# Research to Improve Census Imputation Methods:
# Item Results and Conclusions[1]

Sally Obenski, James Farber, and Gary Chappell
Sally Obenski, U.S. Census Bureau, Washington, DC 20233-9200

## Abstract

This paper presents the summary results from a research effort to examine potential improvements to characteristic imputation in the decennial census. It includes results from research conducted on various methods, including the use of administrative records assignment and spatial analysis. It discusses our approach to optimize the strengths of the different methods by developing hybrid methods. We also discuss statistical, operational, and policy implications.

**Key Words:** feasibility, hybrid, administrative records, hot deck, matching

## 1. Introduction and Statement of Problem

Since 1960, the Census Bureau has used a method known as the "hot deck" to impute information that has been successfully collected from nearby housing units for information missing from neighboring units. A hot deck is a data table (or "matrix") in which values of reported donor item responses, stratified by selected characteristics, are stored and updated on a flow basis. These donor responses are used as needed to assign values of the variable(s) in question to donees, that is, people (or housing units) with similar characteristics that did not respond to the required question(s). These imputed values generally come from the nearest household ("nearest neighbor") with similar characteristics. Hot deck imputation (also known as "allocation") is used for most items when other information on the same person or housing unit record is not available to be used to determine the value of the missing item. It is often possible to use related responses from the same record to impute the missing item, a process known as "assignment." Note that in the research described in this paper, assignments were not within scope.

At least two external panels noted concerns about the hot deck approach and asked that the Census Bureau consider alternatives (National Research Council, 2004, 442-444; Tanur et al., 2003). One of the panels specifically recommended research into administrative data (Tanur et al, 2003). 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.

## 2. Research Objectives and Scope

The objective of this research was to identify a method or combination of methods that would improve allocation of census short-form characteristics for testing in the 2006 Census Test (Obenski 2005). These items are household relationship, sex, age, race, Hispanic origin, and housing tenure (renter/owner).[2] At the same time, the method(s) must be operationally feasible with manageable policy implications. The scope of the research included only the housing unit population. The population in group quarters, such as prisons or nursing homes, was out of scope. Additionally, evaluation was done at the household and state levels. This allows us to focus on individual and distributive accuracy at the levels that are most consistent with our mission of providing accurate characteristics. Given the complexity of this

---

[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] Type of Vacancy, a short-form item indicating the status of a vacant unit, was not included in this report.

research and limited resources, we agreed that a minimum of three states would be examined by all imputation methods (Delaware, Georgia, and New York) and that additional states would be added in a prespecified order. This paper presents a summary of the results.

## 3. Alternative Characteristic Imputation Methods and Application

Table 1 lists the methodologies that were tested and the items each method could impute (Chen, 2005).

**Table 1. Alternative Item Imputation Methods by Type of Imputation**

| Method | Tenure | Relationship | Sex | Age | Race | Hispanic Origin |
|---|---|---|---|---|---|---|
| Traditional Hot Deck | X | X | X | X | X | X |
| Administrative Records Direct Assignment | | | X | X | X | X |
| Spatial Analysis | X | | | | | |
| CANCEIS | | X | X | X | | |

### A. Traditional Hot Deck.

The Census 2000 edit and imputation system was used to test the traditional hot deck process in this research. This system combines both editing and imputation in the main program. Missing item values (as well as reported values not satisfying the edits) that are not assigned a value by the edit rules are imputed from hot deck matrices. Each state is processed separately, and its records are edited sequentially in a specific geographic order. Housing unit and population records satisfying specified quality conditions are used as donors and update the matrices. When an item value needs to be imputed, a value is assigned from a donor currently in the matrices defined by the item's edit. Generally this means that imputations come from closely preceding housing units with similar reported characteristics within the same state.

The edit system applied the individual item edits and conducted the required item imputation in a single program. Some edits made use of related information to assign values instead of relying on allocation matrices. For example, first names were used when present to assign a value for missing sex, and last names were used to help to impute Hispanic origin. In Census 2000, the hot deck used supplemental information from the long form to edit the short-form items. For example, if tenure was missing but the long form listed a rent payment value, then the hot deck would impute tenure as "renter." This is notable because in this research the hot deck ran exactly as it did in Census 2000, meaning it used long-form data when possible to impute short-form items. Because the 2010 census will not include a long form, the hot deck had an advantage in this research that it would not have in 2010. However, a follow-up analysis found that the results were not changed much due to this situation and the overall conclusion of the research was not changed at all.

### B. Administrative records data involving direct assignment.

For census person records that are missing age, sex, race, or Hispanic origin and can be matched to administrative records that contain this information, the administrative records value is assigned to the missing item on the census person record. This research used the Social Security Administration's Transaction File supplemented with race and Hispanic origin information from Census 2000 respondent records (Farber and Miller, 2003). Administrative records assignment is not an allocation method but more like an edit because it uses supplemental information about the same person from other sources, including previous census data, to make assignments. Using administrative records to assign imputed values adhered strictly to all requirements to protect the confidentiality of a person's information (Clark and Gates, 1999).

An administrative records assignment can only be done when a matching administrative record is available. Person record matches are highly dependent on names, dates of birth, and addresses being accurately reported in the census. Probabilistic matching techniques are used to match these variables in the census to administrative records (Killion, 2002). Not all census records that need item imputation can be matched to administrative records. Therefore, administrative records assignment requires another method to impute missing responses on census records not matched to an administrative record, or matched to one that cannot provide the needed information.

Only census records that matched to administrative records contributed to the comparative results in this study. Administrative records assignment had the advantage of having more information available for allocation than the other methods.

## C. Spatial Analysis.

Spatial analysis (Thibaudeau, 2002) captures the correlation between the tenures of adjacent neighbors over moderately extended geographic areas. The tract (approximately 2000 housing units) is the geographical level of choice. Based on the correlation, a statistical model is generated that imputes a missing tenure by reconstructing it from the correlation. The method tends to preserve the pattern of tenure dispersion better than hot deck. One major disadvantage of spatial analysis is that it only could be applied to the tenure item, making it an unlikely candidate from an operational perspective. As mentioned, this method did not have access to any of the supplemental tenure-related information that was available on the long form and used by the hot deck.

## 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). This system was originally named the "Nearest-neighbor Imputation Methodology (NIM)" and is based on the Fellegi-Holt principles (1976) of making the minimum number of changes necessary to the failed record to allow that record to pass all edit rules. NIM was used in the 1996 Canadian Census. The Canadian Census Edit and Imputation System (CANCEIS), based on NIM, was first used in the 2001 Canadian Census. The key features of CANCEIS are the Decision Logic Tables and the Imputation Engine. The Decision Logic Tables contain the rules that determine whether the variables in a given record pass edit rules determined by subject-matter experts. The Imputation Engine then attempts to find potential donor households, stratified by household size, that are "similar" to the households with failed records based on matching the characteristics on both sets of records that passed the edit rules. The precision with which these matches

are made is determined by setting certain system parameters. Once a match is made, the donor household supplies the value(s) needed to the receiving record that will allow that record to pass the edit rules.

One key 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. Another advantage is that the selection of donors from the pool of potential donors is based on a probability function, thus allowing for measurement of variability introduced by the method. Like spatial analysis, it did not have access to any supplemental information as administrative records did. One drawback about CANCEIS is that it is a proprietary Statistics Canada product, which would necessitate special arrangements for any official use by the Census Bureau.

## 4. Methodology for Evaluating Alternative Item Imputation Methods

The major methodological steps for evaluating the alternatives are discussed below.

## A. 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 "missing" for the purpose of this analysis. The truth deck is designed 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 and generally preserves the missing data patterns as observed in Census 2000.

## B. Identify pre-specified evaluation criteria.

Changing a decennial census operation has substantial implications. Therefore, the guiding principle was that the method had to substantially outperform the traditional imputation method in order to be considered for operational feasibility and policy assessments. The evaluation criteria are listed below and discussed in the results section.

- Individual Item Consistency with Truth Deck and Distributive Accuracy—Is the method more accurate than the traditional hot deck?

- Operational Feasibility—Can we implement the method in a census production environment?

- Policy Review—Will the method(s) be publicly and legally acceptable?

## C. Develop an evaluation workflow.

The next methodological step was to determine the evaluation workflow to help ensure consistent evaluative approaches among tested methods.

## D. Run the methods and compare the results.

The next step was for each method to be run against the truth deck for the three base states (Delaware, Georgia, and New York), and then against as many more states as resources allowed. Based on the accuracy of the imputations from each method, candidate methods were identified for the operational feasibility and policy assessments.

## 5. Research Results

The overall results were that administrative records assignment showed substantial increases in accuracy over the hot deck for the truth deck cases that could be matched to administrative records. The other methods—spatial analysis and CANCEIS—showed modest improvements in some instances but did not demonstrate overall superiority to the hot deck. Administrative records assignment was the only method that performed sufficiently better than the hot deck on four out of the six items to warrant the operational and policy assessments.

We checked the percentage of agreements with the truth deck for each item and method at the person level. The summary of the consistency check can be found in Table 2. We then examined distributive accuracy. Those results are summarized in Tables 3 and 4 and details are discussed below.

As the tables show, spatial analysis and CANCEIS were more accurate than the hot deck in some instances, but we concluded that it simply was not practical to implement alternative methods item by item.

## A. Individual Item Consistency and Distributive Accuracy Results by Item

To evaluate accuracy at the lowest level, the person level, we measured how consistent the values imputed by each method were to the truth deck values. We then looked at distributive accuracy at more aggregated levels. To measure individual accuracy with more statistical rigor, we used Cohen's Kappa Item Level Measure of Agreement (Agresti, 1990). This statistic measures the strength of agreement between the imputed values and the true values. We examined distributive accuracy by computing the absolute relative error of each method compared to the truth deck. We only provide the total distributional error comparisons. The purpose of these measures was to identify a method or methods that are clearly superior to the hot deck and that should be considered for implementation in the 2006 Census Test.

Due to resource constraints, only four states were fully tested for each of the candidate methods. These states were Delaware, Georgia, New York and Florida. All of the tables show combined results for all four states. The results for each state are similar to the combined results and are not shown.

The following tables provide a summary of our research results that suggest the use administrative records to assign missing values where possible, followed by the hot deck. The most accurate results have been bolded.

As shown in Table 2, administrative records assignment yielded highly accurate allocations at the person level for sex, age, race, and Hispanic origin. The consistency rates for cases that could be matched to administrative records were higher than for any other method. For relationship, the overall consistency rates for CANCEIS and hot deck were comparable. For tenure, the overall consistency rate was somewhat better for hot deck.

### Table 2. Summary of Allocation Consistency Rates for All Four States

| Method | Tenure | Relationship | Sex | Age | Race | Hispanic Origin |
|---|---|---|---|---|---|---|
| Traditional Hot Deck | **64.8%** | 75.6% | 92.3% | 25.8% | 75.7% | 95.4% |
| Administrative Records Assignment | | | **99.2%** | **98.8%** | **96.5%** | **99.0%** |
| Spatial Analysis | 57.5% | | | | | |
| CANCEIS | | **76.1%** | 58.3% | 26.9% | | |

Table 3 provides the Kappa statistic results. Kappa shows the predictive value of the information used in each imputation method, such as the demographic data of administrative records or the nearest neighbor concept of hot deck. Kappa can take any value from –1 to +1, with values greater than zero implying that the imputation method is better than a random imputation, and values less than zero implying that the imputation is worse than random imputation. As can be seen, the age, sex, race, and Hispanic origin assignments made from administrative records are far superior to the null model of independence. However, the hot deck also performs well, with reduction in error ranging from 0.18 percent for age to a high of 0.85 percent for sex.

### Table 3. Allocation Kappa Statistics for All Four States

| Allocated Characteristic | Hot Deck | Spatial Analysis | AR Assignment | CANCEIS |
|---|---|---|---|---|
| Age | 0.18 | | 0.99* | 0.19 |
| Sex | 0.85 | | 0.91* | 0.16 |
| Relationship | 0.63 | | | 0.63 |
| Tenure | 0.46* | 0.36 | | |
| Race | 0.57 | | 0.94* | |
| Hispanic Origin | 0.76 | | 0.94* | |

* The Kappa statistic is statistically significantly different from the other Kappa statistics for the characteristic at the 0.01 significance level.

Next, we examined distributive accuracy by comparing the distributions produced by each method to the truth deck distribution and computing the absolute relative errors for each item imputed by each method. Table 4 summarizes the absolute relative error results. Again, when administrative records were available, administrative records assignment produced distributions with substantially lower

relative errors than the other methods, except for sex where administrative records assignment did slightly better than CANCEIS.

### Table 4. Summary of Total Relative Error for Administrative Records Matched Cases

| Method | Tenure | Relationship | Sex | Age | Race | Hispanic Origin |
|---|---|---|---|---|---|---|
| Traditional Hot Deck | **35.8%** | 28.1% | 3.1% | 83.4% | 118.0% | 12.6% |
| Administrative Records Assignment | | | **0.1%** | **4.6%** | **40.2%** | **2.5%** |
| Spatial | 46.3% | | | | | |
| CANCEIS | | **20.7%** | 0.2% | 34.4% | | |

Table 5 illustrates the relative error for a simulated hybrid of administrative records assignment followed by a hot deck. The reduction in relative error for the hybrid may be conservative since in actual production, the administrative records assigned cases could possibly become donors for use in the hot deck allocation matrices. On the other hand, any reduction in hot deck's relative error will be contingent on how many cases administrative records can match. The more allocation done by administrative records assignment, the lower the relative error of the hybrid will be. The hot deck generally produces more accurate allocations as the number of hot deck allocations decreases.

### Table 5. Summary of Total Relative Error for All Four States for Administrative Records Assignment, Hot Deck, and Hybrid

| Method | Sex | Age | Race | Hispanic Origin |
|---|---|---|---|---|
| Traditional Hot Deck | 3.1% | 83.4% | 118.0% | 12.6% |
| Administrative Records Assignment | 0.1% | 4.6% | 40.2% | 2.5% |
| Hybrid | **0.1%** | **31.1%** | **51.1%** | **2.3%** |

The records on the truth deck represented ideal census respondents because they responded to every question. In particular, most of them reported highly accurate names and dates of birth, which along with address were key variables for matching to administrative records. Therefore the truth deck cases are easier to match to administrative records than

the census records that did not fully report all items. To ensure that we are clear about the difference between our ability to match truth deck cases versus true Census 2000 item imputation cases, we compare the match rates for these two universes in Table 6. Table 6 illustrates the matching challenge facing administrative records assignment in a production environment. The match rates to administrative records for the Census 2000 person records that required item imputation are lower than for the truth deck cases for every item, and much lower for sex and age. The low match rate for sex is primarily due to the large percentage—90 percent—of cases that the hot deck assigned in an edit based on name, which means only the hardest-to-match cases required allocation. As for age, preliminary research indicates that we can substantially increase match rates by optimizing the matching algorithm for census production.

**Table 6. Match Rates to Administrative Records of Truth Deck Cases and Actual Census 2000 Imputed Cases in New York**

| Universe | Sex | Age | Race | Hispanic Origin |
|---|---|---|---|---|
| Truth Deck | 84% | 86% | 85% | 88% |
| Actual Imputed Cases | 10.9% | 2.0% | 71.7% | 71.3% |

**B. Operational Feasibility for Administrative Records Assignment**

Based on the statistical results, the team selected administrative records assignment as the primary imputation method with the unmatched cases falling to hot deck. Consequently, we conducted an operational assessment on administrative records assignment based on processing data collected in this research. This initial operational assessment indicated no major obstacles to implementing administrative records assignment in a census production environment. Processing the entire nation took about 68 hours, but this time could be reduced through additional parallel processing (2 states were run at a time for this test.) However, census records used in the research already contained the Protected Identification Keys (PIKs) used to merge census records to

administrative records. Therefore, for future production activities, additional processing time will be needed to prepare the administrative records and census files for matching. Our initial assessment is that this process could be implemented in a production environment before edit and imputation, and thus would not negatively affect the imputation schedule.

This operational assessment is based on our best available estimates but a census production environment often presents unique and unforeseen challenges. Therefore it will be critical to fully evaluate the operational performance of the hybrid method in the production environment of the 2006 Census test.

**C. Policy Implications for Administrative Records Assignment**

The results of this research indicated that complementing our traditional hot deck imputation method by adding an administrative records direct assignment phase would improve the accuracy of item allocation in the census. However, we need to assess the policy implications arising from this use of administrative records. Our initial assessment based on this research is that there are no major policy hurdles to continue exploring this use of administrative records in the 2006 Census test. Prior to and following the census test, we will continue to research the policy implications of using administrative records for item imputation in the decennial census.

**6. Limitations**

- The truth deck was based on complete Census 2000 households for which no allocation was required. The truth deck cannot perfectly reflect the actual Census 2000 universe of cases that needed item imputation. Therefore, we do not know how any of the alternative methods would perform with the actual universe of interest.

- Only one replicate of the truth deck was run. We believed this to be sufficient because of the large number of cases flagged for imputation on the truth deck.

- The method for creating truth deck may itself create an unknown bias that favors one method over another.

- The research conclusions were only based on four states.

- The research did not consider joint distributions among related items but examined items individually.

- The hot deck had access to supplemental information that spatial analysis and CANCEIS did not.

- By definition, administrative records assignment had access to supplemental information because it is using information about the person on the administrative record to make its assignments. Unlike the other methods, administrative records assignment is an edit rather than an allocation method.

- Administrative records assignment of race and Hispanic origin is based on Census 2000 data, which may change by 2010.

- The effectiveness of administrative records assignment depends on the ability to match the census records needing imputation to administrative records.

- Based on this research, no single method can replace the hot deck because it is the only method capable of imputing missing values for all short-form items.

## 7. Conclusions and Recommendations

Several important limitations in this research were noted and considered. Despite these limitations, the group that conducted this research found that administrative records assignment was the only method consistently and demonstrably more accurate than the hot deck. Administrative records assignment was also assessed as operationally feasible for testing in 2006, and no policy obstacles were identified in this initial assessment that precluded further research and testing. However, administrative records assignment can be used only for sex, age, race, and Hispanic origin, and only for census records that match to administrative records. Another method must be used for allocation of non-matched cases and for allocation of tenure and relationship.

Therefore, we recommend that a hybrid method consisting of administrative records assignment followed by the traditional hot deck process be further tested in the 2006 Census Test. We recognize that records actually requiring imputation will be more difficult to match than the simulated truth deck imputation cases. In fact, administrative records match rates for the Census 2000 records from New York on which short-form data were imputed were substantially lower than for the truth deck cases. However, research is underway to improve these match rates. Nevertheless, by first applying administrative records assignment to records missing data, hot deck accuracy will likely be improved by reducing the number of imputation cases falling to the hot deck.

## 8. References

Agresti, A (1990), *Categorical Data Analysis,* New York: Wiley.

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

Chen, I. (2005), "Census 2006 Test Implementation Research Plan," Internal Census Bureau memorandum.

Clark C. and Gates, G. (1999), "Memorandum on Restricted Access Policy for Administrative Records," Internal Census Bureau memorandum.

Farber, J. and Miller E. (2003), "Matching Census 2000 to Administrative Records, Survey Research Methods Section Proceedings of the Joint Statistical Meetings.

Fellegi, I.P. and Holt, D (1976) "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical

Association, March 1976,Volume 71, No. 353, 17-35.

Killion, R.A. (2002), "Programming Documentation for Creation of a Race-Enhanced Census Numident Using Census 2000 Race Data," Planning, Research and Evaluation Division Administrative Records Research Memorandum Series #60, Washington, DC: Bureau of the Census.

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

Obenski, S. (2005), "Recommendation for the 2006 Census Test Production Method for Housing Unit Characteristic Imputation," internal Census Bureau memorandum.

Tanur, J.M. 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.