}+\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']\cdot\mathbf{1}$$

where MSE is the mean square error from the model,

n is the total number of observations (both complete and imputed),

$\mathbf{1}$ is a column vector of m ones where m is the number of imputed observations,

$\mathbf{I}$ is a m by m identity matrix,

$\mathbf{A}$ is a m by r matrix containing the r predictor variable values for the m imputed observations and

$(\mathbf{X}'\mathbf{X})^{-1}$ is the correlation matrix of the parameter estimates from the model.

We always fit the models to the set of complete data from the entire DO. As a result, the mean square error and the correlation matrix of parameter estimates used in the calculation of the model component is the same for estimating imputation variances at the DO and the individual tract levels. The model used in the calculations depends on the type of person age being averaged. For instance, we use different models for imputing the age of a spouse of the householder and the age of a child of the householder. For imputing householder age, our imputation procedure uses two models. We calculate the model component variance estimate for average householder age by performing the above calculation separately for each of the two models. When making this calculation for each one of the models, we only use the group of imputed householders that correspond to the model. The m in this case is the number of householders whose imputed age is derived from model 1 for the model 1 calculation and the number of householders whose imputed age is derived from model 2 for the model 2 calculation. The n is the total number of householders. Once we have the variance component estimates from the two models, we add them together to obtain the model component variance estimate for average householder age.

The second component of variation we refer to as the simulation component. We use this component to estimate the variation in average ages caused by adding randomly selected residuals to the imputed age values. We also capture the variation caused by imputing a person's missing relationship to the householder prior to imputing missing age (Thibaudeau, et al., 1997). We produce our simulation component variance estimate by replicating the imputation process 1,000 times. Our estimate $s^2_{sim}(\bar{Y})$ is then calculated as follows:

$$s^2_{sim}(\bar{Y}) = \frac{1}{1000}\sum_{i=1}^{1000}(\bar{Y}_i - \bar{\bar{Y}})^2$$

where $\bar{Y}_i$ is the average age for replicate i and $\bar{\bar{Y}}$ is the average of the 1,000 replicate average ages.

In Table 5, we show estimates of the standard errors due to imputation for average householder, spouse of the householder and child of the householder age. The table displays standard errors for both the Bergen County and the Sacramento DOs.

Table 5. District Office Level Standard Errors Due to Imputation

		Average Age	Standard Error	Percent Imputed
Bergen Co.	Householder	50.5	0.0110	4.3%
	Spouse	47.0	0.0090	3.1%
	Child	15.3	0.0076	2.9%
Sacramento	Householder	46.0	0.0055	2.8%
	Spouse	44.2	0.0053	2.2%
	Child	11.6	0.0042	2.9%

We can see from Table 5 that there is very little variation due to imputation in the average ages. The major reason is that the number of persons with an imputed age is a very small percent of the total number of

persons. In Table 6, we show for two tracts, tract 23401 from the Bergen County DO and tract 4906 from the Sacramento DO, the same standard error estimates. Here the percentages of persons with an imputed age are higher and the standard errors are higher.

Table 6. Tract Level Standard Errors Due to Imputation

		Average Age	Standard Error	Percent Imputed
Tract 23401	Householder	47.2	0.1609	12.7%
	Spouse	44.2	0.1902	12.6%
	Child	14.7	0.1258	4.7%
Tract 4906	Householder	41.1	0.2834	13.4%
	Spouse	38.8	0.2093	10.2%
	Child	10.0	0.1290	9.3%

V. Conclusion

We have developed our model-based approach to imputing for missing person age on the Census 2000 short form with the expectation that we would be able to make improvements in maintaining multi-variable relationships found in the data. Based on our comparisons with the 1990 Census hot-deck method, we believe that the improvements are evident. By using our model-based approach, we can directly determine which variables have the greatest influence on a person's age. Once the variables are determined, we can use them along with the parameter estimates taken from the models to predict values for the missing ages. These predicted ages should exhibit the multi-variable relationships found in the nonmissing data. We have shown this to be true with the relationship between age, tenure, and enumerator visitation. Householders who are renters and are visited by an enumerator have lower average ages than the overall population of householders. A large portion of the householders with missing age are enumerator-visited renters; therefore, the average imputed age for householders should be lower than the overall average householder age. We have also stated that the age of the householder is the most important predictor for finding a missing age for another person in the household. Consequently, we would expect that imputing lower householder ages would also produce lower imputed ages for other household members.

We have also seen that the average age of all of the householders, complete and imputed together, can be noticeably lower after using our model-based imputation method than the average age after imputation by the hot-deck method. We have found this to be true for certain groups of householders within tracts that have a relatively high percentage of imputed ages. Once again, the higher average householder age produced after imputation using the hot-deck method is at least partially caused by omitting the relationships between age, tenure and enumerator visitation.

In addition to predicting values for missing ages, we have demonstrated that we can estimate the variation in average ages due to the imputation of these missing values. We are able to calculate variance estimates derived from using our multiple regression models and from other sources such as the adding of randomly selected residuals to the predicted ages. The variance estimates themselves appear to be very small, mostly due to the low percentage of imputed ages.

We feel that our model-based method is an improvement over the hot-deck method for imputing a value for the missing age of person on the 2000 Census short form. As we have shown, our method can determine and preserve the multi-variable relationships between the age of a person and other available information in the data. Improvements can be made in the existing hot-deck procedure by implementing the model-fitting techniques shown in our method when finding the characteristics to use when matching a person with missing age to the nearest neighbor.

REFERENCES

Neter, J., Wasserman, W. and Kutner, M. H. (1990), *Applied Linear Statistical Models, Third edition*, Boston: Irwin.

Thibaudeau, Y., Williams, T. and Krenzke, T. (1997), "Multivariate Item Imputation for the 2000 Census Short Form," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 371-376.

U.S. Bureau of the Census (1994), "Summary of the 1990 Decennial Census Edit and Imputation Procedures for the 100% Population and Housing Items," memorandum from John H. Thompson to Susan M. Miskura.

* This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage dicussion.