Measuring the Discriminatory Power and Bias of Imputation Methods Designed for Imputing Status and Occupancy Status[1]

Yves Thibaudeau, Inez l. Chen, and Robert D. Sands
U.S. Census Bureau
Yves.Thibaudeau@census.gov

To compensate for missing data in surveys and censuses, the U.S. Census Bureau traditionally implements versions of the nearest-neighbor hot-deck. However, at this time an array of possible imputation methods is under investigation for status, occupancy status, and count imputation in the 2010 decennial census. In addition to the hot-deck, we are considering direct or modeled information retrieval from administrative records, and imputation based on statistical spatial models. In the investigative effort, we set to quantify the performance of these imputation methods. The investigation proceeds from a propensity analysis of the Census 2000 missing data pattern. In this paper we focus on performance measurements in terms of discriminatory power and bias. We quantify the discriminatory power of a method by the cross-product ratio (CPR) –i.e. the ratio of the odds of correctly imputing status or occupancy status. The CPR is tailored for situations where two values may be imputed. The two possible values for status are "delete" or "nondelete". The two possible values for occupancy status are "vacant" and "occupied". In addition to explaining the meaning of he CPR, the paper recalls the investigative role of another performance measurement, the bias.

1. Background

Housing unit (HU) status takes two possible values for each record on the address list:

1. Delete: this housing unit status characterizes an address or structure that is not a housing unit according to the definition of the Census Bureau.

2. Nondelete: this housing unit status characterizes any address or structure identifying a unique housing unit based on Census Bureau definition.

The occupancy status characterizes a nondelete housing unit by one of two values:

1. Occupied: this occupancy status characterizes a housing unit as being the usual residence for at least one individual on the Census Day.

2. Vacant: this occupancy status characterizes a housing unit that was not occupied as a usual residence by one individual on the Census Day.

This paper describes and discusses the merits of two imputation performance measures tailored to measure the performance of an imputation method for imputing unit status for housing units subjected to status or occupancy imputation:

1. The log odds, also called the log cross-product ratio (LCPR).

2. The bias.

When taking performance measurements based on these two measures, both the actual (or reported) value of the missing data and its imputed values must be available for a set of housing units and/or addresses. So, in order to take performance measurements, a simulation of the missing data mechanism for housing unit and occupancy status is conducted. The simulation artificially creates missing housing unit and occupancy statuses. The missing statuses are then imputed using specific imputation methods. For each imputation method the actual and observed values of the statuses are available. Given this information, the two performance measures presented in this memorandum measure the performance of the method.

The first performance measure –the LCPR– is directly linked to the number of correct and incorrect imputations at the address level. The concept behind the LCPR is best understood by first reviewing the definition of the cross-product ratio. The LCPR is just the logarithm of the

[1] Disclaimer: This report is released to inform the interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

cross-product ratio. But, the LCPR has better statistical properties than the cross-product ratio.

Given a method for imputing status, the cross-product ratio is the ratio between two odds:

1. The odds of imputing a "delete" vs. imputing a "nondelete" status when the true status of a unit/address is delete.

2. The odds of imputing a "delete" vs. imputing a "nondelete" when the true status of a unit/address is nondelete.

Similar interpretation is also applicable to imputing occupancy status.

The second performance measurement is the bias. It centers on the concept of error balancing over a geographical area. For this research the geographical areas are the states.

For a given imputation method, the bias is the expected difference between number of deletes erroneously imputed as nondeletes and the number of known nondeletes erroneously imputed as deletes. A mathematical definition for the bias is given in section 3.

These two measures are complementary. The LCPR measures the accuracy of an imputation method at the unit level. But, the bias measures accuracy at the level of a geographical area. In this paper, the geographical areas are the states.

Note that an imputation method can have a high LCPR, therefore it is accurate at the unit level, and still produce a large bias over a given geographical area, such as a state. The converse is also true. Chambers (2001) suggests that accuracy at aggregate level – in this case the state level - takes precedence over accuracy at the unit level. This memorandum does not assume such a hierarchy. It focuses on defining estimators for these measurements in the context of the simulation.

Formal definitions of the LCPR and the bias and appropriate estimators are given in section 3.

2. Missing Data Simulations Based on Census 2000

To estimate the LCPR and bias for several imputation methods, a simulation of the missing data patterns observed in Census 2000 is conducted. The simulation is a set of 100 repetitions of simulated missing HU status and missing occupancy status, at the state level. A detailed description of the simulation is given in DSSD 2006 Census Test Memorandum Series J2-03.

Each state included in the simulation is divided into several "equal-propensity cells". The objective when creating equal-propensity cells is to group the HU's and/or addresses in subsets such that each entity in a subset has the same probability of having a missing HU status, or missing occupancy status. The cells are constructed based on the values of several covariates collected or created during Census 2000.

Then, in each cell, the units are sub-sampled at a rate equal to the rate HU status (or occupancy status) is observed to be missing. The sampled units are flagged. The actual or reported values of HU status for the flagged units and the corresponding imputed values produced by an imputation method provide the information for estimating the LCPR and the bias of the method.

The full implementation of the simulation repeats the sub-sampling cycle in each equal-propensity cell 100 times. For each imputation method, this process leads to 100 estimates of the LCPR and bias for HU status and occupancy status. The 100 are combined to produce a final estimate for each performance measurement.

3. Formal Definitions

This section defines the LCPR and the bias in the context of a parametric model. Later sections present the natural estimators for the LCPR and the bias. To simplify the presentation, only status imputation is considered. The application of the model to occupancy imputation is similar.

Suppose addresses with known status are sub-sampled from an equal-propensity cell. Then a specific imputation method imputes the statuses of each sampled address. So, these addresses have an observed or reported status and an imputed status.

For one such address there are four possible outcomes. The outcomes are each represented by

a cell of Table 1. The probabilities for each of these outcomes are $p^{1,1}$, $p^{2,1}$, $p^{1,2}$, $p^{2,2}$.

Table 1 – Probabilities of Observed and Imputed Status for a Random Address

		Imputed Status		
		Del	*Non delete*	*Delete+ Nondelete*
Obs Status	*Del*	$p^{1,1}$	$p^{1,2}$	$p^{1} = p^{11} + p^{12}$
	Non delete	$p^{2,1}$	$p^{2,2}$	$p^{2} = p^{21} + p^{22}$

The following conditions hold:

$$0 < p^{1,1}, p^{1,2}, p^{2,1}, p^{2,2} < 1$$
$$p^{1,1} + p^{1,2} + p^{2,1} + p^{2,2} = 1 \qquad (3.1)$$

For each address selected at random, $p^{1,1}$ is the probability that the address is an observed "delete" and its imputed status is also a "delete". Similar interpretations apply for the other probabilities in Table 1. Given this notation, formal definitions are given.

Definition 1. The Cross-Product Ratio

The *cross-product ratio (or odds ratio)* Φ is

$$\Phi = \frac{p^{1,1}\, p^{2,2}}{p^{1,2}\, p^{2,1}} \qquad (3.2)$$

Definition 2. The Log-Cross-Product Ratio

The *log cross-product ratio (or log odds ratio)* Γ is

$$\Gamma = \log(\Phi) = \log\left(\frac{p^{1,1}\, p^{2,2}}{p^{1,2}\, p^{2,1}}\right) \qquad (3.3)$$

Definition 3. The Bias

The *bias* B is

$$B = N\left(p^{1,2} - p^{2,1}\right) \qquad (3.4)$$

N is the number of addresses for which status had to be imputed.

There are three distinct ranges of values for Γ and Φ. Each range has an intuitive interpretation as it quantifies the accuracy of the imputation method.

The interpretation of the LCPR is as follows:

1. $\Gamma = 0$ (Or equivalently $\Phi = 1$)

This LPCR value identifies an imputation technique not extracting any discriminatory knowledge from the available information

2. $0 < \Gamma < \infty$ (Or equivalently $1 < \Phi < \infty$)

A LCPR in this range points to an imputation technique using the available information advantageously. The higher the LCPR is, the higher the discrimination power.

3. $-\infty < \Gamma < 0$ (Or equivalently $0 < \Phi < 1$)

This situation is unlikely for a reasonable imputation technique. A LCPR in this range suggests the imputation technique is purposely designed to produce incorrect imputations.

In practice the imputation method may make use of information on the nearest neighbor, or information mined from administrative records. The LCPR effectively quantifies the accuracy of the method in terms of its power to discriminate between deletes and nondeletes.

The remaining sections give estimators for the LCPR and the bias. Appropriate variance estimators and confidence intervals for the LCPR and the bias are derived.

4. Single-iteration Estimator of the Log Cross-product Ratio and Its Variance

The goal of a simulation as described above is to estimate the probabilities in Table 1. More efficient estimates are available when the simulation is repeated several times, relative to estimates obtained from a single simulation. These more efficient estimates are composite;

i.e. they are constructed by pooling the single-simulation estimates.

This section presents the derivation of estimates based on a single simulation. Composite estimators are derived in subsequence sections.

Let N_i be the number of units designated for status imputation from a given state at simulation *i*. In the document, it is assumed that the total number of simulations is 100. So, $i = 1, \ldots, 100$. Given simulation *i*, the sample equivalent of Table 1 is exhibited in Table 2.

Table 2 - Counts of Addresses Cross-Classified by Observed and Imputed Status (Simulation *i*)

		Imputed Status		
		Del	*Non del*	*Delete+ Nondelete*
Obs Status	*Del*	$N_i^{1,1}$	$N_i^{1,2}$	$N_i^1 = N_i^{1,1} + N_i^{1,2}$
	Non Del	$N_i^{2,1}$	$N_i^{2,2}$	$N_i^2 = N_i^{2,1} + N_i^{2,2}$

The following holds

$$N_i = N_i^{1,1} + N_i^{1,2} + N_i^{2,1} + N_i^{2,2} \qquad \textbf{(4.1)}$$

This notation allows for the formulation of the formal model underlying each simulation.

Conditional on N_i^1 and N_i^2, the following model generates the other entries in Table 1**:**

$$N_i^{1,1} \sim Binomial\left(N_i^1 , {p^{1,1}}/{p^1} \right) \qquad \textbf{(4.2)}$$

$$N_i^{2,1} \sim Binomial\left(N_i^2 , {p^{2,1}}/{p^2} \right) \qquad \textbf{(4.3)}$$

This model leads to defining a natural estimator for the LCPR Γ:

$$\hat{\Gamma}_i = \log\left(\frac{N_i^{1,1} N_i^{2,2}}{N_i^{1,2} N_i^{2,1}} \right)$$

$$= \log\left(N_i^{1,1}\right) + \log\left(N_i^{2,2}\right) - \log\left(N_i^{1,2}\right) - \log\left(N_i^{2,1}\right) \qquad \textbf{(4.4)}$$

An estimator for the variance of $\hat{\Gamma}_i$ conditional on $\left(N_i^1, N_i^2\right)$ is derived by linearization (Valliant Rust, 2003; Bishop Fienberg Holland, 1975 p. 377):

$$\hat{Var}\left(\hat{\Gamma}_i \,\middle|\, N_i^1, N_i^2\right) = \frac{1}{N_i^{1,1}} + \frac{1}{N_i^{1,2}} + \frac{1}{N_i^{2,1}} + \frac{1}{N_i^{2,2}} \qquad \textbf{(4.5)}$$

The expression on the right hand side (RHS) of (4.5) has only four terms because the covariance terms involved in the linearization simplify.

5. Multiple-iteration Estimator of the LCPR and Its Variance

The preceding section defined $\hat{\Gamma}_i$, an estimator of Γ based on a single simulation *i*. It also defined $\hat{Var}\left(\hat{\Gamma}_i \,\middle|\, N_i^1, N_i^2\right)$, an estimator of the variance of $\hat{\Gamma}_i$ conditional on N_i^1 and N_i^2, the reported numbers of deletes and nondeletes in simulation *i*.

In this section, these definitions are extended to obtain estimators based on 100 simulations. A natural estimator of Γ integrating information from all the simulations is the average of the $\hat{\Gamma}_i$'s:

$$\hat{\bar{\Gamma}} = \frac{\sum_{i=1}^{100} \hat{\Gamma}_i}{100} \qquad \textbf{(5.1)}$$

An estimator for the variance of $\hat{\bar{\Gamma}}$ can serve to construct confidence intervals for $\hat{\bar{\Gamma}}$. The following variance estimator is derived for that purpose.

$$\hat{Var}\left(\hat{\bar{\Gamma}}\right)$$

$$=\left(\frac{1}{100}\right)^2 \sum_{i=1}^{100}\left[\hat{Var}\left(\hat{\Gamma}_i \,\middle|\, N_i^1, N_i^2\right)+\left(\hat{\Gamma}_i-\hat{\bar{\Gamma}}\right)^2\right] \quad \textbf{(5.2)}$$

A 90% confidence interval for Γ is

$$\left[\ \hat{\bar{\Gamma}}\ -\ 1.64\sqrt{\hat{Var}\left(\hat{\bar{\Gamma}}\right)},\ \hat{\bar{\Gamma}}\ +\ 1.64\sqrt{\hat{Var}\left(\hat{\bar{\Gamma}}\right)}\ \right] \quad \textbf{(5.3)}$$

6. Single-iteration Estimator of the Bias and Its Variance

Conditional on the row totals $\left(N_i^1, N_i^2\right)$, an estimator for the bias of a given imputation method is

$$\hat{\mathrm{B}}_i\ =\ N_i^{1,2}\ -\ N_i^{2,1} \quad \textbf{(6.1)}$$

An estimator for the variance of $\hat{\mathrm{B}}_i$ is derived from an identity. Conditional on the row totals $\left(N_i^1, N_i^2\right)$, the following holds.

$$Var\left(\hat{\mathrm{B}}_i \,\middle|\, N_i^1, N_i^2\right)$$

$$=Var\left(N_i^{1,2} \,\middle|\, N_i^1, N_i^2\right)+Var\left(N_i^{2,1} \,\middle|\, N_i^1, N_i^2\right) \quad \textbf{(6.2)}$$

A computationally tractable formula for an estimator of $Var\left(\hat{\mathrm{B}}_i \,\middle|\, N_i^1, N_i^2\right)$ is obtained by substituting basic binomial estimators for the variances on the RHS of (6.2). The resulting variance estimator is given bellow.

$$\hat{Var}\left(\hat{\mathrm{B}}_i \,\middle|\, N_i^1, N_i^2\right) \quad \textbf{(6.3)}$$

$$=N_i^{1,2}\left(1-\frac{N_i^{1,2}}{N_i^1}\right)+N_i^{2,1}\left(1-\frac{N_i^{2,1}}{N_i^2}\right)$$

7. Multiple-iteration Estimator of the Bias and Its Variance

The estimators are

$$\hat{\bar{\mathrm{B}}}\ =\ \frac{\sum_{i=1}^{100}\hat{\mathrm{B}}_i}{100} \quad \textbf{(7.1)}$$

$$\hat{Var}\left(\hat{\bar{\mathrm{B}}}\right)$$

$$=\left(\frac{1}{100}\right)^2 \sum_{i=1}^{100}\left[\hat{Var}\left(\hat{\mathrm{B}}_i \,\middle|\, N_i^1, N_i^2\right)+\left(\hat{\mathrm{B}}_i-\hat{\bar{\mathrm{B}}}\right)^2\right] \quad \textbf{(7.2)}$$

The template for the derivation of (7.2) is given in Appendix B.

A 90 % confidence interval for B is

$$\left[\hat{\bar{\mathrm{B}}}-1.64\sqrt{\hat{Var}\left(\hat{\bar{\mathrm{B}}}\right)},\ \hat{\bar{\mathrm{B}}}+1.64\sqrt{\hat{Var}\left(\hat{\bar{\mathrm{B}}}\right)}\ \right] \quad \textbf{(7.3)}$$

8. General Trends for Three Status Imputation Methods

The lCPR and bias were estimated for three status imputation methods.

1. Administrative Record Modeling

2. Log-Linear Spatial Modeling

3. Basic Hot-Deck

Administrative Record Modeling (ARM) predicts the missing status of a unit on the basis of information extracted from administrative records, and information available from census data. The information extracted from administrative records could be only the knowledge of whether or not there exists a matching record for a specific HU. This information can be useful because we expect a HU with no matching record to be more likely to have a delete HU status.

Log-Linear Spatial Modeling (LLSM) draws information only from census measurements to predict status. Thibaudeau (2002) propose an

approach for LLSM in the context of item imputation, similar to that used in the current situation for HU status imputation. This approach for LLSM is a simplified version of spatial modeling as proposed by Besag (1974).

The version of the Hot-Deck (HD) in this case is a bilateral sequential hot-deck (Fay 1999). It imputes HU status by borrowing it from the nearest qualifying neighbor, in the order of the census master address file.

The results of the simulation on selected states suggest ARM outperformed both LLSM and HD for the LCPR of HU status imputation. This means administrative record databases, supported by modeling, deliver superior predictive power, at the unit level.

At the same time, the results of the simulation suggest HD outperform the other methods for the bias. This may be the result of clustering. Most deletes are within short distance of each other. So, they form clusters. In these clusters, because the sequences of delete/nondelete are symmetric, the falsely imputed deletes and falsely imputed nondeletes produced by the HD tend to balance each other.

9. Discussion

The paper presents tools that were used for investigating methods for imputing status and occupancy status in the context of accuracy at the unit and at the state level. A simulation of plausible missing data mechanisms in a census enumeration was the basis for quantifying the two types for three methods of imputation.

It is important to be aware of the limitations of the simulation. The missing data mechanisms in the simulation are based on a theoretical assumption. It is assumed data are missing at random (MAR). The actual missing data mechanisms during Census 2000 probably do not conform to this assumption.

The results of the paper are realistic only to the extent the imputation methods are robust to changes in assumptions. More research, including sensitivity analyses, is necessary to assess the robustness of the methods and the generality of the results.

References

Besag, J. (1974). Spatial Interaction and Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society,* B, **36**, 2.

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press.

Chambers, R. (2001). Evaluation Criteria for Statistical Editing and Imputation. *National Statistics Methodological Report* (Unumbered).

Fay, R. E. (1999). Theory and Application of Nearest Neighbor Imputation in Census 2000. *Proceedings for the Section on Survey Research Methods, American Statistical Association.*

Thibaudeau, Y. (2002). Model Explicit Item Imputation for Demographic Categories. *Survey Methodology*, 19, 1.

Valliant, R., Rust, K. (2003). *Inference for Complex Surveys*. Lecture Notes.

Williams, T. (2005). *Memo to Donna Kostanich, Assistant Division Chief, Decennial Statistical Studies Division. Series J2-03*