2010 Census Count Imputation – Research Results using Spatial Modeling

Robert D. Sands and Richard A. Griffin, U.S. Census Bureau[1]
Robert D. Sands, robert.d.sands@census.gov, (301) 763-4255

Keywords: Hot Deck, Categorical Data Model, Truth Deck

1. Introduction

After the completion of all 2010 Census data collection operations a small number of housing units remain for which either the count of the number of residents, or the housing unit status as occupied, vacant or non-existent is not known. In order to improve the accuracy of the count of persons and housing units in the U.S., it is necessary to impute this missing data.

Section 2 provides background about the Census. Section 3 briefly describes the traditional method of count imputation used in the Census. The spatial modeling method is described in Section 4. The truth deck used for the research is described in Section 5. The results of the research and comparisons of spatial modeling with traditional methods are discussed briefly in Section 6. Conclusions are drawn in Section 7 and Section 8 discusses limitations of the research.

2. Background

At the beginning of each decade, the U.S. Census Bureau conducts a decennial census to obtain the population count for congressional apportionment, distribution of federal funds, and redistricting for the following ten years. The first question on the 2010 Census questionnaire determines how many people were living in this housing unit (housing unit size) on, Census Day, April 1, 2010. The remaining questionnaire items ask the name, sex, age, date of birth, Hispanic origin, race, and relationship to the householder of every person residing in all occupied housing units. In addition, the telephone number and tenure (owner or renter) of the householder is determined.

The 2010 Census is a massive and complex operation. Not every occupied housing unit provides all the requested information. For the 2010 Census, the Census Bureau will conduct imputation procedures for housing unit records lacking a status designation of occupied, vacant, or nonexistent as well as impute the housing unit size for housing units which are known or are imputed to be occupied but lack a population count. These procedures, collectively, are referred to as "count imputation". In the 2000 Census, about 0.55% or 691 thousand housing units required count imputation (Alberti, 2004). After count imputation is complete the "characteristic edit and imputation" is performed. This operation will perform edits of each of the 2010 Census questionnaire items and provide imputed values for questionnaire items that are missing or that failed the edit.

3. Hot Deck

Since the 1960 Census, the Census Bureau has been using an imputation method commonly known as the "hot deck" or the "nearest neighbor hot deck" (Fay, 1999) to perform count (Kilmer, 2002) and characteristic imputation. In the nearest neighbor hot deck housing units are ordered by address identifier to maintain geographic proximity of the sequence. A "hot deck" is maintained of the most recently encountered housing units that qualify as "eligible donors". When a housing unit is encountered that is missing the required information ("donee") the most recent eligible donor that shares certain other characteristics with the donee then imputes its information to the donee. The methodology has worked well, but is not without its limitations.

One problem is that the hot deck is not based on any explicit model and only recently have the theoretical

[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 authors and not necessarily those of the U.S. Census Bureau.*

properties of the methodology been closely examined or a proposed variance estimation procedure been rigorously justified (Chen and Shao, 2000). In addition, recent developments have raised the possibility that another approach might perform better than the hot deck. These developments include: i) an extensive statistical literature for handling missing data (Thibaudeau, 2002; Little and Rubin, 2002; Allison (2001); Kalton and Kasprzyk, 1986), ii) the potential use of administrative data to supplement 2010 Census operations (Bauder and Johnson, 2003; Bye and Judson, 2004) and iii) considerable advances in computer technology (Sands, 2003)

In preparation for the 2010 Census, the Census Bureau launched an extensive research effort to study the possibility of using an alternative imputation methodology to the hot-deck (Chen, 2005; Griffin, 2004). These alternatives included modified versions of the hot deck, administrative records-based methods, and methods based on the spatial modeling approach. This paper is focused mainly on the Spatial Models with limited comparison to a hot deck approach.

4. Spatial Modeling

During the period 2004-2006, an investigation of a spatial modeling approach for the imputation of missing housing unit size, occupancy or housing unit status was undertaken using the housing unit and person data from the 2000 Census. The methodology is defined in Thibaudeau (2002), which concerns the imputation of demographic categorical variables. This approach is based on modeling the conditional probability for the status of a housing unit given the status of its closest completely reported neighbor. Instead of looking for the best individual nearest neighbor housing unit, as in the hot deck, a model generates imputations based on the information available from all the nearest neighbor associations in the local area such as a census tract.

In the current discussion, each housing unit address in a census tract is linked with its nearest neighbor whose housing unit size information is completely reported. The linking of a housing unit with its closest completely reported neighbor of the same structure type (single- or multi-unit) is achieved through the ordering of the 2000 Census housing unit file by a numerical address identifier. This numerical address identifier maintains the geographic proximity of housing units.

At the census tract level, the distribution of estimated probabilities for the missing housing unit size/status is obtained from the empirical frequency of that item among all completely reported housing units sharing the same type of nearest neighbor and other characteristic(s) with the housing unit missing housing unit size. For each housing unit requiring imputation of the missing item, a random draw is made from the aforementioned probability distribution to assign a particular value of the missing item. This arrangement has the effect of capturing the association or transition probability of each housing unit with its nearest neighbor within a tract. The modeling of the local nearest neighbor relationship gives the methodology its spatial aspect.

The spatial modeling methodology discussed in this paper was constructed for the 2010 Census count imputation research using the preliminary unedited 2000 Census short form data set combined with the final edited 2000 Census short form data set (Kilmer, 2004).

A categorical data model was designed using attributes of the housing unit and those of the nearest neighbor housing unit not requiring count imputation. The primary objective was to select variables that are good predictors of housing unit size. In addition, the resulting model design could only employ three or four variables because using more variables would lead to small cell counts in the diminutive census tract. Such model over-specification would then produce excessive perturbation in the cell probabilities.

4.1 Model Variable Selection Procedure

First, the selection process (Griffin, 2005) examined the bivariate relationship of each of eighteen “predictor” variables, individually, with housing unit size. The analysis of a series of bivariate relationships was employed first to determine the several variables to be used for the creation of a model in order to eliminate the unproductive variables early and ultimately to affect a more efficient use of resources.

The bivariate analyses used a cross classification of all the 2000 Census housing units not requiring count imputation within each census tract by housing unit size and the particular predictor variable. The uncertainty coefficient, U, (SAS Institute, 1999; Agresti, 2002; Theil, 1970) was calculated, using (1), for each predictor variable within each of the Census tracts in the U.S.

$$U = \frac{\sum_i \sum_j \pi_{ij} \log\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}} \right)}{\sum \pi_{+j} \log\left(\pi_{+j}\right)} \qquad (1)$$

Where π_{ij} is the empirical probability that the housing unit has the ith housing unit size and the jth value of the predictor variable.

The uncertainty coefficient can range from 0 to 1 with higher values indicating better predictors. The median U across all Census tracts was used to rank the variables on their ability to predict housing unit size (Table 1). It should be noted that the analysis employed *all* U.S. housing units not requiring count imputation in the 2000 Census and therefore no standard errors are applicable.

The eighteen prospective predictor variables are housing unit mail return status, housing unit structure type, housing unit address type, nearest neighbor household type, nearest neighbor housing unit size, nearest neighbor housing unit tenure, nearest neighbor housing unit (householder) race, nearest neighbor housing unit (householder) Hispanic origin, nearest neighbor housing unit presence of minor children, nearest neighbor housing unit presence of persons over 65, and eight collection block-level statistics.

The categories for housing unit mail return status are: housing unit did return a 2000 Census questionnaire and housing unit did not return a 2000 Census questionnaire. For housing unit structure type they are: housing unit is in a single-unit structure and the housing unit is in a multi-unit structure. For housing unit address type the categories are: city-style address, rural route address, Post Office box, incomplete address, and missing address.

In the following, the identical set of categories exist for both the housing unit and the nearest neighbor housing unit. The ten categories for household type are: non-existent, vacant, married-couple family, male-householder family, female-householder family, male-householder non-family, male-householder living alone, female-householder living alone, female-householder non-family, and occupied but some relevant person characteristics unknown. The nine categories for housing unit size are: non-existent, vacant, 1, 2, 3, 4, 5, 6, and 7+ persons. The categories for tenure are: owner and renter. The categories for race of the householder are: White, Black, American Indian/Alaska Native, Native Hawaiian/Pacific Islander, and Other Race. The categories for Hispanic origin of the householder are: Hispanic and non-Hispanic. The categories for presence of children are: no children under 18 present and one or more children under 18 present. The categories for presence of persons 65+ are: no persons 65+ and at least one person 65+.

The eight collection block-level statistics designate a collection block as either high or low within the state on the proportion of i) vacant housing units, ii) non-existent housing units, iii) multi-unit structure type, and iv) mail return status using two high/low cut-off percentiles: 0.75 and 0.90, respectively. For example, if a collection block is above the 75th percentile within the state for percentage of vacant housing units then all the housing units in that collection block are designated as high for block-level housing unit vacancy.

4.2 Spatial Model Variable Selection

Table 1. Median Tract-level Uncertainty Statistics, U, from the Two-Way Classifications of All U.S. Census 2000 Fully-Reported Housing Units By Housing unit size and each of Eighteen Housing unit Characteristics
(N=60 Thousand Census Tracts containing approximately 120 Million Housing units)

Housing unit Characteristic	**U**
housing unit mail return status	**0.047**
nearest neighbor household type	**0.030**
nearest neighbor housing unit size	**0.029**
nearest neighbor housing unit ten.	**0.018**
nearest neighbor housing unit race	**0.016**
nearest neighbor housing unit children present	**0.015**
nearest neighbor housing unit persons over 65	**0.013**
nearest neighbor housing unit Hispanic origin	**0.012**
housing unit structure type	**0.012**
block non-existent housing units-75th percentile	**0.011**
block vacant housing units-75th percentile	**0.010**
block multi-unit structure type-75th percentile	**0.008**
block vacant housing units-90th percentile	**0.007**
block non-existent housing units-75th percentile	**0.006**
block multi-unit structure type-90th percentile	**0.006**
block mail return status-75th percentile	**0.005**
housing unit address type	**0.004**
block mail return status-90th percentile	**0.003**

The mail return status of the housing unit was at the top of the ranking with a U = 0.047. Nearest neighbor housing unit size and nearest neighbor household type finished next on the list with uncertainty statistics in the 0.03 range. Other nearest neighbor housing unit classifications such as the presence of minor children, presence of persons over 65, householder tenure, Hispanic origin and race had 0.012 <= U <= 0.018. All eight of the collection block statistics had a U <= 0.011. The type of address of the housing unit finished near the bottom with a U=0.004. Lastly, it should be noted that housing unit structure type is embedded in the definition used to create all the nearest neighbor variables. That is, the nearest neighbor is defined as the nearest housing unit, not requiring count imputation, of the same structure type. Therefore, the nearest neighbor prefix variables also implicitly reflect the spatial association within structure type.

The relative predictive power of the mail return variable appears to reflect the fact that receipt by mail, at the Census Bureau, of a 2000 Census questionnaire almost never originated from a vacant or non-existent housing unit.

Also, a modest spatial association of housing unit size (as well as of household type) among neighboring housing units exists at the tract-level.

Consequently, i) *mail return*, ii) *nearest neighbor housing unit size*, iii) *nearest neighbor household type*, and iv) *housing unit structure type* (included for historical reasons) were selected as the prospective variables from which to construct categorical data models to be used for the remainder of the research.

4.3 Spatial Model Construction

Several different hierarchical log linear models were then constructed from the four variables listed above along with the housing unit size variable (defined only for fully reported housing units). The models used different combinations of the variables and also featured unsaturated (reduced) versions. The reduced models were tested because of the prevailing view exemplified by Agresti (2002, page 316) that "*In practice, unsaturated models are preferable, since their fit smoothes the sample data and has simpler interpretations*".

Since the nearest neighbor housing unit size and the nearest neighbor household type variables both cover aspects of housing unit size it was necessary to choose only one of these for the log linear modeling. This occurred because the high correspondence between the two variables leads to the existence of a number of structural zero cells in the model (e.g., a housing unit cannot both be nearest neighbor household type = 1 (Vacant) and nearest neighbor housing unit size = 2 persons). The structural zero cell situations, in turn, lead to difficulty in reaching convergence for log linear models. The variable nearest neighbor household type was selected (and nearest neighbor housing unit size eliminated) because it performed better in terms of the uncertainty statistic. Note that housing unit size is the variable we are trying to predict in the spatial modeling. Also note that we did not choose any models using any of the predictor variables with rank lower than third in Table 1 except for structure type. As stated earlier, selecting more variables could create too many sparse cells in some tracts and also increase model complexity.

Consequently, we decided to simulate the following three spatial model count imputation procedures on the tracts in the three test states. Only three test states were employed initially because several models, most notably model 3 below, required significant computer resources for the simulations.

i) *Three variable saturated model* using housing unit size, mail return status and nearest neighbor household type. Spatial Modeling performs imputation within "cells" established in each census tract. In the 3 variable model, the cells are defined as a cross–classification of housing units on the two levels of mail return status and the ten levels of nearest neighbor household type. Each of these twenty cells is further divided into nine housing unit sizes. The cross-tabulation of the housing units not requiring count imputation provides empirical counts of housing units falling into each cell and housing unit size category. In a cell, the proportion of reported (non-count imputation) housing units with a particular size is multiplied by the number of missing (count imputation) cases. This product is then added to the count of reported housing units to yield the expected frequency of cases in a size category in a cell. These expected frequencies are used to estimate the **probabilities** that a housing unit requiring count imputation falls into a particular housing unit size category. Finally, the vector of cell housing unit size probabilities is used to make a random draw of housing unit size for each unit requiring count imputation:

$$\hat{x}_{ijk} = r_{ijk} + \frac{r_{ijk}}{\sum_{s} r_{ijs}} m_{ij} \qquad \tilde{\pi}_{ijk} = \frac{\hat{x}_{ijk}}{\sum_{s} \hat{x}_{ijs}} \qquad (2)$$

where: $r, m, \hat{x}, \tilde{\pi}$ are, respectively, the empirical counts of reported (non-count imputed) housing units, empirical counts of missing (count imputed) housing units, expected frequency, and estimated probability. The symbols i, j designate the ith mail return, jth nearest neighbor household type, respectively. The symbols k, s both designate the housing unit size category.

ii) *Four variable saturated model* using housing unit size, mail return status, nearest neighbor household type**,** and structure type (single-unit and multi-unit).

iii) *Three variable reduced model* automated selection procedure using housing unit size, mail return status and nearest neighbor household type. For each tract the selection procedure (Griffin, 2006) worked as follows: i) The first model attempted is the conditional independence model. This model indicates that mail return is related to housing unit size and nearest neighbor household type is related to housing unit size, but mail return status is not associated with nearest neighbor type conditional on housing unit size. If the Likelihood Ratio Goodness of Fit Chi-Square statistic (LR χ^2) is not significant at the 5% significance level (i.e., the null hypothesis of a good model fit is not rejected), then the conditional independence model is used for this tract, ii) Otherwise, the all two-way interaction model is attempted. This model has all three two-way interactions, but no three-way interaction. If LR χ^2 is not significant at the 5 percent level, then the all two-way interaction model is used, iii) Otherwise, the three variable saturated model, which includes the one three-way interaction, is used.

5. Truth Deck

To objectively evaluate the various imputation models it was necessary to create a housing unit data set known as the "truth deck". A detailed description of the process of truth deck creation is available in Williams (2005). A brief description of the process follows.

The intent of the truth deck was to reflect the missing data patterns that occurred in the short form data set resulting from 2000 Census operations. The first step was to determine, for each state, using the full 2000 Census data set a collection of "equal propensity cells" that partition housing units into groups that have the same probability of requiring count imputation. For each state, various 2000 Census operational, housing unit and geographical variables defined the equal propensity cells. In the second step, the subset of 2000 Census housing units not requiring count imputation underwent a simulation of the missing data pattern using the propensity cell probabilities. Specifically, for each fully reported housing unit with covariates identifying it as belonging to a particular equal propensity cell, a random draw was taken from the uniform [0,1] distribution. If the random draw was less than or equal to the propensity cell probability the housing unit's simulated status was flagged as requiring count imputation. Otherwise the housing unit retained its designation as not requiring count imputation. This simulation was repeated 100 times creating 100 replications of simulated count imputation status for each fully reported housing unit. The data set containing each fully reported housing unit along with the housing unit's 100 count imputation status "flags" was referred to as the truth deck.

This set-up allowed each particular count imputation method or model to be run on the truth deck producing for each flagged housing unit 100 replications of an imputed housing unit size (non-existent, vacant, 1, 2, 3, 4, 5, 6, 7+ persons). The housing unit size chosen by the particular method was then compared to the housing unit's actual reported size. Various statistical measures of the method's accuracy relative to the truth deck were then computed. Finally, these accuracy statistics were used to compare imputation methods.

6. Comparison of Spatial Modeling Models with the Hot Deck Models

Two phases of the comparison of the imputation method alternatives were conducted. The first used only three states (Kilmer, 2006a; 2006b) while the second employed all 50 states plus the District of Columbia (Kilmer, 2005). The results observed in the evaluation statistics in the two phases were very similar.

In the first phase, three hot deck methods and the three spatial models described in Section 4.3 were compared. Following this phase, two methods were eliminated from further comparison. The eliminated hot deck method had featured some minor modifications that did not distinguish it from the 2000 Census hot deck. The three variable reduced spatial model was also eliminated because its performance relative to the two saturated spatial models did not justify the substantial processing resources required (Griffin, 2006).

The discussion of the results that follows will focus on the 50 states plus the District of Columbia (all states) comparison.

The all states comparison looked at two spatial models: i) three variable saturated model using housing unit size, mail return and nearest neighbor household type and ii) four variable saturated model using housing unit size, mail return, nearest neighbor household type**,** and structure type. In addition, two hot deck methods were compared: i) the 2000 hot deck and ii) the modified hot deck. The modified hot deck benefited from several changes suggested by preliminary findings in the 2010 Census research effort.

Several evaluation statistics were used to rank the various models. Although numerous classifications are possible, the statistics can be, for our purposes, divided into two major groups: i) *aggregate* - measures of closeness for overall counts (or shares of the total) for geographic areas or groups and ii) *individual* - measures of closeness for each individual case.

The most straightforward aggregate measure of accuracy was how well the particular model, relative to the 270 million person truth deck, counted the population. The two spatial models were the closest to the reported truth deck population total. A similar measure of the accuracy of aggregate counts, for sub-national areas (Sands and Kohn, 2006), replicated these results. Also, Kilmer (2005) using the non-parametric Friedman two-way analysis of variance by ranks (Chen et al, 2005) showed that when the four imputation alternatives are ranked within each state by their absolute difference with truth deck population count, the two spatial models finish on top.

Insight into the apparent superiority in aggregate accuracy of the spatial models compared to the hot deck methods can be gained from an examination of the individual accuracy results.

Two key measures of individual accuracy are the nominal and ordinal statistics. The nominal statistic quantifies the number of errors made between the reported and imputed housing unit size. The ordinal statistic assigns a penalty based on the magnitude of difference between reported and imputed housing size.

Again employing the Friedman ranks analysis, when these four imputation methods were ranked in each state for the nominal as well as for the ordinal statistics, the modified hot deck finished higher in most states for the nominal statistic and for the ordinal statistic. The three variable spatial model was at or near the bottom of the ranking for all states.

At the housing unit level the three variable (and to a lesser extent the four variable) spatial model is the least able to correctly impute a particular housing unit's size. On the other hand, both hot decks tend to overcount vacant housing units and undercount occupied housing units.

7. Conclusion

Even though the spatial model is wrong more often for individual housing units the errors tend to cancel out, whereas the hot decks apparently have a tendency to impute a vacant housing unit too often leading to errors that predominate on the low side. The reason for this is unclear.

The result demonstrated in this paper does replicate those found by Thibaudeau (2002). In that study of the use of spatial modeling to impute tenure of the householder (using data from the 2000 Census Dress Rehearsal), Thibaudeau found that a spatial modeling approach was more effective than the hot deck at handling distributional differences between the housing unit and its nearest neighbor in the characteristic being imputed.

8. Limitations

The truth deck, to which all imputation methods are compared, is based on the 2000 Census data set and therefore the truth is relative to the character of the 2000 Census. Further, the truth deck's construction is based on the assumption that the subset of Census records that were in need of count imputation and therefore not in the truth deck reflect the characteristics of the fully reported housing units contained in the truth deck. The truth deck assumes that the housing units needing count imputation are missing at random.

9. References

Alberti, N.S. (2004). *Data Processing in Census 2000*. Census 2000 Testing, Experimentation, and Evaluation Program Topic Report No.7, TR-7. Washington, DC: U.S. Census Bureau. http://www.census.gov/pred/www/rpts/TR7.pdf

Allison, P.D. (2001). *Missing Data*, Series on Quantitative Applications in the Social Sciences, 07-136, Thousand Oaks, CA: Sage.

Agresti, A. (2002). *Categorical Data Analysis, Second Edition*, New York: Wiley-Interscience.

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

Bye, B. V., and D.H. Judson. (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.

Bishop, Y.M.M., Feinberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*, Cambridge, MA: The MIT Press.

Chen, I. (2005). *2006 Census Test: Imputation Research Plan*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J1-02, January 25, 2005.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, **16**, 113-131.

Chen, I., Kilmer, A.D., Shores, R.W., and Seiss, M. (2005). *2010 Census Imputation Research : Statistics to be Computed for Count Imputation Research Options*, DSSD 2006 CENSUS TEST

MEMORANDUM SERIES J2-xx.

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, 112-121.

Griffin, R.A. (2004). Potential Methodologies for Count Imputation for the Decennial Census, *Proceedings for the Section on Survey Research Methods*, American Statistical Association.

Griffin, R.A. (2005). *2010 Census Imputation Research – Variable Selection for Spatial Modeling*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-06, December 21, 2005.

Griffin, R.A. (2006). *2010 Census Imputation Research – Documentation of Spatial Models*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-10, April 5, 2006.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, **12**, 1-16.

Kilmer, A.D. (2002). *Census 2000 Specifications for Imputing Housing Unit Status and Population Counts*, DSSD CENSUS 2000 PROCEEDURES AND OPERATIONS MEMORANDUM SERIES #Q-79.

Kilmer, A.D. (2004). *2010 Census Imputation: Imputation Research Files*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J1-01, October 14, 2004.

Kilmer, A.D. (2005). *2010 Census Count Imputation Research: Results and Analysis of Count Imputation Methodologies for All States*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-??, December 13, 2005.

Kilmer, A.D. (2006a). *2010 Census Count Imputation Research: Results and Analysis of Count Imputation Methodologies for Three States*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-??, May 2, 2006.

Kilmer, A.D. (2006b). Results of Research of Count Imputation Methods for Selected States. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, forthcoming.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, New York: Wiley.

Sands, R.D. (2003). A Simple and Efficient Approach to the Cross-tabulation of Large SAS® Data Sets. *Proceedings of the NESUG 16 Meeting*, NorthEast SAS Users Group, cc003. http://www.nesug.org/html/Proceedings/nesug03/cc/cc003.pdf

Sands, R.D. and Kohn, F.E. (2006). *2010 Census Imputation Research – Evaluation of Population Shares for States and Congressional Districts for Alternative Count Imputation Methodologies*. DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-12. July 11, 2006.

Sands, R.D. and Shores, R.W. (2006). *2010 Census Imputation Research – Documentation for Cell Frequencies Count Imputation Implementation*. DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-1x. April 20, 2006.

SAS Institute (1999). *SAS/STAT® User's Guide*, Version 8, Cary, NC: SAS Institute Inc., p.1297.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, **76**, 103-154.

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

Thibaudeau, Y., Chen, I., and Sands, R.D. (2005). Measuring the Discriminatory Power and Bias of Imputation Methods Designed for Imputing Status and Occupancy Status, *Proceedings for the Section on Survey Research Methods*, American Statistical Association.

Williams, T. (2005). *2010 Census Imputation Research – Methodology for Developing the Truth Deck*, DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2-03, August 25, 2005.