MODEL EXPLICIT ITEM IMPUTATION FOR CENSUS 2000

Yves Thibaudeau, U.S. Census Bureau, Washington, DC 20233

1. Background

We have invested a good deal of research effort to develop a model-based imputation methodology that provides a practical alternative to the nearest neighbor hot-deck methodology developed for the 1990 census. We have made good progress and we have set a benchmark for our item imputation procedure using the 1990 census data for the district office (DO) of Sacramento for purpose of evaluation. We chose this particular DO since it is one of the sites where we are currently conducting our Census dress rehearsal and we look forward to validating our benchmark with dress rehearsal data.

Throughout this short summary we review the specific imputation contingencies for the item imputation in 1990 for the Sacramento DO and we recall the base principles of the 1990 imputation methodology. Then we point out a systematic inconsistency in the imputation of the Hispanic origin item, and we explain how and why the 1990 methodology produced this inconsistency. Finally we introduce our model-based imputation procedure and we show how it can finesse around this pitfall. These results make up the first benchmark for our methodology.

We have identified 215,214 households after edit with values for the household items in Sacramento in 1990. The household items are tenure, sex, race, and Hispanic origin of the householder. The rates of imputation for the household items in Sacramento in 1990 are 2.3 %, 0.6 %, 1.4 %, and 6.8 % for the tenure, sex, race, and Hispanic origin (HO) items respectively. Although no missing items should be ignored HO is clearly the single most potentially damaging source of bias. In all 14,516 households did not report their HO in 1990 (table 1). Of this number 1019 imputations resulted in Hispanic HO's. This amounts to 7.0 % of all the households subject to imputation of HO. The remaining 93 % of the imputations generated non-Hispanic HO's. What is puzzling is that 27,348, or 13.6 % of the households who did report HO on their census form declared a Hispanic HO (table 1). If we exclude the possibility of a non-report bias this implies that households reporting the HO item are two times more likely to be Hispanic than those who don't.

Under this scenario Hispanics are only half as likely to omit the HO item relative than the rest of the population. However, this goes against other behavioral evidence. For instance consider the cases of households who did not return their census form. An enumerator must visit these households and request a value for each item. According to our records 17.0 % of the households enumerated under these circumstances are Hispanic (table 2). That is a higher proportion than among the mail-returns. Based on this observation one might conjecture that Hispanics tend to be more response-averse than Non-Hispanics, and thus a dramatically lower rate of omission for the HO item among Hispanics as suggested above is unlikely. Of course we can not draw conclusions from this observation alone. Nevertheless, we think it is likely that the Hispanic households were imputed at a rate lower than their actual prevalence, and we present corroborating evidence as well as a rationale for this assertion.

2. An Example of Structural Bias: The Imputation of Hispanic Origin in Sacramento

In 1990 we developed a nearest neighbor procedure for the imputation of HO conditional on race. In other words, when HO is not reported, the procedure retrieves the HO of the last household with a reported HO whose race agrees with that of the household with unreported HO. The vast majority of Hispanics belong to two broad race categories (table 3): 38.0 % of reported Hispanic households who reported the race item are White while 57.0 % are "Others " (mostly Mexican). Thus when HO is unreported for a household with a reported race "Others", the nearest neighbor procedure retrieves the last record corresponding to a household of race "Others" with reported HO, and this household becomes a **donor** of its HO status to the household with unreported HO.

The problem is that the overall rate of "others" drops from 7.0 % (table 4) of the population among the households with reported HO to less than 1 % of the population of households with unreported HO. In other words the race contributing almost 60 % of the Hispanic households is practically eradicated among the households with unreported HO. Thus the donors for imputing HO come from races accounting for less than half of the Hispanics which explains the dramatic drop among the households subject to imputation of

HO.

We must emphasize that the category "Others" is very loosely defined. In Sacramento most of the individuals in this race are Latinos and may decide legitimately to check the "white" box. However, for the sake of the argument, let's assume that there is no ambiguity in the delineation of the race categories. Then we can find another plausible explanation for the drop of the rate of Others. We found that the value of HO is strongly correlated with the report status of the race item. That is 79.2 % (table 5) of the households who do not report the race item are of Hispanic origin. A simple ratio estimator based on the report status of the race item would raise the proportion of Hispanic households among the unreported cases from 7 % to more than 10 %.

3. The Model-Explicit Approach

We give a summary of our model-explicit approach. The idea behind our approach is to take advantage of the same supporting information required by the nearest neighbor approach, but to simulate the unobserved process of the items so that the imputation procedure is completely transparent. The nearest neighbor approach stipulates that a household with unreported race be imputed with the race of the neighbor. This approach is not transparent in the sense that an observer quickly realizes that the imputed households always have the same race as their neighbors. That is not very realistic.

We replace the deterministic rules of the nearest neighbor approach with more natural rules, namely probabilistic rules. So when a household omits to report race, and that household lives next to a black household, the odds of being Black are greater than if it were next to a White household, but the final imputation depends on other factors, such as the racial mix of the tract. Race of the neighbor is only one predictor in our model. The predictive power of the race of the neighbor is adjusted to the local tract. There are two other predictors adjusted at the tract level: tenure of the neighbor, and HO of the neighbor. These three predictors are the most correlated variables with their respective response variables. There are also correlations between the response variables which are included in the model. For instance, race is negatively correlated with HO, i.e. blacks are less likely to be Hispanic.

The basic structure of our model is log-linear. Interaction factors between variables are limited to second order interactions. We include the interaction factors between the predictors and their corresponding response variables; that is race of the neighbor interacts with race of the householder, HO of the neighbor with HO of the householder, and tenure of the neighbor with tenure of the householder. In addition all the second order factors between the response variables are included. The model is hierarchical, that is all first order effects are included.

Our model is dichotomous. Race is either Black or non-Black, HO is either Hispanic or not Hispanic, and tenure is owner or renter. When more races, origins or tenure statuses need to be imputed, we can repeat our procedure within the broad categories delineated through the first imputation, or we can revert to the nearest neighbor approach. This incremental approach avoids the pitfalls created by having to deal with too many race categories at once. For instance, merging the White and "Others" race categories protects us from the structural bias of the nearest neighbor procedure when imputing HO. A second imputation pass can be set up to separate the White and the "Others" race categories.

Unlike the nearest neighbor approach, repeating the imputation operations will most likely not result in the same imputations. The model approach simulates the stochastic process underlying the unreported items. The counts related to the items are random variables. This approach gives rise to two errors: a model error and a simulation error. Table 6 gives the distributional characteristics of the number of Hispanic households accounting for the uncertainty generated by the cases with unobserved Hispanic origin. This distribution is called the predictive distribution, and it reflects the model error, that is the likelihood of a departure from the count of Hispanics predicted by the model. Graph 1 is a histogram of 600 simulations of the count of Hispanics under the predictive distribution. In other words, graph 1 is a graphical estimate of the predictive distribution.

The simulation error is created by our imputation scheme. We deliberately create noise while simulating the unreported items. The total error is the error resulting from the combination of both error sources (model, and simulation). The total error is the distance between the predicted count of Hispanic households and the imputed count generated by our methodology. Table 7 gives the distributional characteristics relating to the total error and graph 2 is a histogram of 600 simulations providing an estimate of the distribution of the total error. It should be observed that the expected

number of Hispanic households given by our imputation procedure and the number obtained in 1990 are more than 15 standard deviations apart in terms of the distribution of the actual count (table 6).
It is therefore extremely unlikely that the 1990 number was produced by this distribution. We estimate that we failed to account for at least 600 Hispanic households in Sacramento in 1990.

Accordingly, we estimate that the bias in the number of Hispanic households we reported in the census for Sacramento in 1990 is around 600, and this is a conservative estimate. Note that at the same time, the standard deviation for the error of the new method is around 40 (table 7). Furthermore, preliminary analysis seems to indicate that the standard deviation for the 1990 hot deck is of the same magnitude. Therefore, by all accounts the 600 Hispanic households shortfall is very significant. In addition, this shows that, when using the nearest neighbor hot deck, the major concern should be the bias not the the standard deviation or the variance. The bias is an order of magnitude larger than the standard deviation. Any future use of the hot deck in censuses, and in surveys absolutely requires an evaluation and a correction of the potential bias. Ignoring the problem can lead to serious flaws in the ensuing analysis.

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 discussion.

Table 1. Distribution of the Hispanic Origin Item vs Report Status for all Households

	Non-Hispanic Households	Hispanic Households	Row Totals
Hispanic Origin Item Reported	86.4 %	13.6 %	200698
Hispanic Origin Item Imputed	93.0 %	7.0 %	14516
Grand total			215214

Table 2. Distribution of Hispanic Origin vs. Mail-Return Status for the Cases with Reported Hispanic Origin

	Non-Hispanic Households	Hispanic Households	Row Totals (Reported HO only)
Mail-Returns	87.9 %	12.1 %	138756
Non-Mail Returns	83.0 %	17.0 %	61942
Total			200698

Table 3. Distribution of Race vs. Hispanic Origin for the Cases with Reported Race and Hispanic Origin Items

	Non-Hispanic Households	Hispanic Households	Grand Total
White	77.0 %	**38.0 %**	
Black	11.4 %	1.3 %	
American Indians	1.1 %	1.5 %	
A. P. I	10.3 %	2.2 %	
Others	0.2 %	**57.0 %**	
Column Totals	172933	25763	198696

Table 4. Distribution of Race vs. Report Status of Hispanic Origin for the Cases with Reported Race Item

	Hispanic Origin Item Reported	Hispanic Origin Item Not Reported
White	71.9 %	70 %
Black	10.1 %	17.9 %
American Indian	1.2 %	1.3 %
A.P.I.	9.2 %	9.9 %
Others	**7.5 %**	**0.86 %**
Total	198696	13567

Table 5. Distribution of Hispanic Origin vs. Report Status of the Race Item for the Cases with Reported Hispanic Origin

	Non-Hispanic Households	Hispanic Households	Row Totals
Race Item Reported	87.0 %	13.0 %	198696
Race Item Not Reported	20.8 %	79.2 %	2002
Grand Total			200698

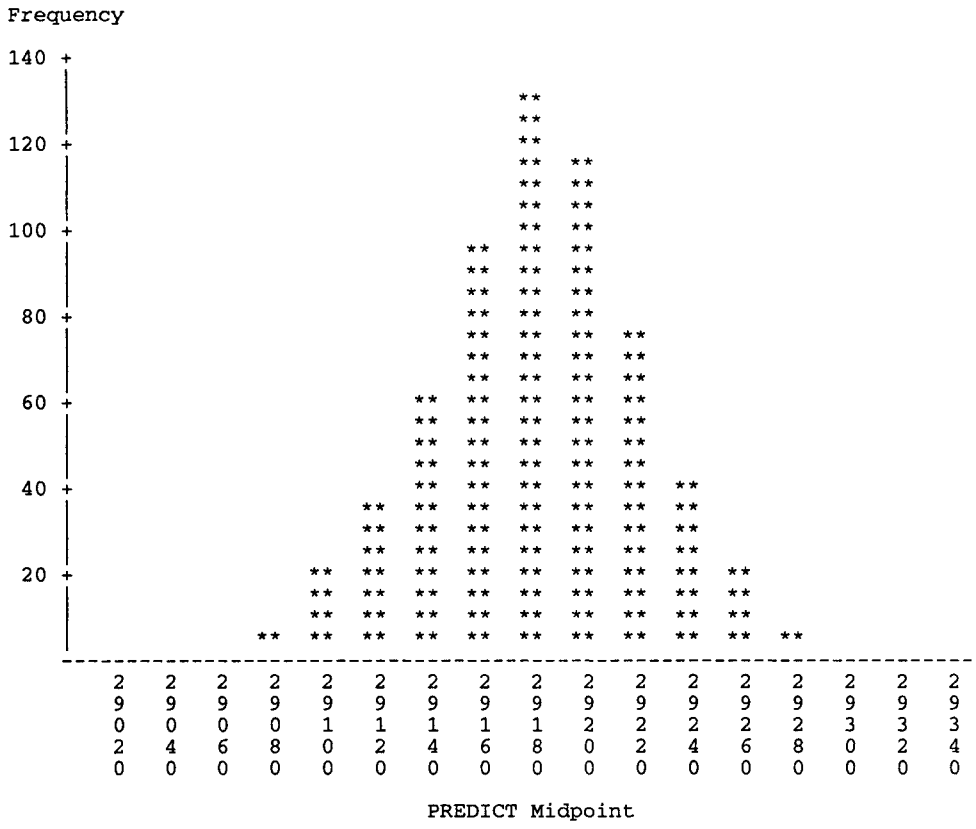
Table 6. Estimated Distributional Characteristics for the Number of Hispanic Households Accounting for the Cases with Unreported Hispanic Origin

Reported in 1990	Mean	Mode	Median	Std. Error
28367	29182	29194	29182	40.0

Table 7. Estimated Distributional Characteristics for the Error on the Number of Hispanic Households with Model-Based Imputation

Reported in 1990	Mean	Mode	Median	Std Error
0	6.5	7	7	55.5

Graph 1. - Frequency of the Predicted Hispanic Population (600 Simulations)



Graph 2. - Frequency of the Error between the Imputed and Predicted Populations