Using Administrative Records for Imputation in the Decennial Census[1]

James Farber, Deborah Wagner, and Dean Resnick
U.S. Census Bureau
James Farber, U.S. Census Bureau, Washington, DC 20233-9200

Keywords: Missing data; matching; statistical modeling; hot deck

1. Introduction

Administrative records (AR) provide a potential source of comprehensive, inexpensive and accurate data for a number of uses in Census Bureau programs. One promising application is imputation of missing data in the decennial census. The deterministic algorithms that have traditionally been used for imputation rely heavily on the similarity of housing units in small geographic areas. As the country has diversified, however, the assumption that neighbors are alike is growing increasingly tenuous. For this and other reasons, the Census Bureau has undertaken research into alternative methods that may provide more accurate imputed data.

This paper discusses two new imputation methodologies based on AR. One method is a direct assignment method, in which administrative data imputes missing data on a matching address or person record. The second method builds a statistical model that relates administrative data to census data and uses the predictions from those models to fill in missing census data. These methods were implemented on a set of truth decks and the results evaluated for accuracy and feasibility.

2. Background

There are two primary types of missing data imputation in a decennial census. The first is known as characteristic imputation or item imputation, which fills in data for respondents who answered some of the census questions but not all of them. The characteristics that may be missing are:

- sex
- age
- race
- Hispanic origin
- relationship
- tenure (i.e., owned or rented)
- vacancy type (for vacant units only)

The second type of imputation is count imputation, which has three categories:

- status imputation: for addresses where the very existence of a housing unit is unknown
- occupancy imputation: for housing units that exist but it is not known if they are occupied or vacant
- household size imputation: for occupied housing units with an unknown household size.

In Census 2000, both characteristic and count imputation were done using a sequential hot deck. In general, the hot deck imputed data from a "nearest neighbor" with similar characteristics to the housing unit, household, or person that required imputation. The "nearest neighbor" concept assumes strong serial correlation of housing units and people across the physical landscape; people who live near each other tend to be alike. The validity of this assumption can be called into question by the increasing diversity of the U.S.

The hot deck has other undesirable aspects, as noted by the Committee on National Statistics' Panel to Review Census 2000 (2004). First, the panel noted that the hot deck relied on a single donor rather than obtaining more information from the local area, which could result in the hot deck assigning a rare and unusual value from that single donor. Second, the hot deck may not fully incorporate the multivariate nature of imputation. Third, the hot deck may have difficulty doing simultaneous imputation of several variables that are correlated, such as race and Hispanic origin (Cresce, 2002). Finally, the hot deck cannot produce an error estimate.

A number of alternative imputation methods exist, including the use of AR. AR are data collected by government agencies to administer programs. Some examples of AR are the Medicare program and the income tax system. The Census Bureau has a long history of using AR for statistical programs like the intercensal estimates. AR provide a comprehensive, timely and inexpensive source of data with a number of potential statistical applications, including imputation. The AR acquired by the Census Bureau contain many of the characteristics imputed by the hot deck, specifically race, age, sex and Hispanic origin. In addition, AR may be useful for count imputation because AR contain addresses provided by people participating in programs, thus providing evidence

[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.

of the existence, occupancy status and household size of many addresses. Further information on research and production activities involving AR at the Census Bureau, are given in Leggieri *et al.* (2002), Farber and Miller (2003), and Resnick (2003).

3. Methodology

The Census Bureau undertook a research effort to examine alternative imputation methods (Cresce *et al.*, 2005). One part of this research involved simulating imputation using AR and evaluating the accuracy of the results. Truth decks were created from fully reported Census 2000 records that flagged data as missing based on the patterns of missingness observed in the census. The truth decks attempted to fully capture the correlations between missing items, such as race and Hispanic origin, between people in the same household, and among addresses in local areas, such as the tendency for vacant units to cluster. One truth deck was created for the simulation of characteristic imputation, and another truth deck was used for simulation of count imputation (Williams, 2005). The truth decks strengthened evaluation of the various alternative methods because the results of each method could be compared to the truth as reported by Census 2000 respondents.

Other alternative imputation methods as well as the Census 2000 hot deck were also researched. This paper discusses only the methods that involved AR. Different AR imputation methods were used for characteristic imputation and count imputation. Only one method was researched for characteristic imputation, and two AR-based methods were tested for count imputation.

3.1 Direct Assignment for Characteristic Imputation

In general, direct assignment of administrative data occurs by matching the census person record to an AR person, finding the necessary data on AR, and using them to impute the missing census data. It is important to note that direct assignment may leave a number of cases unresolved; any census record that does not match to a unique administrative record cannot be imputed by direct assignment and will require an additional imputation method. Thus the evaluation of this method focuses not only on the accuracy of the imputed data but also on the match rate. It would not be operationally efficient to implement a new method that provides highly accurate imputations but for few cases. The specific method used in the imputation research is described below.

3.1.1 Match Census 2000 person records to administrative records

For efficiency, this research used the results of a match between Census 2000 and AR that had been done for another project, covered in detail in Farber and Miller (2003). That match had been done by comparing name, date of birth, sex, and address of census records and AR via probabilistic matching techniques, and identifying matching person records. A match identifier was placed on the AR files to facilitate future projects, such as this imputation research. For the purposes of this research, it was sufficient to use the results of the previous match even though that match was not designed to optimize the accuracy of imputation. It is possible that the imputation results could have been more accurate if we had revisited the matching techniques but we believe these gains would have been marginal and not worth the large effort to re-match. Therefore, when a truth deck person record was flagged as having missing characteristics, we simply looked at the match identifiers in AR and if we found the identifier for that census person, then we considered that person as matched to AR.

3.1.2 Impute missing characteristics for matching census records from administrative data

For matching records, we pulled data from AR to impute any of the following person characteristics:

- Age
- Sex
- Race
- Hispanic origin

When one of these items is flagged as missing on the truth deck and the census person record matched to AR, we directly assigned the missing value as the value found in AR.

AR do not contain sufficient information on housing tenure or vacancy type, two additional census characteristics that may require imputation. An additional imputation method would be needed to impute these characteristics.

3.2 Direct Assignment for Count Imputation

As with characteristic direct assignment, count direct assignment involves matching the census address to an AR address, finding the necessary data on AR, and using them to impute the missing census data. And again, count direct assignment may leave some cases unresolved because the census address did not match an AR address. These unresolved cases would require an additional clean-

up imputation method.

3.2.1 Match census addresses to administrative records addresses

Each census address is uniquely identified by a variable called the Master Address File Identification number (MAFID). All of the address records on the count truth deck contained MAFIDs. AR also contain MAFIDs because the Census Bureau geocodes AR addresses, mostly by matching them to Master Address File addresses. Hence the address match was very simple: because MAFIDs were on both files, the match was a simple merge on MAFID of the two input files.

3.2.2 Impute for missing status

The AR used in this research come from programs that are administered to people. An AR address gets a MAFID if and only if a person reports that address to the agency administering the program. Thus it is likely that AR addresses physically exist.

To impute missing status, we matched each truth deck case flagged as needing status imputation to AR on MAFID and imputed the address as "occupied" if it matched, with a household size equal to the AR person count at the address.

For census addresses that did not match to AR, we could not impute status using direct assignment and instead imputed the result produced by the AR modeling method described in section 3.3 of this paper.

3.2.3 Impute for missing occupancy status

To impute missing occupancy status, we matched each truth deck case flagged as needing occupancy imputation to AR on MAFID and imputed the address as "occupied" if it matched, with a household size equal to the AR person count at the address.

For census addresses that did not match to AR, we could not impute occupancy using direct assignment and instead imputed the result produced by the AR modeling method.

3.2.4 Impute for missing household size

To impute missing household size, we matched each truth deck case flagged as needing household size imputation to AR on MAFID and if it matched, we imputed the household size as the AR person count at the address.

For census addresses that did not match to AR, we could not impute a household size using direct assignment and instead imputed the result produced by the AR modeling method.

3.3 Statistical Modeling for Count Imputation

Statistical modeling uses a quantified relationship between a dependent variable and independent variables to predict outcomes for cases where the dependent variable is not observed. In count imputation, the dependent variable is the status, occupancy status or household size of the address, depending on which type of count imputation is needed. The independent variables may come from AR or census data. We look at the census units where the count imputation outcomes are known to quantify their relationship with AR data and census data, and then use that relationship to predict the status, occupancy status or household size for census units where these outcomes are unknown.

3.3.1 Modeling status and occupancy status

We modeled status and occupancy status using a logistic regression model calibrated with a sample of linked AR and Census 2000 data. We identified a set of variables that we believed were related to the status or occupancy status of each census address and performed exploratory data analysis to confirm our expectations. Some variables were dropped because the data analysis demonstrated no clear relationship with status or occupancy status. The final set of possible predictor variables consisted of:

- Flag indicating if the census address matched to an AR address
- Flag indicating if the census address was in a multi-unit structure
- Flag indicating if the census address was enumerated in nonresponse followup
- Flag indicating if the census address had an enumerator return in Census 2000
- Flag indicating if either neighbor of the census address was a delete (for status) or vacant (for status and occupancy status)
- Percent of the block that consisted of deleted addresses (for status) or of vacant addresses (for status and occupancy status)
- Percent of the tract that consisted of deleted addresses (for status) or of vacant addresses (for status and occupancy status)

We then fed these possible predictors into SAS PROC Logistic with the forward stepwise option and no interaction terms, which generated the parameters of the best-fitting model. Tables 1 and 2 at the end of this paper give details on the parameter values for the delete model and the vacant model, respectively. The parameters are

within our expectations. In particular, the odds of a census address being a delete or vacant drop significantly if the address matched to an AR address. As stated above, because AR addresses are provided by people participating in government programs, they are likely to exist and be occupied.

We then applied the models to the truth deck cases that required status and occupancy imputation, which produced a probability of each outcome for each case. We imputed a specific outcome by generating a random number for each case and comparing the random number to the predicted probability of each outcome.

Some truth deck cases flagged as missing status or occupancy were ultimately imputed as occupied by AR modeling and thus required household size imputation as well. Due to resource constraints, we ran the household size model described below specifically for the cases flagged for household size imputation. This AR household size model was not appropriate for status and occupancy cases that ultimately needed a household size. Instead, we imputed a specific household size randomly from the census household size distribution in the block for status and occupancy cases that were modeled as occupied and needed household size imputation.

3.3.2 Modeling household size

We modeled truth deck cases flagged as missing a household size using a Poisson regression model calibrated with linked AR and Census 2000 data. The basic idea was to model the census household size using the AR household size along with local-area census statistics.

The census data already included well-defined household sizes because the census data are collected for all household members at the same time. Because AR data are collected to administer a program and not for statistical purposes, they do not necessarily adhere to a census-like concept of a household. The first step in the modeling process was to create our best estimate of a household for each AR address. We aggregated AR person records by address, broken down into two categories: a count of person records on a tax return for that address, and a count of person records not listed on a tax return but instead on some other AR source file, such as Medicare. People listed on the same tax return are more likely to represent a true household at the address, while people listed at the address but not on a tax return possibly reside elsewhere.

In addition, for housing units with tax returns, we noted the number of children at home exemptions, and computed the excess of this number over the total number of dependents born after April 1, 1978. This allowed for the possibility that older dependents, college students for example, did not physically live at the address and hence have a different relationship to the true census household size than other people listed on the tax return.

We also generated two fields to reflect the instances where persons were listed at more than one address within AR data: the first for people captured on more than one tax return having distinct addresses, and the second for people captured on more than one non-tax file having distinct addresses. The computed values for these fields were the sum of the complements of the simple probability that each person was actually living at a given address: $\sum_{p}(1-1/n_{p,t_p})$, where $n_{p,tp}$ is the number of distinct addresses for person p of type t_p: either on a tax return or not. The adjusted counts attempted to account for person records that did not have a distinct address in AR and hence should possibly not be counted fully in determining the census household size at the address.

In sum we had the following variables specific to each AR housing unit:

- Person count from a tax return
- Person count from non-tax AR files
- Person count of older tax return dependents
- Adjusted person count based on multiple tax return addresses
- Adjusted person count based on multiple non-tax AR addresses

In addition to these variables from AR, we also computed the average census household size at the sub-block and block levels. Based on some exploratory data analysis, we used the sub-block mean household size as a predictor for census households in sub-blocks with six or more occupied housing units. Otherwise, we used the block mean census household size as a predictor.

Prior to computing the regression parameter estimates, a simple transformation of the dependent variable was required. Because a Poisson random variable can take all integer values from zero to infinity but only occupied housing units were to be modeled, we adjusted the census household size by subtracting one from the known household size, giving an adjusted count representing household count in excess of one. This was the dependent variable used in the regression.

We also made a log transformation of the dependent variable. We considered making a log transformation of the independent variables, but this harmed the predictive

accuracy of the estimated model. In addition, to simplify the interpretation of the parameter estimates, no interaction terms were included in the model. The model looked like this: $\ln \lambda = \beta_0 + \sum_{i=1}^{6} \beta_i x_i$, where

$\lambda \sim$ Poisson parameter representing expected household size,

$\beta_0 \sim$ Intercept term,

$\beta_1 - \beta_6 \sim$ Parameters corresponding to independent variables $x_1 - x_6$,

and $x_1 - x_6 \sim$ The independent variables described above.

Table 3 at the end of this paper gives the results of the regression that was run using SAS PROC Logistic, which uses a maximum likelihood search methodology to determine the parameter estimates under the assumption that the response variable was distributed as Poisson with parameter λ = (Expected Household Size in Excess of One Person). It is interesting to note that nearly all of the parameters have similar magnitudes, from about 0.220 to 0.275. Also the direction of these estimates agrees with our expectations, which helps to validate the model.

We computed the parameter estimates using a sample of truth deck cases, some of which did not have a matching AR address. For these cases, we set the value of all of the AR-derived independent variables to zero, so that the only remaining independent variable was the average census household size in the local area. After estimating the parameters, we used the model equation to generate the predicted probability of each possible household size for the truth deck cases flagged as needing household size imputation. We used random numbers to select a specific household size, with top-coding at 99.

4. Results

This section gives a brief overview of the results. More detailed results are provided in Obenski *et al.* (2005).

4.1 Match Rates between Administrative Records and Census Data

It would not even be worth considering a new imputation method based on AR direct assignment if AR do not match a large number of census cases. Thus the first level of evaluation of the AR-based imputation methods is to compute match rates. For characteristic imputation, the match rates vary depending on the variable that needs imputation but in general the match rates between truth deck cases and AR were around 90 percent. This means that we could directly assign AR data for about 90 percent of the truth deck cases that needed imputation. Count imputation had about an 80 percent match rate between AR and truth deck cases that were flagged for imputation.

As mentioned above, the truth decks were built from fully reported data, which were generally easier to match to AR than the true census cases that needed characteristic or count imputation. The real test of AR is the match rate for the true Census 2000 imputation cases. As expected, the match rates are lower. For Census 2000 cases that did not report race or Hispanic origin, the match rates between census person records and AR are around 70 percent. For those without a reported age or sex, the match rates were extremely low, 10 percent or lower, but mainly because of the matching techniques used in the prior match project described in section 3.1.1 above. Preliminary results indicate the match rates could be substantially improved with minor tailoring.

For Census 2000 count imputation cases, the match rates between census addresses that actually needed count imputation and AR addresses are around 50 percent.

For both characteristic and count imputation, our initial assessment is that the match rates between AR and census data are large enough to merit the operational costs of using AR, should the accuracy of AR imputation also prove sufficient.

4.2 Accuracy of Simulated Imputations

The second level of evaluation of the AR imputation methods is the accuracy of the resulting imputations. Because the truth deck was constructed from fully reported census data, the evaluation described in this paper consists simply of comparing the true value to the value imputed by the AR methods.

4.2.1 Direct assignment for characteristic imputation

Table 4 summarizes the accuracy of the characteristic imputations from AR direct assignment. The agreement rate represents the percentage of truth deck cases for which AR imputed the correct value for each characteristic. That is, the AR imputed value agreed with the value originally reported by the respondent.

Table 4. Agreement Rates for AR Direct Assignment for Characteristic Imputation

Characteristic	Agreement Rate
Age	96%
Sex	99%
Race	96%
Hispanic Origin	98%

The characteristic imputations provided by AR direct assignment are highly accurate. Age is especially accurate; the agreement rate means that AR had the exact age provided by the census respondent in 96 percent of the cases. The AR imputation was within two years of the true value for more than 99 percent of the cases.

4.2.2 Direct assignment for count imputation

Table 5 summarizes the accuracy of the count imputations from AR direct assignment. The agreement rate represents the percentage of the time that AR imputed the correct status, occupancy status or household size.

Table 5. Agreement Rates for AR Direct Assignment for Count Imputation

True Status of Census Unit	Agreement Rate
Delete	0%
Vacant	0%
Occupied	100%

Of those units that AR direct assignment imputed as occupied, the imputed household size perfectly matched the true census household size for about 55 percent of the cases.

These striking results reflect that the strength of AR for count imputation is to impute whether or not a housing unit is occupied and to impute its household size. AR cannot directly assign a census count imputation case as a delete or vacant because the AR addresses are reported by people who participate in government programs. And hence AR addresses are overwhelmingly occupied.

4.2.3 Statistical modeling for count imputation

Table 6 summarizes the accuracy of the count imputations from AR modeling. The agreement rate represents the percentage of truth deck cases for which the AR-based model imputed the correct status, occupancy status or household size.

Table 6. Agreement Rates for AR Modeling for Count Imputation

True Status of Census Unit	Agreement Rate
Delete	84%
Vacant	72%
Occupied	81%

Of those units that AR modeling imputed as occupied, the imputed household size perfectly matched the true census household size for 36 percent of the cases.

Unlike AR direct assignment, AR modeling could impute all three possible count imputation outcomes and did so with a high degree of accuracy.

5. Conclusions

The application of AR for imputation in a decennial census is one of the ideal uses of AR because the AR person data are highly accurate for direct assignment. In addition, the accuracy and comprehensiveness of AR provide a solid foundation on which to model census data for count imputation purposes. If the results provided by AR prove sufficiently accurate and if operational and policy issues are surmountable, then the next step is to attempt to actually use one of these AR-based methods for characteristic or count imputation in the 2006 Census test.

6. References

Cresce, A. (2003) "Comparison of Edit and Imputation Procedures for the Question on Hispanic Origin: 1990 Census and Census 2000", *Proceedings of the 2002 Joint Statistical Meetings*, Government Statistics Section, Alexandria, VA: American Statistical Association.

Cresce, A., Obenski, S., and Chappell, G. (2005) "Research to Improve Census Imputation Methods: The Plan to Examine Count and Item Imputation," to appear in *Proceedings of the 2005 Joint Statistical Meetings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

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

Leggieri, C., Pistiner, A., and Farber, J. (2002) "Methods for Conducting an Administrative Records Experiment in Census 2000", *Proceedings of the 2002 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.

Resnick, D. (2003) "Improving the Accuracy of Race and Hispanic Origin in Administrative Records", *Proceedings of the 2003 Joint Statistical Meetings*, Survey Research

Methods Section, Alexandria, VA: American Statistical Association.

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.

Table 1. Parameter Estimates for Delete Model

Independent Variable	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Pr > ChiSq
			Lower	Upper	
Intercept	-4.7106	0.0085	-4.7273	-4.6940	<.0001
Match to AR flag	-1.1352	0.0058	-1.1466	-1.1238	<.0001
Multi-unit flag	-0.3927	0.0072	-0.4068	-0.3786	<.0001
In nonresponse followup flag	-4.0903	0.0073	-4.1047	-4.0760	<.0001
Enumerator form-type flag	4.3663	0.0074	4.3519	4.3808	<.0001
Either neighbor delete flag	2.0309	0.0060	2.0191	2.0427	<.0001
Block proportion of deletes	0.0493	0.0002	0.0489	0.0497	<.0001
Tract proportion of deletes	-0.0069	0.0004	-0.0078	-0.0060	<.0001

Table 2. Parameter Estimates for Vacant Model

Independent Variable	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Pr > ChiSq
			Lower	Upper	
Intercept	-6.6077	0.0146	-6.6363	-6.5792	<.0001
Match to AR flag	-2.1329	0.0038	-2.1403	-2.1255	<.0001
Multi-unit flag	-0.2321	0.0036	-0.2392	-0.2251	<.0001
In nonresponse followup flag	2.5617	0.0059	2.5501	2.5734	<.0001
Enumerator form-type flag	3.5556	0.0139	3.5283	3.5829	<.0001
Either neighbor vacant flag	0.6148	0.0045	0.6060	0.6236	<.0001
Block proportion of vacants	0.0702	0.0002	0.0698	0.0706	<.0001
Tract proportion of vacants	-0.0077	0.0002	-0.0082	-0.0073	<.0001

Table 3. Parameter Estimates for Household Size Model

Independent Variable	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Pr > ChiSq
			Lower	Upper	
Intercept	-1.0520	0.0014	-1.0548	-1.0492	<.0001
Tax-return person count	0.2457	0.0002	0.2453	0.2461	<.0001
Other AR file person count	0.2209	0.0006	0.2197	0.2220	<.0001
Tax-return adjusted person count	-0.2707	0.0033	-0.2771	-0.2643	<.0001
Other AR file adjusted person count	-0.2007	0.0023	-0.2051	-0.1963	<.0001
Mean census household size	0.2749	0.0005	0.2739	0.2759	<.0001
Older dependents counts	0.0865	0.0010	0.0845	0.0884	<.0001