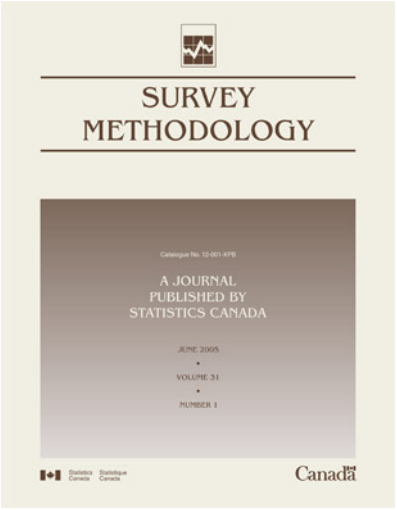
Catalogue no. 12-001-XIE

Survey Methodology

June 2006



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2006

Published by authority of the Minister responsible for Statistics Canada

July 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 32, No. 1, pp. 65-76
Statistics Canada, Catalogue No. 12-001

A Model for Estimating and Imputing Nonrespondent Census Households under Sampling for Nonresponse Follow-up

Elaine L. Zanutto and Alan M. Zaslavsky [1]

Abstract

Sampling for nonresponse follow-up (NRFU) was an innovation for U.S. Decennial Census methodology considered for the year 2000. Sampling for NRFU involves sending field enumerators to only a sample of the housing units that did not respond to the initial mailed questionnaire, thereby reducing costs but creating a major small-area estimation problem. We propose a model to impute the characteristics of the housing units that did not respond to the mailed questionnaire, to benefit from the large cost savings of NRFU sampling while still attaining acceptable levels of accuracy for small areas. Our strategy is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at larger levels of geography. To do this, households are first classified into a small number of types. A hierarchical loglinear model then estimates the distribution of household types among the nonsample nonrespondent households in each block. This distribution depends on the characteristics of mailback respondents in the same block and sampled nonrespondents in nearby blocks. Nonsample nonrespondent households can then be imputed according to this estimated household type distribution. We evaluate the performance of our loglinear model through simulation. Results show that, when compared to estimates from alternative models, our loglinear model produces estimates with much smaller MSE in many cases and estimates with approximately the same size MSE in most other cases. Although sampling for NRFU was not used in the 2000 census, our estimation and imputation strategy can be used in any census or survey using sampling for NRFU where units are clustered such that the characteristics of nonrespondents are related to the characteristics of respondents in the same area and also related to the characteristics of sampled nonrespondents in nearby areas.

Key Words: Missing data; Small area estimation; Iterative proportional fitting; Log-linear models; ECM.

1. Introduction

Sampling for nonresponse follow-up (NRFU) was an innovation for U.S. Decennial Census methodology considered for the year 2000 (U.S. Bureau of the Census 1997a, b). Under current procedures used in 99% of households, the Census Bureau first mails or personally delivers a questionnaire, to be returned by mail. Then field enumerators attempt to contact all mail nonrespondents (about 35% of those mailed). The workload of about 42 million households makes this one of the most expensive census operations.

Sampling for NRFU involves sending field enumerators to only a sample of the nonresponding housing units. This sample is either an unclustered element sample of nonresponding housing units (the "unit sample") or a cluster sample consisting of all nonresponding units in a sample of the census blocks (small areas approximating a city block or some compact rural area, averaging about 15 housing units). This second stage of followup leads to the completion of a questionnaire (through proxy response or imputation, if necessary) for all sample housing units, except those that are resolved to be vacant.

The potential cost savings of sampling are large, but it would require estimating the characteristics of a huge number of nonsampled nonresponding households, posing a major small-area estimation problem (Ghosh and Rao 1994; Rao 2003). We show that using appropriate models to impute the characteristics of the nonsample nonrespondent households, we may benefit from the large cost savings of NRFU sampling while still attaining acceptable levels of accuracy for small areas. Our strategy is to model household characteristics using low-dimensional covariates at detailed levels of geography and more detailed covariates at larger levels of geography. To do this, households are first classified into a small number of types. A hierarchical loglinear model then estimates the distribution of household types among the nonsample nonrespondent households in each block. This distribution depends on the characteristics of mailback respondents in the same block and sampled nonrespondents in nearby blocks. Nonsample nonrespondent households can then be imputed according to this estimated household type distribution.

Although, for complex legal reasons, sampling for NRFU was not used in the 2000 census, our estimation and imputation strategy can be used for small area estimation or imputation in any census or survey using sampling for NRFU where units are clustered such that the characteristics of nonrespondents are related to the characteristics of

1. Elaine L. Zanutto, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, U.S.A. E-mail: zanutto@wharton.upenn.edu; Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A. E-mail: zaslavsky@hcp.med.harvard.edu.

respondents in the same area and also related to the characteristics of sampled nonrespondents in nearby areas. The related methodologies of Purcell and Kish (1980) and Zhang and Chambers (2004) also use loglinear models to estimate small-area cross-classified counts assuming that the total populations are known and that auxiliary cross-classified data is available at the small area level. We have an additional source of information, specifically the characteristics of the nonrespondents in the NRFU sample. This allows us to model the relationship between respondents and nonrespondents directly in some blocks.

Section 2 summarizes proposed strategies for imputing missing data in this situation. Section 3.1 describes our general sampling and estimation procedure. We present our estimation and imputation model in Section 3.2, our smoothing and estimation procedures in Section 3.3, and evaluate our model by simulation in Section 4. Methods for MSE estimation are summarized in Section 5, and Section 6 presents conclusions.

2. Previous Proposals for Imputing Census Nonrespondents

Several methods have been proposed for imputing the characteristics of nonresponding housing units. "Top-down" strategies first estimate counts for aggregates of households and then allocate them to small areas in a manner that maintains consistency with the aggregates. Simple ratio models (Fuller, Isaki and Tsay 1994, henceforth "FIT"), Poisson regression models (Bell and Otto 1994), or more complex loglinear models (as we propose here and in Zanutto and Zaslavsky 1995b, a) are used to estimate counts for small areas and detailed demographic groups for which direct estimates are not possible. Like us, FIT classify households into a modest number of types defined by important characteristics (*e.g.*, number of people, race, tenure) and then estimate the number of households of each type among nonsample nonrespondents. A complete census roster is then generated by imputing the estimated number of households of each type. The main difference between our approach and that of FIT is that by using a loglinear model rather than a stratified ratio model, we obtain more flexibility in the detail of constraints imposed at various levels of geography. Bell and Otto (1994) estimate the number of people over 18 years old of each race (Hispanic, non-Hispanic Black, Other) in each nonsample nonrespondent housing unit but do not consider how to group imputed persons into households or how to impute household-level characteristics such as tenure. These *ad hoc* "top-down" models incorporate at most a few household characteristics and hence do not explicitly model household structure, but they are designed to maintain the consistency of the aggregates that are considered most important.

Schafer (1995) develops a "bottom-up" strategy in which households are built up from individual persons and their characteristics and relationships, each of which must be described by its own model. These models describe the population in more detail and can support full probability (*e.g.*, Bayesian) inferences about unobserved characteristics. However, this approach, unlike the other, requires that a fairly complex set of models be built before any imputations can be made. Furthermore, in this framework it is more difficult to maintain consistency between microdata and aggregate controls. A combined strategy, however, could use our models to produce nearly unbiased estimates by household types and Schafer's models to complete the imputations.

3. Estimation Procedures and Models

3.1 Overview

In the first step of the imputation procedure, counts of the number of nonsample nonrespondent households of each type are predicted using a combination of logistic and loglinear models for each block. This step is the topic of this paper (and of FIT).

For modeling we classified households into types based on a few important characteristics. Here we use 19 types, one of which is "vacant." The remaining 18 are defined by the cross-classification of households by three size categories (1–2 people, 3–4 people, 5 or more people), three race categories (Hispanic, non-Hispanic Black, Other), and two tenure categories (owner, renter).

To predict the number of vacant housing units among nonsample nonresponding units in each block we (and FIT) fitted a logistic regression model, recognizing that the relationship between respondent and nonrespondent households is different for vacant than for nonvacant housing units. Respondent vacants are simply those that were identified as vacant by a postal service letter carrier, leading to mail return of the original questionnaire. Their distribution is likely to depend largely on housing characteristics related to postal delivery, telling us little about the distribution of nonrespondent vacants.

After modeling vacancies, we fitted a loglinear model to predict the distribution of the nonvacant household types in the remaining nonsample nonrespondent households at three geographical levels. The block is the smallest unit and the one for which estimated counts are calculated. The "estimation domain" is the largest unit and is the area in which estimation is conducted independent of other such domains; in our application to the 1990 census, this is the area for which the census was administered from one of 449 district field offices (DO) representing about 200,000

households on average. Finally, we call an intermediate level of geography an "area", comprising a relatively homogeneous collection of contiguous blocks within an estimation domain. In standard Census Bureau geography these might be census tracts, block groups, or Address Register Areas.

We lay out briefly the remaining steps that would be followed to obtain census products using the estimates. In the second step of the imputation procedure the predicted counts would be rounded to integers. Unbiased schemes (*i.e.*, stochastic procedures that in expectation impute the predicted number of units in each cell) for "controlled rounding" (*i.e.*, rounding in a two-way table while preserving marginal totals) were developed by Cox (1987) and George and Penny (1987). However, more research is needed to determine if these methods can be modified to round households counts while preserving all the margins corresponding to effects in the loglinear model. This is an active research topic due to its importance to statistical nondisclosure.

Finally, detailed person and household information would be imputed for nonrespondent households by substituting donor households with similar characteristics. Donors can be chosen from the sampled nonrespondents, the respondents, or a combination of both sources. Finally, tabulations and microdata samples would be prepared from the completed rosters.

3.2 Loglinear Model

We fitted a loglinear model to estimate the prevalence of the various types of households among nonsample nonrespondent households in a DO, using data from the respondents and from the nonrespondents in the NRFU sample for that DO. The model predicts household types for nonsample nonrespondent households in each block by using information about the characteristics of respondent households in the same block and the characteristics of nonrespondent households, measured by the NRFU sample, in surrounding blocks. To accomplish this, the loglinear model contains interactions among the household characteristics that define household type and response status at various levels of geography.

This modeling strategy is motivated by the fact that when a hierarchical loglinear model (*i.e.*, one in which for every included interaction effect, all main effects or interactions marginal to it are also included) is fitted by maximum likelihood, the fitted values for every margin or mean corresponding to an effect in the model are equal to the corresponding observed margins or means (Birch 1963). Therefore, predictions for household types agree with observed rates for the characteristics included in the model, at the levels of geography and response status corresponding to the interactions included in the model. Also, because model predictions for the included effects are constrained to agree with observed rates based on a probability sample (the NRFU sample), the corresponding estimates are consistent and approximately unbiased. (Exact unbiasedness is not obtained because of the nonlinearity of the prediction model and because the number of nonsample nonrespondent households in a block might be associated with some characteristics of the nonresponding households in the block.)

The loglinear model includes nested geographical factors for blocks and areas. It also includes crossed factors representing the demographic characteristics of households: first-stage response indicator (respondent or nonrespondent household), household type index, and model expressions in the variables that define household types. These model expressions are submodels of the fully interacted model which defines household type (*i.e.*, race × size × tenure).

We use the following notation:

i = block index ($i = 1, \ldots,$ number of blocks in the DO),

j = index of household type ($j = 1, \ldots,$ number of types),

r = first-stage (mail) response indicator, $r = 0$ for nonresponding households and $r = 1$ for respondents,

$a = a(i)$ = index for the area containing block i ($a = 1, \ldots,$ number of areas),

$x_k = x_k(j)$, $k = 1, 2, 3, 4$ = model expressions in the variables that define household types where x_1 represents the full cross-classification defining household types, x_2 and x_3 are model expressions which are marginal to x_1, and x_4 is a model expression which is marginal to x_3. (This terminology is explained below.)

We assume a loglinear model of the following form:

$$n_{ijr} \sim \text{Poisson}(m_{ijr}),\ \log(m_{ijr}) = z_{ijr}^T \beta \qquad (1)$$

where n_{ijr} and m_{ijr} are respectively the observed and expected counts for block i, household type j and response status r, and Z is the design matrix corresponding to the following model formula:

$$x_1 + i * x_2 + i * r + r * x_3 + r * a * x_4. \qquad (2)$$

In the standard generalized linear models notation of Wilkinson and Rogers (1973), the "*" operator indicates that the main effects and all interactions that are marginal to the given interaction are included in the model, so that this model contains main effects for model expression x_1, response indicator r, and block indices i and the interactions $i * x_2$, $i * r$, $r * x_3$, and $r * a * x_4$.

Because, in (1), x_4 interacts with area, the smallest level of aggregation for the non-respondent data, it should represent a fairly coarse classification of households including only those household characteristics that are most important to impute accurately at the area level. The x_3 expression may include terms not included in x_4, since it is fitted at a higher level of geography where there is more data available. Similarly, the x_1 expression might include the most interactions, including the interaction of all variables that define household type, since it is fitted at the largest level of geography, using all available data. Finally, x_2, which can be different than x_3 since it interacts with i instead of r, should be less detailed than x_1 since it interacts with block, a much smaller level of geography. These guidelines are motivated by the fact that estimates of interactions with i, r, or a are determined by relatively few observations and should be kept simple. Choosing x_2, x_3, and x_4 as described above should improve the precision of model estimates while preserving the most important margins.

As an example of possible $x_1, \ldots, x_4$ terms, suppose that we define household type by a race × size × tenure cross-classification. Then one possible specification of x_1, x_2 and x_3 is $x_1 =$ race $*$ size $*$ tenure, $x_2 =$ race $*$ size + tenure, $x_3 =$ size $*$ tenure, and $x_4 =$ race + size + tenure. Allowing the $x_1, \ldots, x_4$ terms to be model expressions, rather than just simple interactions, gives us a concise way to represent a model containing all the desired interactions. For example, a model containing an $i * x_2$ term, where x_2 is specified above, includes both a block × race × size interaction and a block × tenure interaction.

A heuristic interpretation of our loglinear model is that we estimate the detailed distribution of household types across the whole area (x_1) and then shift that distribution to allow for the general characteristics of the block (x_2), the general differences between responding and nonresponding households (x_3), and the most important differences between responding and nonresponding households in the particular area (x_4). All interactions could be included except those of the form $r * i * x$, where x represents a model expression in the variables that define household type (*i.e.*, such as x_1, x_2, x_3, or x_4). Interactions of this form depend on the margins determined only by non-respondent households in a single block and these are unavailable in nonsample blocks under the block sample design, and based on a very small sample under the unit sampling design. Therefore our model specification excludes all $r * i * x$ effects, which are always inestimable (or poorly estimated, in the household sample design). This model generalizes two simple theories which are contained as submodels. First, if there are no differences between blocks (*i.e.*, the loglinear $i * x_2$ and $a * x_4$ interactions are zero) then nonrespondent households in each block are imputed according to the overall proportion of nonrespondent households in each of the x_3 categories in the NRFU sample, through the $r * x_3$ effect. In other words, the imputations are made using the same proportions in each block. Second, if there are no differences between respondents and nonrespondents (*i.e.*, no $r * x_3$ or $r * x_4$ interactions) then nonrespondents are imputed in the same proportions as observed in the respondents in each block.

Our general model formulation can accommodate many definitions of area and household type and choices of model expressions. Areas should be defined to be large enough to contain adequate data to estimate the corresponding interactions, but also relatively homogeneous. For example, areas could be defined by a combination of geographical contiguity and stratification by block-level covariates (such as percent minority), in order to obtain more homogeneous areas whose differences could be described by modeling. Generalization to more than two levels of geography within the estimation domain is also straightforward. Thus, for example, we could interact another model expression x_5 with a geographical unit intermediate between the area and the block.

Fitting the model by maximum likelihood, the following quantities are made equal to the corresponding observed values: (1) fitted block counts (through the main effect for block, i), (2) response rates by block (through the $r * i$ term), (3) household characteristic means overall (for x_1 characteristics through the main effect term for x_1) and (4) by block (for x_2 characteristics through the $i * x_2$ term), and (5) household characteristic means for nonrespondents overall (for x_3 characteristics, through the $r * x_3$ term) and (6) for nonrespondents by area (for x_4 characteristics, through the $r * a * x_4$ term). Thus, this model generalizes the model used by FIT of block × type independence, yielding unbiasedness at smaller levels of aggregation, assuming that the margins and averages are estimated unbiasedly from the data. The estimate for area is not exactly the same as the usual unbiased estimate obtained by direct estimation from the NRFU sample because the model makes observed and fitted margins agree for the households in sample. In effect, there is covariance (regression) adjustment that shifts the aggregate to account for observed differences between respondent households in sample blocks and respondent households in nonsample blocks, or in the unit sampling design, between respondent households in blocks with households in the NRFU sample and blocks without households in the NRFU sample.

The idea of modeling household characteristics using low-dimensional covariates at the block level and in more detail at more aggregated levels is similar in concept, although not in details, to the model described in Zaslavsky

(2004). For use of loglinear weights to match sample estimates of aggregates, see Brackstone and Rao (1976), Oh and Scheuren (1983), and Zaslavsky (1988).

3.3 Estimation and Smoothing

We fit the model by maximum likelihood estimation under the Poisson sampling model, which is equivalent to fitting a multinomial logistic regression model. The fitting is complicated by the fact that the data do not form a complete block × response × type table because we have counts by block, but not characteristics for nonsample nonresponding households. In the block sampling design we lack characteristics of all nonrespondents in some blocks and in the unit sampling design we lack characteristics of some nonrespondents in almost all blocks. To fit the model we use a modified iterative proportional fitting (IPF) algorithm adapted to data that are partially classified in a part of the dataset (Appendix).

With some data sets, some parameters may be inestimable because the maximum likelihood estimates lie on the boundary of the parameter space (infinite on the loglinear scale, indicating a zero on the count scale) or because there is no information for the parameter. Tailoring the model specification in each estimation domain to remove inestimable parameters is impractical in a census production setting.

By introducing a small amount of prior information, estimability of all parameters can be guaranteed. To do this, we append a small amount of "pseudo-data" to the data for each area, whose proportions by type are equal to those for some surrounding area (the DO, in our simulations), by adding these counts to the data table before fitting the model. This implements an empirical Bayes analysis for multinomial data with distribution $f(n_1, ..., n_H \mid p_1, ..., p_H) \propto \prod_{i=1}^{H} p_i^{n_i}$, where $n_1, ..., n_H$ are the observed number of households of each type in a block or area. If $\{p_i\}$ have a joint Dirichlet prior distribution, $f(p_1, ..., p_H) \propto \prod_{i=1}^{H} p_i^{\alpha_i - 1}$, $\alpha_i \geq 0$, the resulting posterior distribution for the p_i's is Dirichlet with parameters $\alpha_i + x_i$ (Gelman, Carlin, Stern and Rubin 1995, page 76) and posterior mode proportional to the parameters. Thus, this empirical Bayes procedure is equivalent to adding $\sum \alpha_i$ households to the area, where α_i of these households are of the i^{th} type. We fix the α_i's to be proportional to the observed proportions of each household type in some surrounding area, so the observed distribution of household types is smoothed by mixing it with the distribution for a surrounding area, thus avoiding introducing bias at the level of the larger area. This prior specification induces a prior on the parameters of the loglinear model. See Rubin and Schenker (1987), Zaslavsky (1988), and example and historical references in Clogg, Rubin, Schenker, Schultz and Weidman (1991) for similar use of smoothing.

After estimating the model parameters, the next step is to calculate predicted counts for each household type for the nonrespondent households that are not in the NRFU sample. Using the IPF algorithm, the predictions for the nonsample nonrespondent households are obtained automatically by applying the same fitting proportions to the partially observed part of the table as to the fully observed part of the table, so no further calculation is required (Appendix).

4. Simulations

4.1 Overview

Our simulation study evaluated the bias, variance and MSE of the estimates of estimated demographic aggregates (such as the number of households by race, size and tenure) at various levels of geography, using estimated household compositions for non-respondent households that are not in the NRFU sample. Analytic evaluations are infeasible, given the complexity of the models and sampling scheme, the dependence of the performance of the model on the actual geographical distribution of household types, and the number of variations of the model that could be examined.

We used block-level data from three DOs from the 1990 U.S. Decennial Census; these constituted our estimation areas. The simulations are similar in structure to those described by Schindler (1993) or FIT.

The steps of the simulation are as follows:

1. Blocks or nonrespondent housing units are sampled according to the NRFU sampling scheme.
2. A logistic regression model for vacant households is fitted to the respondent households and the sampled nonrespondent households.
3. The predicted number of nonrespondent households that are vacant is calculated for each block.
4. A model for nonvacant types is fitted using the respondent households and the sampled nonrespondent households.
5. The predicted number of nonsample nonrespondent households of each nonvacant type are calculated for each block.
6. Aggregates of interest are calculated based on the predicted counts, and compared to the truth using loss functions.

In our simulations, repeating these steps 30 times yielded estimates of RMSE (defined in section 4.3) with adequate accuracy to evaluate the performance of our model relative to the alternative models. Specifically, the estimated coefficients of variation of the estimated differences in RMSE for the stratified ratio method (described below) and

loglinear method are less than 0.05, except when the difference between estimated RMSEs is very small, resulting in a large coefficient of variation.

The performance of our proposed model is compared with two alternative estimation methods, under both the unit and block sampling designs. Each method first fits a logistic regression model to estimate the number of nonrespondent households that are vacant in each block. The first alternative, the "unstratified ratio method", imputes households for nonsample nonrespondent households in each block in proportion to the distribution of household types among nonrespondent households in the follow-up sample for the entire DO. The second alternative, the "stratified ratio method", is a version of that in FIT. We first form strata of approximately 82 blocks based on the racial composition of the blocks, as described by FIT. (We use both respondent and nonrespondent data to form strata, assuming, as in FIT, that similar information would be available from administrative records. Stratification based only on respondent information yielded similar results.) Then, in each stratum, nonsample nonrespondent households are imputed to nonvacant types in proportion to the frequency of the type in the follow-up sample for that stratum.

We simulate each estimation method using a NRFU sampling rate of 30%. In each stratum, we simulate NRFU sampling by selecting a 30% simple random sample of blocks for the block sampling design, and a 30% simple random sample of nonrespondent households in each stratum for the unit sampling design. The characteristics of the nonrespondent households in these samples is assumed to be known (*i.e.*, as a result of follow-up operations). For both our loglinear model method and the stratified ratio method, we select a 30% sample of blocks or nonrespondent households using simple random sampling without replacement from each area.

We considered several loglinear model formulations. The best model for both the block and unit sampling designs, by the criteria described in Section 4.3, uses $x_1 = \text{size} * \text{race} * \text{tenure}$, $x_2 = \text{race} * \text{tenure} + \text{size}$, $x_3 = \text{race} * \text{size}$, $x_4 = \text{tenure}$. This model is used in the simulations.

To ensure the model can be fitted in every case and to speed the convergence of the IPF, we smooth the data by adding one hypothetical respondent household ("pseudo-data") to each block. This household is divided among the 18 nonvacant household types according to the overall DO proportions of respondent households. Estimates using 5 households for smoothing were about as accurate as with one, and more aggressive smoothing (adding 10, 15, 20 or 25 households per block) slightly increases errors in the estimates. Also, although adding only a small fraction of a household to each block is sufficient to ensure that the model can be fitted in every case, using less than 1 household per block drastically slowed convergence and slightly increased the error in the estimates.

The three estimation procedures used the same logistic regression model for vacancies. The covariates for each block are the mail nonresponse rate, the percentages of respondent households that are (separately) renters, apartment dwellers, and of a minority race (either Black or Hispanic), the average value of owner-occupied homes, the average monthly rent for rental units, indicator variables for each of the areas, and interactions between percentage of respondent renters and average monthly rent, percentage of respondent renters and average monthly rent squared (mean-centered), percentage of respondent owners and average home values, and percentage of respondent owners and average home values squared (mean-centered). To avoid computational problems arising from blocks with no nonrespondent vacant households, one hypothetical non-respondent household is added to each block divided between vacant and nonvacant according to their proportions in the sampled nonrespondent households in the DO.

4.2 Data

We use short-form data from the 1990 census for three DOs, whose characteristics are described in Table 1. The race of a household is determined by the most prevalent race in the household, usually (98% of households) the only race. In DO 1 we grouped consecutive (and therefore contiguous) block groups (clusters of contiguous blocks) into 94 areas containing an average of 52 blocks and 1100 households. For DOs 2 and 3, block group information was unavailable so we formed areas by grouping consecutive blocks into clusters containing an average of 50 blocks (on average, 548 households per area in DO 2 and 918 households per area in DO 3).

Table 1
Characteristics of the Census District Office Areas Used in the Simulations

	DO1	DO2	DO3
Household	112,966	169,321	149,567
Blocks	4,907	15,470	8,167
Pseudo-areas	94	309	163
Non-Hispanic Black	14.4%	28.5%	1.3%
Hispanic	6.1%	1.0%	6.6%
Other	73.5%	59.4%	81.5%
Owner	63.8%	59.5%	52.6%
Renter	30.2%	29.4%	36.7%
Vacant	6.0%	11.1%	10.7%
Size 1 (1–2 people)	50.4%	46.9%	55.2%
Size 2 (3–4 people)	31.6%	31.6%	26.2%
Size 3 (5+ people)	12.0%	10.4%	7.9%
Response Rate	72.6%	65.3%	56.7%

4.3 Measures of Bias, Variance, and Mean Squared Error

Loss functions for our evaluations are based on the relative error for household category j (a type or combination of types) in geographic area i (a block or collection of blocks):

$$d_{ijs} = \frac{\hat{Y}_{ijs} - Y_{ij}}{Y_{i+}} \quad (3)$$

where Y_{ij} is the true number of households of category j in geographical unit i, $\hat{Y}_{ijs}$ is the corresponding number of households estimated from sample s (including those observed in the sample and estimated by the model), and Y_{i+} is the total number of households in geographical unit i.

We summarize bias in estimated counts for category j and a level of geography (block, area, DO) with Root Mean Weighted Squared Bias (RMWSB):

$$\hat{\mathrm{R}}\mathrm{MWSB}_j^2 = \frac{\sum_i Y_{i+} \left\{ \left(\frac{1}{S}\sum_s d_{ijs} \right)^2 - \frac{1}{S(S-1)} \left(\sum_s d_{ijs}^2 - \frac{1}{S}\left(\sum_s d_{ijs}\right)^2 \right) \right\}}{\sum_i Y_{i+}} \quad (4)$$

where S is the number of samples drawn and $i = 1, \ldots, I$ where I is the number of geographical units. The second term in the numerator removes a bias due to the finiteness of the simulation. From a design-based perspective, we regard the composition of each area as a fixed quantity, and only sampling is random. Then bias is defined as the average difference, over all possible samples, between the truth for an area and the corresponding estimates, essentially the model error for that area. Such error is inevitable since the composition of the nonrespondents in any block is not entirely predictable. A more serious type of bias would involve systematic error in estimates for a collection of blocks with similar composition. Although we have not checked for all possible types of bias in this sense, the model specification protects us against bias at higher levels of aggregation because model estimates are constrained to agree (approximately) with unbiased estimates for areas and DOs.

As a measure of overall error, we calculate the Root Mean Weighted Mean Squared Error (RMWMSE) for each household category j, which is given by

$$\hat{\mathrm{R}}\mathrm{MWMSE}_j^2 = \frac{\sum_i Y_{i+} \left(\frac{1}{S}\sum_s d_{ijs}^2 \right)}{\sum_i Y_{i+}} \quad (5)$$

where Y_{ij}, $\hat{Y}_{ijs}$, Y_{i+}, i, and S are defined as above. (The two "means" refer to mean over geographical units (i) and over samples (s).) We obtain a measure of the standard deviation of the estimates for household category j by calculating the Root Mean Weighted Variance (RMWV):

$$\hat{\mathrm{R}}\mathrm{MWV}_j^2 = \frac{\sum_i Y_{i+} \left\{ \frac{1}{S-1} \left(\sum_s d_{ijs}^2 - \frac{1}{S}\left(\sum_s d_{ijs}\right)^2 \right) \right\}}{\sum_i Y_{i+}} = \hat{\mathrm{R}}\mathrm{MWMSE}_j^2 - \hat{\mathrm{R}}\mathrm{MWSB}_j^2 . \quad (6)$$

Note that these MSE, bias, and standard deviation measures are all estimates of expectations with respect to repeated NRFU sampling from the given finite population of blocks. These loss functions can be applied at various levels of geography, reflecting the fact that the main use of block level estimates is aggregation to form estimates at higher levels of geography. With this in mind, these measures were also chosen because they weight errors by the size of the geographical unit. This leads to consistent estimates of error when aggregating over geographical units, which is appropriate due to the arbitrariness of unit boundaries (Zaslavsky 1993). We base our measures on errors relative to the total area i population rather than the population in the target category only, because the latter denominator inflates the importance of small errors in blocks where the category rarely or never appears.

4.4 Results

For simulated NRFU sampling using both the block and unit sampling designs, estimates of the number of households with each characteristic are calculated at block, area, and DO levels of geography using each of the three estimation methods. The results for each method are represented by the shaded bars in Figure 1 for the unit sampling design. (Results for the block sampling design are not shown here, but the pattern of results are similar with the RMWMSE being about 10% greater for all estimates.) In this figure, each row of bar charts displays the RMWMSE for block, area, and DO level estimates for one of the three DOs. Each group of three bars represents the RMWMSE for estimates of the total number of households for each of the tenure categories, the household size categories and the race categories using each of the three methods. Because all three methods use the same logistic regression model to predict the number of vacant nonsample nonrespondents in each block, the vacant category is omitted from the plots.

RMWMSE with both the stratified ratio method and the loglinear model was much smaller than with the unstratified ratio method for most household characteristics at the block and area level. Therefore, we confine further discussion to comparison of the two former methods.

The most dramatic differences appear for the tenure categories at the block and area levels. In each DO, block and area level estimates of the tenure categories from the loglinear model have much smaller RMWMSE than the estimates from the stratified ratio method, primarily because the former had much smaller bias (RMWSB). Standard deviations (RMWV) were slightly larger for the loglinear model under the unit sampling design, but about equal for the two methods under the block sampling design. The loglinear model had smaller bias for the tenure categories at the area level because tenure is included in the model as an area-level effect, x_4. Stratification on race in the ratio method reduces RMWMSE for the race categories at the block level, but the two methods have comparable RMWMSE for the race categories at the area and DO levels. The stratified ratio method loses its advantage over the loglinear model at the area level because the former does not use any area-level information. Both methods generally produce estimates with comparable RMWMSE at all levels of geography for the size categories.

The statistical significance (under the simulations) of differences in RMWMSE between the methods was evaluated using t–tests. Almost all differences at the block and area levels, excluding the vacant category, have two-tailed $p-$values ≤ 0.001 and therefore cannot be attributed to simulation error.

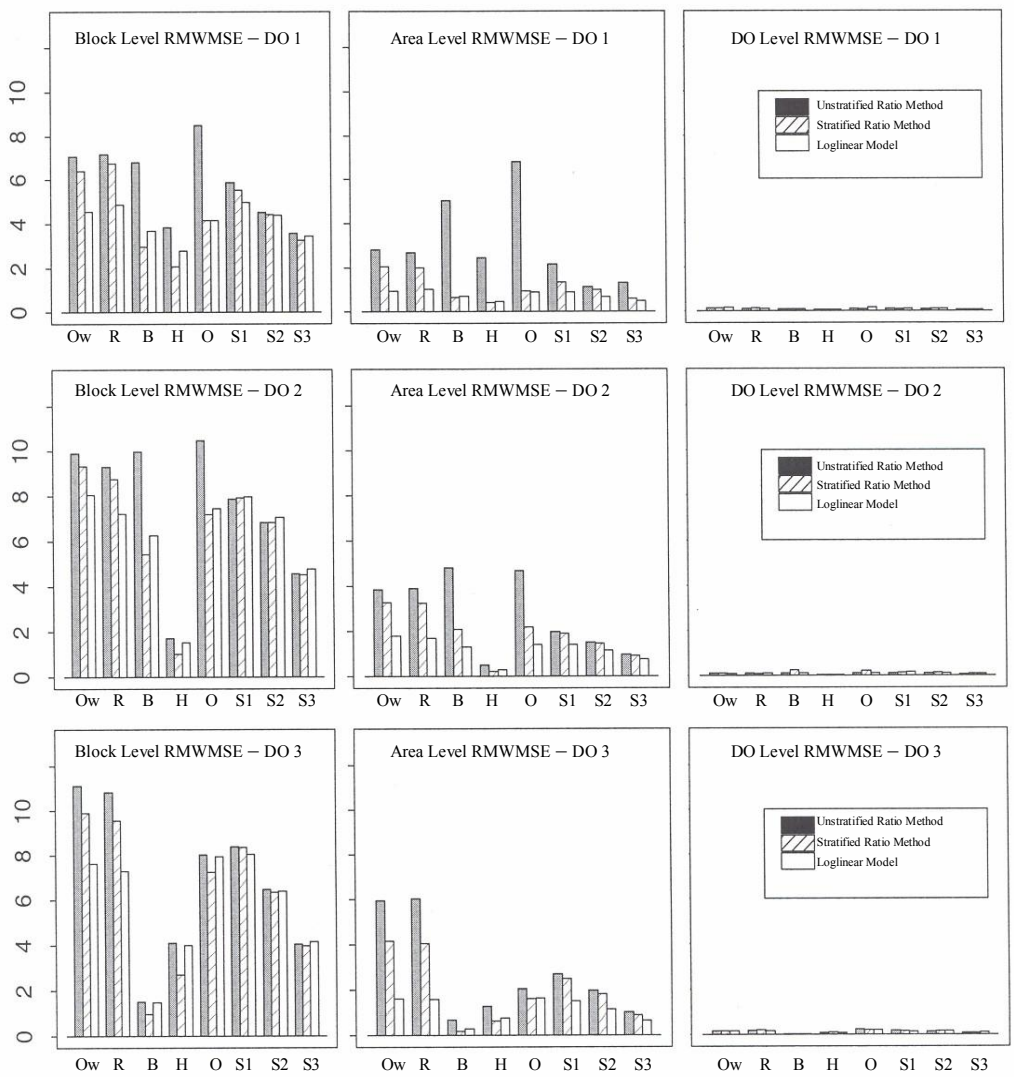


Figure 1. RMWMSE for block, area, and DO level estimates for each household characteristic, using the unit sampling design for DO 1, 2, and 3, with 30 simulated samples ("Ow" = Owner, "R" = Renter, "B" = Black, "H" = Hispanic, "O" = Other race, "S1" = Size group 1 (1 – 2 people), "S2" = Size group 2 (3 – 4 people), "S3" = Size group 3 (5 or more people)).

5. Assessment and Prediction of Model Error

Methods for estimation of MSE of fitted estimates using sample data are briefly summarized here due to space limitation; methods and findings are available from the first author.

First, we developed analytic approximations that predict the effect of changing the sampling rate on the accuracy of our estimates without requiring additional simulations at each rate. These can be useful for sample design. We approximate the RMWMSE of block, area, and DO level estimates at a new sampling rate under both the block and unit sampling designs, assuming simulation results using one sampling rate are already available, by combining estimates of bias and variance at the current sampling rate using two rescaling factors. The first factor reflects the changed proportion of housing units that require estimation under the new sampling rate, which affects the bias and variance of the combined estimates. The other reflects the effect of the sampling rate on the variance of the estimates for the nonresponding units. Simulations demonstrated the accuracy of predictions for RMWMSE using these approximations, except for some extreme extrapolations.

Using these results, we developed a cross-validation procedure to facilitate within-sample estimates of RMWMSE for use in a production setting where the true characteristics of the nonsample nonrespondent households are not known. The follow-up sample in each areas is divided randomly into C cross-validation groups (of blocks for block sampling, and of households for unit sampling). Each cross-validation group is dropped out in turn and the model is fitted to the nonrespondents in the remaining $C-1$ cross-validation groups and the respondents in all C groups. We can then estimate RMWMSE under the design simulated by the cross-validation and project this estimate to the actual sampling rate, or some other rate of interest, using the approximations described in the preceding paragraph. Simulations show that this produces accurate estimates of RMWMSE at block and DO levels of geography, with some overestimation at the area level. This method also provides separate estimates of bias and variance that are shown by simulation to be very accurate. These are useful for assessing model adequacy since a poorly-fitting model would be betrayed by a large component of MSE due to bias.

6. Conclusions

In the preceding sections, we have presented a model-based approach to imputation of the characteristics of nonresponding households in a census that were not sampled for nonresponse followup. In simulations, our loglinear model produces estimates with much smaller error than two alternatives for some estimands, and is about equivalent for others. These conclusions hold for both the block and unit sampling designs. An advantage of our approach is that models can be specified to constrain only a few marginal tables or interactions of characteristics at the finest levels of geography, where the data are sparse, while fitting more detailed distributions of characteristics at higher levels of geographic aggregation at which more data are available. This is consistent with typical practice in release of census data, which include minimal characteristics at the block level but increasingly more detailed characteristics for larger units.

Many important uses of the census involve estimation of the population and its characteristics for small domains such as legislative districts and planning areas for social services (such as schools and clinics) and commercial development. Even though these domains will not always align with the areas used in census estimation, controlling the census estimates to match unbiased estimates at several levels of geography makes it more likely that estimates for policy-relevant domains assembled from wholes or parts of these areas will also be nearly unbiased. Our method has more predictable aggregate properties than complex alternatives such as hierarchical spatial modeling. Although the latter might produce estimates with smaller MSE at the lowest levels of geography, fitting such models and checking their biases at various levels of geographic aggregation would require extensive local tuning which is likely to be impractical in a census production setting.

Our methodology is illustrated here in the context of a NRFU sampling for the U.S. Decennial Census, but our estimation and imputation strategy can be used for small area estimation or imputation in any census or survey using sampling for nonresponse followup with hierarchically structured populations. We can also incorporate administrative records as covariates for predicting the characteristics of the corresponding nonrespondent households (Zanutto and Zaslavsky 2002). In that scenario, data from households in the NRFU sample for which we have both census and administrative records information are used to estimate the systematic differences between the two information sources. Under the same models, we impute the characteristics of nonsample nonrespondent households. Using administrative records through this modeling approach can improve the accuracy of small area (block-level) estimates.

Although the discussion of sampling in the United States census has been politically contentious, nonetheless in the long run it seems likely that some form of estimation will be used for nonrespondents. The potential might be even greater in countries where population estimation already makes substantial use of administrative records (Redfern 1989). Methods such as those described here that can

combine information across data sources while reflecting local diversity will be essential to such efforts.

Appendix

Iterative Proportional Fitting with Partially Cross-Classified Data

A standard approach to fitting loglinear models to partially cross-classified data uses an EM algorithm (Dempster, Laird and Rubin 1977; Little and Rubin 2002, chapter 8), in which in alternate steps (1) the expected counts are imputed under the model and (2) the model is refitted to the observed and imputed data, using iterative proportional fitting (IPF) (Darroch and Ratcliff 1972) for models without closed-form solutions. In the more efficient ECM modification of this algorithm, only a single cycle of the IPF algorithm is taken at each step (Meng and Rubin 1993).

For our application we developed a modified IPF algorithm that is faster than the EM and ECM algorithms for our models, which always include a block × response interaction and never include any block × type × response interactions. We found that our modified IPF algorithm converges in approximately one half to two thirds the number of cycles that ECM requires with less computation per step (Zanutto 1998, Part 1, Appendix A). (Convergence is declared when the predicted and observed values of the minimal sufficient statistics of the model are sufficiently close.)

Our algorithm takes advantage of the fact that partially classified observations contribute to the likelihood only through the total number of nonrespondent households in each block. Therefore, to maximize this part of the likelihood we need only ensure that the fitted number of nonrespondents in each block equals the observed number, which is automatic because the block × response interaction is always included in our model.

The modified IPF algorithm fits the model to the fully classified observations using an ordinary IPF algorithm, ignoring the partially cross-classified observations. For the block sampling design, this means that the model is fitted using the fully observed part of the block × type × response table using an ordinary IPF algorithm, ignoring the partially classified part of the table. Predictions for the partially cross-classified cells are obtained by applying the same fitting proportions to those cells as to the fully observed part of the table. Finally, predictions for the partially cross-classified cells are scaled so that the fitted number of nonrespondents in each block equals the observed number. For the unit sampling design, the same algorithm is used, viewing the collection of respondent households and nonrespondent households in the follow-up sample as analogous to the fully-observed part of the table in the block sampling design and viewing the blocks with no nonrespondents in the follow-up sample as analogous to the out-of-sample blocks in the block sampling design. This gives predictions for nonrespondent households in blocks with no nonrespondents in the follow-up sample. Predictions for nonrespondent households in blocks with one or more nonrespondent households in the follow-up sample are obtained by applying the predicted distribution of household types among sampled nonrespondent households in each of these blocks to the corresponding nonsample nonrespondent households in these blocks. For more details about in the unit sampling case, see Zanutto and Zaslavsky (2002).

We now illustrate the IPF algorithm for the block sampling design under a Poisson model like (1) with $\log(m_{ijr}) = z_{ijr}^T \beta$ where m_{ijr} represents the expected number of households in block i of household type j of response status r, and Z is the design matrix corresponding to the model expression $i * x + i * r + r * x$. This is a simplified version of the model in (2) with only one level of geography and only one "x" representing the full cross-classification defining household types. We observe n_{ijr} if $r = 1$ or if $r = 0$ and $i \in S$, but only n_{i+0} if $i \notin S$, where S represents the set of blocks selected for the NRFU sample.

The IPF algorithm to fit this model starts with initial estimates $\hat{m}_{ijr}^0 = 1$ for all i, j, r and contains the following three steps in cycle t:

$$\text{Step 1: } \hat{m}_{ijr}^{t+\frac{1}{3}} = \begin{cases} \hat{m}_{ijr}^t \left(\dfrac{n_{i+r}}{\hat{m}_{i+r}^t} \right) & \text{if } i \in S \text{ or if } i \notin S,\ r = 1 \\ \hat{m}_{ijr}^t & \text{if } i \notin S,\ r = 0 \end{cases}$$

$$\text{Step 2: } \hat{m}_{ijr}^{t+\frac{2}{3}} = \begin{cases} \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\dfrac{n_{ij+}}{\hat{m}_{ij+}^{t+\frac{1}{3}}} \right) & \text{if } i \in S \\ \hat{m}_{ijr}^{t+\frac{1}{3}} \left(\dfrac{n_{ij1}}{\hat{m}_{ij1}^{t+\frac{1}{3}}} \right) & \text{if } i \notin S \end{cases}$$

$$\text{Step 3: } \hat{m}_{ij1}^{t+1} = \hat{m}_{ij1}^{t+\frac{2}{3}} \left(\frac{n_{+j1}}{\hat{m}_{+j1}^{t+\frac{2}{3}}} \right)$$

$$\hat{m}_{ij0}^{t+1} = \hat{m}_{ij0}^{t+\frac{2}{3}} \left(\frac{\sum_{i \in S} n_{ij0}}{\sum_{i \in S} \hat{m}_{ij0}^{t+\frac{2}{3}}} \right).$$

The scaling factors in each step are based only on observed counts.

These steps are repeated until the estimates of the minimal sufficient statistics for the model, excluding $\hat{m}_{i+r}$ for $i \notin S$, $r = 0$ (*i.e.*, $\hat{m}_{i+r}$ for $i \in S$ and $i \notin S, r = 1$, $\hat{m}_{ij+}$ for $i \in S$, $\hat{m}_{ij1}$ for $i \notin S$, $\hat{m}_{+j1}$, and $\sum_{i \in S} \hat{m}_{ij0}$) are sufficiently close to their observed values. Denoting the step at which this occurs as t^*, the final step in this algorithm is to set

$$\hat{m}_{ijr}^{t^*+1} = \begin{cases} \hat{m}_{ijr}^{t^*} \left(\dfrac{n_{i+r}}{\hat{m}_{i+r}^{t^*}} \right) & \text{if } i \notin S,\ r = 0 \\ \hat{m}_{ijr}^{t^*} & \text{otherwise,} \end{cases}$$

to ensure that estimated block×response margin $(i * r)$ for $i \notin S,\ r = 0$ equals the observed margin.

This IPF algorithm produces estimates that converge to the maximum likelihood estimates of the model parameters (Zanutto 1998, Part 1, Appendix A). The second case in Step 2 is not needed to maximize the likelihood but is included to obtain predictions for the nonsample nonrespondent cells (*i.e.*, $i \notin S,\ r = 0$).

References

Bell, W.R., and Otto, M.C. (1994). Investigation of a model-based approach to estimation under sampling for nonresponse in the decennial census. Unpublished paper presented at the Joint Statistical Meetings, Toronto.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society*, Series B, Methodological, 25, 220-233.

Brackstone, G.J., and Rao, J.N.K. (1976). Raking ratio estimators. *Survey Methodology*, 2, 63-69.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

Cox, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

Darroch, J.N., and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43, 1470-1480.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-22.

Fuller, W.A., Isaki, C.T. and Tsay, J.H. (1994). Design and estimation for samples of census nonresponse. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 289-305.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall Ltd.

George, J.A., and Penny, R.N. (1987). Initial experience in implementing controlled rounding for confidentiality control. In *Proceedings of the Bureau of the Census Annual Research Conference*, Volume 3. Washington, DC: U.S. Bureau of the Census, 253-262.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-76.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second Edition. New York: John Wiley & Sons, Inc.

Meng, X.-L., and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys* (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 143-184.

Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, 48, 3-18.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Redfern, P. (1989). European experience of using administrative data for censuses of population: The policy issues that must be addressed. *Survey Methodology*, 15, 83-99.

Rubin, D.B., and Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. *Journal of Official Statistics*, 3, 375-387.

Schafer, J.L. (1995). Model-based imputation of census short-form items. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: Bureau of the Census, 267-299.

Schindler, E. (1993). Sampling for the count; sampling for non-mail returns. Unpublished report, U.S. Bureau of the Census.

U.S. Bureau of the Census (1997a). Census 2000 operational plan. Washington, DC.

U.S. Bureau of the Census (1997b). Report to Congress–the plan for Census 2000. Washington, DC.

Wilkinson, G.N. and Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, 22, 392-399.

Zanutto, E. (1998). *Imputation for Unit Nonresponse*: *Modeling Sampled Nonresponse Follow-up*, *Administrative Records*, *and Matched Substitutes*. Ph.D. thesis, Harvard University, Cambridge, Massachusetts.

Zanutto, E., and Zaslavsky, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 608-613.

Zanutto, E., and Zaslavsky, A. M. (1995b). Models for imputing nonsample households with sampled nonresponse followup. In *Proceedings of the Bureau of the Census Annual Research Conference*. Washington, DC: U.S. Bureau of the Census, 673-686.

Zanutto, E., and Zaslavsky, A.M. (2002). Using administrative records to improve small area estimation: An example from the U.S. Decennial Census. *Journal of Official Statistics*, 18, 559-576.

Zaslavsky, A.M. (1988). Representing local area adjustments by reweighting of households. *Survey Methodology*, 14, 265-288.

Zaslavsky, A.M. (1993). Combining census, dual-system, and evaluation study data to estimate population shares. *Journal of the American Statistical Association*, 88, 1092-1105.

Zaslavsky, A.M. (2004). Representing the Census undercount by multiple imputation of households. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Eds. A. Gelman and X.-L. Meng). West Sussex, England: John Wiley & Sons, Inc. 129-140.

Zhang, L.-C., and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society*, Series B, 66, 479-496.