Research to Improve Census Imputation Methods: The Plan to Examine Count and Item Imputation[1]

Arthur R. Cresce Jr., Sally M. Obenski and Gary B. Chappell
U.S. Census Bureau, Washington, DC 20233-9200
Gary B. Chappell, U.S. Census Bureau, Washington, DC 20233-9200

Abstract

This paper provides background on the issue of missing information in the Census and the long-standing use of the traditional hot deck methodology. It discusses the emergence of a major interdivisional project to examine alternative methodologies, including the use of administrative records and spatial analysis. An overview of the methods is provided, as well as the rationale for the approaches taken and the challenges we faced, including the lack of "truth." Finally, the paper discusses the evaluation of the research results in terms of numerical and distributive accuracy, operational feasibility and public acceptance, in seeking to optimize the strengths of different methods.

Key Words: hot deck, administrative records, spatial analysis, truth deck

1. Background

The Census Bureau began imputing values for missing responses with the 1950 census. The fact that this also was the first census to be processed by an electronic computer was not a coincidence. Since that time, the Census Bureau has worked to create or refine computerized techniques for imputing responses. One procedure developed to take advantage of the processing power of computers was the use of a "hot deck" – first implemented in the 1960 census – to impute values from nearby housing units. A hot deck is a data table (or "matrix") in which values of reported donor responses, stratified by selected characteristics of the individuals, are stored and updated on a flow basis and used as needed to assign values of the variable(s) in question to donees, that is, people (or housing units) with similar characteristics who did not respond. This means that values imputed generally come from the nearest household ("nearest neighbor") with similar characteristics. Hot deck imputation (also known as "allocation") is used in most cases when it is not possible to assign values either from other information provided by the respondent or from information provided by other household members. This method is applied not only to population characteristics (for example, age, race, educational attainment, and income) but also to housing characteristics (such as housing tenure), housing unit status (whether an address actually identifies a unique housing unit), occupancy status, and population count. The hot deck is also known as a "sequential" hot deck because housing units are sorted geographically first, after which the hot deck sequentially stores and allocates values as it passes from one housing unit to the next.

As computer capacity grew, so did the effort to make hot-deck imputation more accurate by making the matching criteria more sophisticated. For example, more dimensions (characteristics) were used to match donors and donees. The number of values stored in each hot deck cell increased to accommodate situations where there was a large non-response and the risk of assigning values from the same donor repeatedly was higher. For example, in Census 2000, the hot decks used to assign an Hispanic origin were stratified by age and race, and they also were divided into three separate hot decks by whether the person had a Spanish surname (as determined from a dictionary of names): 1) donors with a Spanish surname, 2) donors with a non-Spanish surname, and 3) donors with no surname or with a name that was not clearly either Spanish or not Spanish.

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the author(s) and not necessarily those of the U.S. Census Bureau.

2. The Problem

The Committee on National Statistics' Panel to Review Census 2000 (National Research Council, 2004) noted several concerns about the "sequential hot deck (SHD)" approach. First, the panel noted that the SHD relied on a single donor rather than obtaining more information from the local area, which could result in the hot deck assigning an "odd" value from that single donor. Second, SHD may not fully incorporate the multivariate nature of imputation. Third, this method may have difficulty doing simultaneous imputation of several variables that are correlated. Finally, SHD fails to produce an error estimate that could be used to determine a measure of characteristic (item) non-response variance so that it could be included as a component of the overall variance estimate. (National Research Council, 2004, 442-444) An external panel, tasked with reviewing the Census Bureau's Program of General Census and Survey Research and Support and three of its specific programs (Missing Data and Imputation, Small Area Estimation, and Ethnographic Applications), had similar concerns (see Tanur *et al*., 2003). The panel recommended a "state-of-the-art imputation engine for basic demographic characteristics of households and the individuals within them." They also recommended the use of administrative records to improve the imputation models (Tanur *et al*., 2003, 13).

These concerns along with a general goal of improving census coverage prompted the formation of a research group to examine whether alternative allocation methods might better account for missing information in the decennial census.

3. Types of Imputation

Two main paths of research are being conducted on imputation: 1) count imputation and 2) characteristic imputation.

The count imputation path comprises the following types of imputation:

- Household Size Imputation. When Census Bureau records indicated that a housing unit was occupied but the number of residents could not be determined from the information available, a population count for the unit was imputed.
- Occupancy Status Imputation. When Census Bureau records indicated that a housing unit existed but whether it was occupied or vacant could not be determined from the information available, occupancy status (occupied or vacant) was imputed; then, if the unit was imputed to be occupied, a household size was imputed.
- Housing Unit Status Imputation (referring to whether the address actually identified a unique housing unit). When Census Bureau records contained an address, but there was insufficient information about whether the address represented a valid, nonduplicated housing unit, the unit's status as an occupied housing unit, a vacant housing, or a nonexistent housing unit was imputed. If the unit was imputed as occupied, its household size was imputed. Examples of addresses not considered to represent housing units include buildings used only for business purposes.

Present research efforts on characteristic imputation will be applied only to the Census 2000 short-form characteristics – household relationship, sex, age, race, Hispanic origin, housing tenure (renter/owner), and occupancy type.

4. Alternative Imputation Methodologies

The following are the main methodologies being tested:

A. Administrative records data involving direct assignment.

For records that can be matched to administrative records (see Farber *et al*., 2005), and under circumstances permitted by relevant legal authorities and polices, the information will be assigned to the matched census record. Methodologies using administrative records will adhere strictly to all requirements to protect the confidentiality of respondent's information.

B. Administrative records data involving modeling.

For records that cannot be matched to administrative records, modeling will be implemented only for count imputation topics (status, occupancy, and count). More details on this method can be found in Farber *et al.* (2005).

C. Spatial Analysis.

In this technique (see Thibaudeau, 2002), characteristics are obtained from neighboring housing units to generate an imputed value through statistical modeling, systematically capturing the relationship among characteristics of interest. Spatial analysis in this application refers to the relationship between the missing data item and other characteristics from geographically close households. This process allows imputation errors to be assessed.

D. Canadian Census Edit and Imputation System (CANCEIS).

In 1992, Statistics Canada introduced a new method of imputation for demographic variables (Bankier, 1997; Bankier, 2000). The key features are first to identify the nearest neighbor from which to borrow information and then to determine the minimum number of variables to impute for the record requiring imputation. This procedure reverses the traditional imputation procedure of determining what variables to impute first and then finding the information from comparable neighbor households. The advantage of CANCEIS is that, rather than looking at only one or two variables at a time, it maximizes the number of variables viewed simultaneously, resulting in a better preservation of the joint distribution of variables before and after imputation.

E. Modified traditional hot deck.

Improvements such as capping the maximum allowable household size and performing imputation in phases using a pre-defined geographic sort will be evaluated.

5. Strategy for Technically Evaluating Alternative Imputation Methodologies

The plan for evaluating each type of imputation is composed of the following steps:

- Create a "truth deck" for each state. The "truth deck" files are made up of households for which no imputation was needed under the Census 2000 edit and imputation process for anyone in the household. Certain fields are flagged as being "missing" for the purpose of this analysis. The truth deck is intended to reflect as much as possible the results of the Census 2000 operations, so the truth deck identifies about the same percentage of cases requiring imputation based on the missing data patterns observed in Census 2000. About the same percentage of records with reported data is flagged and treated as if the reported data are missing for analysis purposes. The construction of this comparison file is discussed below. Separate sets of truth deck files were created for count imputation and for characteristic imputation.

- Run each of the methods, including the traditional hot deck, against the truth deck file for each state.

- Compare the resulting distributions against the reported values; calculate statistics that can be used to compare the results for each alternative methodology, including the traditional hot deck; and analyze the operational feasibility of the method in a decennial census environment.

6. Creation of Truth Decks

Three different truth decks are being created:

A. Count Imputation Truth Deck.

The count imputation truth deck file (see Williams, 2005) is a housing-unit-level file stratified by selected characteristics at the block-group level. The truth decks for each state use available census information to stratify all records into different groups, based on selected operational, characteristic, and geographic variables. Classification variables may vary from state to state. A uniform probability of missing: 1) status, 2) occupancy, or 3) count information is assigned to all records within each stratum. The probability is the ratio of the number of cases requiring each type of count imputation to the total number of cases within each stratum. Each record is randomly flagged as needing imputation or not based on the probabilities in each stratum. The flagging process is replicated 100 times to account for variability in the random selection

of records being flagged. Afterwards, this truth deck is used as the data file on which to run different count imputation methodologies and conduct analyses allowing comparison of imputed values to actual values. For each method described above, at least one imputation is run for each replication.

B. Housing Unit Population Characteristic Imputation Truth Deck.

The housing unit population characteristic imputation truth deck file is a person-level file for imputing race, Hispanic origin, age, sex, relationship, and tenure. Even though tenure is a housing-unit level characteristic, we use the same truth deck. The methodology is similar to that for the count imputation truth deck but with different flagging procedures. For person characteristics, the flagging is set at the person level. Tenure is set at the housing-unit level.

The creation of this truth deck is based on the observed missing data patterns in Census 2000 (i.e., missingness is not random, but rather patterned). If a characteristic is usually missing with another characteristic for a certain segment of the population, the pattern is retained when flagging records to indicate which items are to be treated as if needing imputation. The primary goal in selecting households to flag for simulated imputation is to preserve: 1) the relations among imputation rates for different items for the same person, and among imputation rates for items for different people within the same household; and 2) the relations among the imputation rates for items and the values of the nonmissing items for the person and other people within the household. County and household size are the two most important factors for determining the missing data pattern. The missing data pattern varies from county to county and, within a county, it varies from one-person households to two-person households, from two-person households to three-person households, etc.

More specifically, the first step in truth deck creation was to build a donor file of Census 2000 households for which tenure was not imputed, and in which no person had any short-form characteristics imputed: race, Hispanic origin, age, sex or relationship. (Note: items edited, changed or assigned due to consistency checks are treated as non-imputed.) Next, we created an imputation file containing records from households for which tenure was imputed, and/or any person had race, Hispanic origin, age, sex or relationship imputed. This imputation file is the basis for determining what household patterns of imputed/non-imputed items exist, and how frequently they occur by county. In other words, our approach considers the household pattern of which characteristics were imputed and the values of the characteristics that were not imputed. Thus donor households matching on the values of the characteristics that were **not** imputed are flagged for simulated imputation for the characteristics that **were** imputed.

We determined that key criteria must be satisfied in order to find acceptable donor households for a given household imputation pattern: occurrence within same county; same household size; matching on tenure, if it is not imputed; and matching person-by-person on each characteristic that was not imputed. The characteristics themselves are stratified into several broad groupings: six for race; two for Hispanic origin; four for age; two for sex and five for relationship. To improve the donor pool, we do not require a match on sex or age if a match occurs on relationship.

The household imputation pattern also depends on the ordering of the person records within the household. For each imputation pattern, we created alternative patterns by resorting the person patterns. Each donor then was compared to every alternative pattern within a county. Where insufficient donors were available to match to the imputation cases, we successively broadened the search, first by comparing to every alternative pattern as above, but not requiring the donor to be in the same county. Next, we searched within the county but relaxed the criteria such that we only required the donor to match on household size and householder race. Finally, we searched through the state with relaxed criteria. As a limitation, if a pattern did not get enough donors through the above stages, we do not search further for any more donors. (Note: we also do not get donors for patterns for household sizes larger than 16.)

Households in which all person characteristics are imputed constitute a special case for us. To address these cases, we employed a modified, two-stage methodology for selecting donors. In the first stage, we created cells based on household size, tract, multi-unit status and tenure. Using these cells, we found the proportion of "fully imputed" cases, then randomly selected the necessary number of donors among those records not already selected as donors. Cases where tenure was also imputed were handled identically, except that the cells were created exclusive of the tenure variable. After running the first stage of the methodology, residual cases for which an insufficient number of donors was selected fall to the second stage. In this stage, we ran the same methodology, except we no longer required selected donors to occur in the same tract, merely the same county.

One characteristic imputation truth deck file was created per state, unlike count imputation where 100 replicates were deemed necessary to mitigate variability. Our one-replicate rule was principally due to the high levels of geography in the file, which precluded the likelihood that atypical demographic distributions would occur. Empirical evidence for several states also showed general proportional differences only in the range of the 4th decimal place.

C. Vacant Housing Unit Truth Deck.

A separate truth deck consisting of vacant housing units only is being created to evaluate whether the imputed results from spatial analysis or the Census 2000 hot deck are closer to the reported values for the type of vacancy.[2] The methodology to create the vacancy truth deck will be similar to the methodology of the other truth decks.

[2] Type of vacancy consists of the following categories: 1) for rent; 2) for sale only; 3) rented or sold, not occupied; 4) for seasonal, recreational, or occasional use; 5) for migrant workers; and 6) other vacant.

7. Applying the Alternative Methodologies

The table below summarizes the methodologies applied to count imputation:

Table 1. Imputation Methodology Used for Count Imputation

Method	Count Imputation		
	Status	Occupancy	Count
Administrative Records Assignment	X	X	X
Administrative Records Modeling	X	X	X
Spatial Analysis	X	X	X
CANCEIS			
Traditional Hot Deck	X	X	X
Revised Hot Deck	X		X

The next table summarizes the methodologies applied to characteristic imputation:

Table 2. Imputation Methodology Used for Characteristic Imputation

Method	Characteristic Imputation					
	Race	Hispanic Origin	Age	Sex	Tenure	Relation-ship
Administrative Records Assignment	X	X	X	X		
Administrative Records Modeling						
Spatial Analysis					X	
CANCEIS			X	X		X
Traditional Hot Deck	X	X	X	X	X	X
Revised Hot Deck[3]						

[3] The Revised Hot Deck will not affect characteristic imputation and will not be available for inclusion into the first phase of characteristic imputation evaluation. As Table 2 indicates, only Traditional Hot Deck currently can impute all types of missing information. If a new method is recommended to replace the hot deck for some types of imputation, then clearly a hybrid approach to imputation will be required to impute situations which the new method cannot.

8. Fully Evaluating the Results

The research question to be answered is – can we develop an imputation method that is superior to the current hot deck method and meets the following criteria/guidelines:

- Numerical and Distributive Accuracy
- Operational Feasibility and Cost Effectiveness
- Public Acceptance

Evaluation of numerical and distributive accuracy of the results from these methodologies is being conducted on two levels. On one level, we are using descriptive analyses involving the examination of distributions (for example, single year of age, race by Hispanic origin, and age by sex by relationship) to compare the impact of each methodology on characteristics of interest. On another level, we are employing statistical measures that summarize the accuracy of the imputations based on each method.

Additional information is being collected and analyzed concerning operational feasibility issues such as:

- The complexity of the method/process
- The impact and interrelationship of external systems and subsystems
- The number of operating systems, run times, file formats and programming languages used
- Security issues (e.g., use of administrative records needing an approved environment for processing)
- The degree of human intervention
- The number of machine-to-machine transfers required.

Based on the experience of implementing the hot deck methodology, any "new" method (or methods) chosen will need to be fully and clearly explained to ensure that the public and data users understand the method, its usefulness, and any other implications.

9. Time Frame for Analysis

Alternative methodologies are currently being run on the various "truth deck" files, including the above-mentioned evaluation measures, and these methodologies will be compared and analyzed. One or more methodologies will be selected with the goal of developing the specific individual methodology or a hybrid of methodologies for testing in the 2006 Census Test. For details on item imputation research results, see Obenski *et al*. (2005).

10. Limitations

We have noted the following limitations thus far:

- Because not all methods generate all needed fields (even within the count/status and characteristic areas; see Table 1), it will be a challenge to compare them.
- Because we will generate multiple measures for each method, it will be a challenge to develop an overall "summary" of these measures.
- The truth deck reflects decennial respondents or nonresponse follow-up interview results; thus, their status as "truth" is an assumption that is not always correct. However, they are taken as the standard for comparison.
- The method for creating truth decks may itself create an unknown bias that may favor one methodology over another.
- Treating housing units as independent evaluation units ignores the fact that properties of addresses and people are geographically clustered.

11. References

Bankier, M. (1997). *Documentation of the New NIM Prototype*. Social Survey Methods Division Report. Ottawa: Statistics Canada.

Bankier, M. (2000). *Imputing Numeric and Qualitative Variables Simultaneously*. Social Survey Methods Division. Ottawa: Statistics Canada.

Bauder, M. and Judson, D. H. (2003). *Administrative Records Experiment in 2000 Household Level Analysis.* Washington, DC: U.S. Census Bureau.

Bye, B. V. and Judson, D. H. (2004). *Results from the Administrative Records Experiment in 2000.* Census 2000 Testing, Experimentation, and Evaluation Program Synthesis Report No. 16, TR-16. Washington, DC: U.S. Census Bureau.

Farber, J., Wagner D., and Resnick, D. (2005), "Using Administrative Records for Imputation in the Decennial Census," to appear in *Proceedings of the 2005 Joint Statistical Meetings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Obenski, S., Farber J., and Chappell, G. (2005), "Research to Improve Census Imputation Methods: Item Results and Conclusions," to appear in *Proceedings of the 2005 Joint Statistical Meetings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

National Research Council (2004). *The 2000 Census: Counting Under Adversity.* Panel to Review the 2000 Census. Constance F. Citro, Daniel L. Cork, and Janet L. Norwood, eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Tanur, Judith *et al*. (2003). *External Panel Review of the Statistical Research Division of the Census Bureau: Executive Summary.* February, 26, 2003.

Thibaudeau, Y. (2002). Model Based Item Imputation for Demographic Categories. Survey Methodology, 28, 135-143.

Williams, T. (2005), "The Development of Truth Decks for the 2010 Census Count Imputation Research," to appear in *Proceedings of the 2005 Joint Statistical Meetings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.