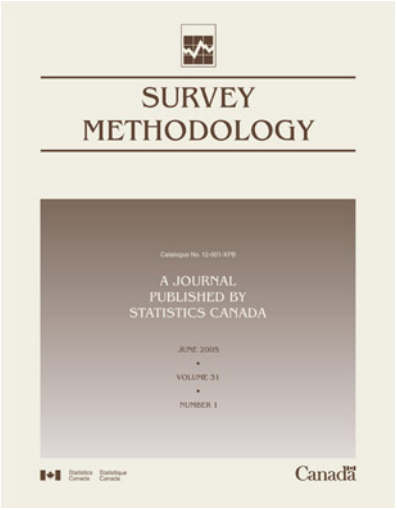
Catalogue no. 12-001-XIE

Survey Methodology

December 2002



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

December 2002

Published by authority of the Minister responsible for Statistics Canada

January 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. 28, No. 2, pp. 135-143
Statistics Canada, Catalogue No. 12-001

Model Explicit Item Imputation for Demographic Categories

Yves Thibaudeau [1]

Abstract

We propose an item imputation method for categorical data based on a MLE derived from a conditional probability model (Besag 1974). We also define a measure for the item non-response error that is useful to evaluate the bias relative to other imputation methods. To compute this measure, we use Bayesian iterative proportional fitting (Gelman and Rubin 1991; Schafer 1997). We implement our imputation method for the 1998 dress rehearsal of Census 2000 in Sacramento, and we use the error measure to compare item imputations between our method and a version of the nearest neighbor hot-deck (Fay 1999; Chen and Shao 1997, 2000) at aggregate levels. Our results suggest that our method gives additional protection against imputation biases caused by heterogeneities between domains of study, relative to the hot-deck.

Key Words: Nearest Neighbor; Conditional probability approach; Bayesian iterative proportional fitting.

1. Introduction

Let S represent a demographic categorical count requested from a census, or needed to compute a survey statistic, and suppose S can be computed from the records of a survey file f, when the records are complete. Also, suppose f is ordered in such a way that proximity in the order of f corresponds to geographical proximity. Consider the situation where f includes records with unreported items. We propose to estimate S with $d(A(f))$, where $A(f)$ is an imputation method that produces a complete survey file, and $d(\cdot)$ estimates S by replacing the unreported items with their values imputed with $A(f)$. $A(f)$ is based on a likelihood that models transitions between two neighbors in f, and associations between the items to be imputed and the relevant domains of study (Cochran 1977, page 34) defined by partitions of the population. $A(f)$ is meant as an advantageous alternative to the popular sequential hot-deck (Kovar and Whitridge 1995), which is a version of the nearest neighbor hot-deck (Fay 1999; Chen and Shao 1997, 2000) that attempts to minimize geographical distance between a unit with unreported items and a suitable imputation donor, while also guaranteeing the distributional homogeneity of the observed and the imputed items with respect to each domain of study. When the domains of a same partition tend not to geographically overlap, borrowing imputation items from a near-by neighbor preserves homogeneity. But, when small domains tend to be dispersed within large domains, the methodologist faces a dilemma. Then, she must choose between hot-deck rules that lead to borrowing the imputed items from geographically close units, leaving the possibility of imputation biases reflecting the local heterogeneity between domains, and domain-specific rules, which guarantee distributional homogeneity by domain, but may not minimize geographical distance. $A(f)$ is an alternative designed to preserve domain integrity, while also simulating the distributional profile of an imputation donor sharing some characteristics with a geographical neighbor. We motivate the design of $A(f)$ with examples and a theoretical description. In this section we review a classification of current hot-deck methods for item imputation with their operating principles, so that we can properly compare them with $A(f)$ in later sections. We also give details on the dress rehearsal of Census 2000 in Sacramento, our test bed throughout the paper.

Fay (1999), and Sande (1981) identify the sequential hot-deck (SHD) as the first category of hot-decks, which we call the "pure" SHD. They add a second category, the fixed-cell hot-deck (FCHD), which we call the pure FCHD. Fay defines a third category of hot-decks: the nearest neighbor hot-deck (NNHD). Chen and Shao (1997, 2000) give an abstract definition of the NNHD in terms of a measure of proximity $|\ |$, based on a covariate x. With the NNHD, a "donor" is any unit such that $|x_r - x_d|$ is minimal, where x_r corresponds to the receiving unit (receiver), and x_d corresponds to the provider of the imputations (donor). By constructing the appropriate measure, and defining a suitable x, we recover both the pure SHD and the pure FCHD as special cases of the NNHD. The pure SHD imputes a receiver item by replacing it with the corresponding item from the closest unit for which it was reported, in the order of f. The pure FCHD relies only on the value of variables that we call the class variables to divide the units between post-strata that are homogenous with respect to the items to be imputed. A donor is chosen at random from the same post-stratum as that of the receiver, irrespective of the order of f.

Fay (1999), and Fay and Town (1998) propose the concept of exchangeability to validate the NNHD. For categorical data two units in f are exchangeable if they are uncorrelated and identically distributed, given the

1. Yves Thibaudeau, Mathematical Statistician, Statistical Research Division, US Census Bureau, 4700 Silver Hill Road, Stop Code 9100, Washington, DC 20233-9100. E-mail: yves.thibaudeau@census.gov.

information available prior to imputing. The operational assumption of the NNHD is that a unit and its nearest neighbor(s) are exchangeable. For the pure SHD it means two contiguous units in f are exchangeable. For the pure FCHD it means that units sharing the same values for their class variables anywhere in f are exchangeable. We define a third instance of the NNHD, which we call the hybrid sequential hot deck (HSHD). To guarantee exchangeability the HSHD requires proximity both in terms of the order of f, and in terms of the class variables.

We use the term "nearest neighbor" in the abstract sense of the NNHD, unless specified otherwise. We use the terms "closest neighbor" to designate the nearest neighbor of the pure SHD, and "closest complete neighbor" to mean the survey unit with no unreported items that is closest in the order of f. In the case of the Sacramento dress rehearsal, the Census Bureau uses a HSHD to estimate householder counts by tenure, race, origin (Hispanic origin), and sex. The householder, usually an adult, is unique for each housing unit, and is determined by the ages, relationships, and order of the persons on the census questionnaire. The HSHD substitutes unreported items with the values of these items corresponding to the last householder who reported them and is in the same post-stratum (Treat 1994). The sorted order of f maintains the proximity of geographical neighbors. The intent behind the HSHD is to define nearest neighbors who are close, both in geography and "in kind". Throughout the paper, we continue to use the term householder, although its meaning may extend to a generic survey unit.

The design of the HSHD is well suited for item imputation in populations geographically clustered by domain. Then the need for class variables is limited. But difficulties arise when the geographical boundaries between the domains begin to blur. Designing a HSHD with good discrimination power in those conditions is an attempt at walking a fine line between specifying enough class variables to account for heterogeneities between domains, and specifying too many, which could yield post-strata so narrowly defined in terms of domain that they don't capture the local geographical character of the receivers. Complicating the situation is the fact that the demographic composition of the population may change as the geography changes, and thus a particular scheme for the HSHD might need to be revised, as the geography changes. In the face of these difficulties $A(f)$ is innovative in the sense that, instead of searching for an ideal nearest neighbor, it generates imputations through a model-based simulation that integrates information relating to the local geography, as well as to domain partitions. $A(f)$ integrates both kind of information by calibrating the parameters of a log-linear model on the basis of the strength of the correlations between the covariates and the variables subject to imputation. Our parameter estimation strategy is the same as that of Zanutto and Zaslavsky (1995a, b). However, because they have access to a representative sample of complete non-respondents, these authors can obtain estimates of the imputation probabilities by implementing a one-step EM algorithm (Dempster, Laird and Rubin 1977). In our situation, we don't assume access to a representative sample, and we implement the full EM algorithm. Implicitly we make an assumption of items "missing at random" (MAR) (Little and Rubin 1987, page 16).

To analyze the results obtained with $A(f)$, and to compare them with those of the HSHD, we derive error measures related to $A(f)$ based on approximations computed using a Bayesian algorithm first introduced by Gelman and Rubin (1991). There are fundamental objections to Bayesian methodologies. Fay (1992) shows that variance estimation based on multiple imputations (Rubin 1996) can lead to inflated estimates of variance, whereas in the same situation the jackknife estimator (Rao and Shao 1992) avoids biases. Meng (1994) suggests that Fay's example stems from a poor communication between an imputer who has specific model information, and an analyst who only has knowledge of the estimation process. In the language of Meng, this situation is uncongenial. While requirements for coordination between imputer and analyst are restrictive, imputation based on exchangeability also has dangerous pitfalls, as we show in section 2. In addition the Bayesian approach allows for asymptotic approximations of error measures through mechanical algorithms, while a strict frequentist approach might require tedious expansions, as we show in section 5.

Our objective is to present $A(f)$, and to show its comparative advantages over the HSHD, using the Sacramento dress rehearsal as an example. In this case f contains records for the 138,271 physically enumerated householders (Kostanich 1999), of whom 90,156 returned a census questionnaire by mail or were visited by an enumerator at a first attempt, and 48,115 were selected in a sample. We implement our method at the level of the tract, a connected unit of geography containing on average 1,300 householders in f.

The paper is organized as follows. In section 2 we illustrate the difficulties of designing a HSHD methodology that guarantees exchangeability. In section 3, we define $A(f)$, and in section 4 we present a likelihood for the model parameters. In section 5, we show how to implement $A(f)$ and derive a measure of error to make comparisons with the HSHD. Section 6 presents and motivates the basic model for the dress rehearsal, and section 7 gives results for both $A(f)$ and the HSHD in this case. In section 8, we summarize the differences and we make recommendations.

2. Assessing Exchangeability with Respect to a Partition by Domains of Study

We illustrate the difficulties inherent in designing a HSHD that preserves exchangeability between domains of study (Cochran 1977, page 34) with an example, where tenure (ownership) is the measurement of interest, and the

relevant domains of study are defined by race. To impute tenure, the Census Bureau uses the class variable "household type" to post-stratify *f* in five post-strata defined by the presence/absence of a live-in spouse for the householder, and the size of the household (1, 2, 3+) (Wilson 1998). The intent is to define post-strata that establish distributional homogeneity in terms of ownership at the level of the post-stratum, rendering the domain boundaries of a relevant partition uninformative within each post-stratum.

We examine the post-stratum comprising all the householders without a live-in spouse, and living in households of 3 or more. We call it post-stratum 3. For the purpose of this example, we have removed from *f* all the householders with unreported tenure, and each nearest neighbor is exclusive to a single householder. Table 1 gives householder frequencies for eight exhaustive race-tenure categories for post-stratum 3. Table 1 also gives the rate of ownership for their nearest neighbors, cross-classified by their race and by the same eight race-tenure categories of the corresponding householders. We observe that, on average, when a householder is either in the Black-owner or in the Black-renter category, his nearest neighbor is at least 25% more likely to be an owner when this nearest neighbor is White, than when he is Black. It is tempting to explain this differential rate by geographical differences. However, table 2, which shows the rates of ownership of the householders in post-stratum 3, cross-classified by their own race and that of their nearest neighbors, reveals that in fact Blacks with White nearest neighbors have a slightly lower rate of ownership than Blacks with Black nearest neighbors. What this means is that, if the probability of not reporting tenure is constant for all Blacks, then imputing their tenure by substituting the tenure of their nearest neighbor over-estimate ownership for Blacks in post-stratum 3.

These distributional disparities between householders and their nearest neighbors reflect a lack of exchangeability. A McNemar test leads to a formal rejection of the exchangeability hypothesis. There are 1,784 Black householders with White nearest neighbors. In 1,187 instances, tenure is tied. Among the 597 non-tied cases, the owner is White in 396 cases. Under the exchangeability hypothesis, ownership goes to either race with probability one-half. But the proportion of Whites among the owners is eight standard deviations above one-half. This example illustrates the difficulties in designing a valid NNHD that maintains exchangeability. In the next section we present our imputation method, which is devised for this type of situation.

3. An Imputation Method Based on Demographic Transition Probabilities

Besag (1974) describes the conditional probability approach to spatial processes. This approach gives a framework for probabilistically modeling the values of "sites", in terms of the values of their "neighbours" to construct a spatial process. Besag (1974) also suggests making a unilateral approximation to simplify this construction. Then, the value of each site depends only on a finite number of "predecessors". This approach is natural in our situation since f provides a unilateral ordering of householders who play the roles of sites and predecessors, in turn. Specifically, we construct a first-order process where each householder is a site, and the complete closest neighbor is

Table 1
Number of Householders and Rates of Ownership of the Nearest Neighbors in Post-Stratum 3 by Race of the Nearest Neighbor and Joint Race and Tenure of the Householder

	Race-Tenure Category of the Householder							
	White Owner	White Renter	Black Owner	Black Renter	Asian Owner	Asian Renter	Other Owner	Other Renter
Number of Householders in Post-Stratum 3	3,347	5,197	1,319	3,630	872	1,196	681	1,637
Rate of Ownership of the White Nearest Neighbors	0.556	0.564	0.562	0.299	0.561	0.287	0.540	0.163
Rate of Ownership of the Black Nearest Neighbors	0.379	0.189	0.427	0.211	0.443	0.202	0.471	0.158
Rate of Ownership of the Asian Nearest Neighbors	0.589	0.332	0.667	0.320	0.668	0.262	0.535	0.302
Rate of Ownerships of the Other Nearest Neighbors	0.423	0.251	0.497	0.237	0.595	0.177	0.463	0.152

Table 2
Rates of Ownership of the Householders in Post-Stratum 3 by Race of the Householder and Race of the Nearest Neighbor

	Race of the Nearest Neighbor			
	White	Black	Asian	Other
Rate of Ownership of the White Householders	0.415	0.358	0.384	0.337
Rate of Ownership of the Black Householders	0.257	0.264	0.304	0.267
Rate of Ownership of the Asian Householders	0.441	0.441	0.400	0.360
Rate of Ownership of the Other Householders	0.309	0.297	0.337	0.234

its only predecessor. In this set-up, the value of a site is the state of a householder, which we define shortly. We refer to the conditional probability for the value of a site given that of its predecessor as the transition probability from the state of the closest complete neighbor to the state of the householder. Our imputation methodology is based on the MLE of the transition probabilities at the level of a tract. In this section we describe the imputation methodology, and in the next section we introduce a likelihood for the transition probabilities.

Consider a population of householders in f representing a tract. Let Ψ represent a set of C categorical variables that characterize each householder. The variables are labeled $1, \ldots, C$, and have respectively $K_1, \ldots, K_C$ categories. Let $\Psi^{\times}$ denote the Cartesian product of the categorical variables in Ψ. Then, $\Psi^{\times}$ is the state space of the householder and has K states, where $K = \prod_{i \in \Psi} K_i$. Similarly, let Ξ be the set of E categorical variables defining the closest complete neighbor in f. The variables are labeled $1, \ldots, E$, and have $F_1, \ldots, F_E$, categories. $\Xi^{\times}$ is the state space of the closest complete neighbor and has F states, where $F = \prod_{i \in \Xi} F_i$. The items represented in Ξ are also represented in Ψ. Let the state of the householder be $\boldsymbol{s} \in \Psi^{\times}$, where $\boldsymbol{s}$ is a vector whose components represent the variables in Ψ. Similarly, $\boldsymbol{t} \in \Xi^{\times}$ is the state of the closest complete neighbor. Under the assumptions above, let $P(\boldsymbol{s} \mid \boldsymbol{t})$ represents the transition probability from $\boldsymbol{t}$ to $\boldsymbol{s}$ in the order of f. Now suppose a householder only reported the categorical variables in a subset $\mathrm{Z} \subset \Psi$. Let $\boldsymbol{v} \in \mathrm{Z}^{\times}$ be the vector of reported variables. Let $\sigma(\Psi, \mathrm{Z}, \boldsymbol{v}) \subset \Psi^{\times}$ be the subset containing all the values of $\boldsymbol{s}$, such that $\boldsymbol{s}$ agrees with $\boldsymbol{v}$ on the variables in Z. Define

$$P(\boldsymbol{s} \mid \boldsymbol{t}, \mathrm{Z}, \boldsymbol{v}) = \frac{P(\boldsymbol{s} \mid \boldsymbol{t})}{\sum_{\boldsymbol{u} \in \sigma(\Psi, \mathrm{Z}, \boldsymbol{v})} P(\boldsymbol{u} \mid \boldsymbol{t})}; \; \boldsymbol{s} \in \sigma(\Psi, \mathrm{Z}, \boldsymbol{v}). \quad (1)$$

To impute the items in the set difference $\Psi - \mathrm{Z}$ according to $A(f)$, we roll dice weighted by the values of the MLE of $P(\boldsymbol{s} \mid \boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})$, for each householder in marginal state $\boldsymbol{v}$ and with closest complete neighbor in state $\boldsymbol{t}$. Under our assumptions, the MLE of $P(\boldsymbol{s} \mid \boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})$ contains all the information available from f on the unreported items. In the next section we formulate a likelihood for $P(\boldsymbol{s} \mid \boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})$.

4. A Likelihood for the Transition Probabilities

Let $N(\boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})$ be the number of householders who only reported the items defining the marginal state $\boldsymbol{v}$ involving only the items in $\mathrm{Z} \subset \Psi$, and with closest complete neighbor in state $\boldsymbol{t}$. Let $\boldsymbol{N}$ be a vector with the $N(\boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})$'s as its components, at the level of a tract. Let $\boldsymbol{P} = [P(\boldsymbol{s} \mid \boldsymbol{t})]$ be the vector comprising the $P(\boldsymbol{s} \mid \boldsymbol{t})$'s ordered lexicographically by $\boldsymbol{t}$ and $\boldsymbol{s}$. Based on the assumptions described above, we have the following likelihood for the transition probabilities.

$$L(\boldsymbol{N}; \boldsymbol{P}) = \prod_{\boldsymbol{t} \in \Xi^{\times}} \prod_{\mathrm{Z} \subset \Psi} \prod_{\boldsymbol{v} \in \mathrm{Z}^{\times}} \left(\sum_{\boldsymbol{s} \in \sigma(\Psi, \mathrm{Z}, \boldsymbol{v})} P(\boldsymbol{s} \mid \boldsymbol{t}) \right)^{N(\boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})}; \quad \boldsymbol{P} \in \Theta_{\boldsymbol{P}}. \quad (2)$$

The running indices in (2) are $\boldsymbol{t}, \mathrm{Z}, \boldsymbol{v}$, and $\boldsymbol{s}$. If every item is reported, then Ψ is the only instance of Z with $N(\boldsymbol{t}, \mathrm{Z}, \boldsymbol{v}) \neq 0$, for some $\boldsymbol{t}$ and $\boldsymbol{v}$. In that case (2) is analogous to the likelihood of the transition probabilities of a first-order Markov chain (Bishop, Fienberg and Holland 1975 page 263). In general, we model $\Theta_{\boldsymbol{P}}$ as a log-linear subspace. For this purpose it is more convenient to work with an expression equivalent to (2) that has a simpler algebraic representation. We introduce the nuisance parameter $\boldsymbol{U} = [U(\boldsymbol{t})]$, where $\boldsymbol{U}$ is a probability vector, that is $\sum_{\boldsymbol{t} \in \Xi^{\times}} U(\boldsymbol{t}) = 1$, and $0 < U(\boldsymbol{t}) < 1$, for all $\boldsymbol{t} \in \Xi^{\times}$. $\boldsymbol{U}$ represents the prevalences of the states of the closest complete neighbors. Let $Q(\boldsymbol{s}, \boldsymbol{t}) = U(\boldsymbol{t}) \times P(\boldsymbol{s} \mid \boldsymbol{t})$, and $\boldsymbol{Q} = [Q(\boldsymbol{s}, \boldsymbol{t})]$. Then $\boldsymbol{Q}$ is a probability vector with $\boldsymbol{K} \times \boldsymbol{F}$ components lexicographically ordered by $\boldsymbol{t}$ and $\boldsymbol{s}$. We set up Θ, the parameter space of $\boldsymbol{Q}$, as a hierarchical log-linear model (Agresti 1990, page 143; Bishop, Fienberg and Holland 1975, page 67). Then, if we design Θ so that it includes the interactions of all orders between the variables in Ξ, (2) is equivalent to the following likelihood in terms of $\boldsymbol{Q}$.

$$L^{*}(\boldsymbol{N}; \boldsymbol{Q}) = \prod_{\boldsymbol{t} \in \Xi^{\times}} \prod_{\mathrm{Z} \subset \Psi} \prod_{\boldsymbol{r} \in \mathrm{Z}^{\times}} \left(\sum_{\boldsymbol{s} \in \sigma(\Psi, \mathrm{Z}, \boldsymbol{r})} Q(\boldsymbol{s}, \boldsymbol{t}) \right)^{N(\boldsymbol{t}, \mathrm{Z}, \boldsymbol{r})}; \quad \boldsymbol{Q} \in \Theta. \quad (3)$$

That is, if Θ has the architecture described above, a specific choice for Θ unambiguously defines $\Theta_{\boldsymbol{P}}$ in (2), and since the items of the closest complete neighbor are always reported, the factorization $L(\boldsymbol{N}; \boldsymbol{P}) = L^{*}(\boldsymbol{N}; \boldsymbol{Q}) \times R(\boldsymbol{N}; \boldsymbol{U})$ holds, for some $R(;)$. (3) is easier to manipulate than (2) since it corresponds to the likelihood of the cell probabilities associated with a partially classified contingency table (Little and Rubin 1987, page 181). Under mild conditions on the non-response mechanism (for example, strictly positive and constant probabilities for each response configuration (Thibaudeau 1988)) the likelihoods in (2) and (3) are identifiable and asymptotically unimodal. Multimodality is theoretically possible for finite samples, but it does not appear to occur in the cases studied in the paper, where the proportions of unreported items are small.

5. Finding the MLE and Deriving Measures for the Non-Response Error

In this section, we recall how to compute $\hat{\boldsymbol{P}}$, the MLE of $\boldsymbol{P}$, and we derive measures of errors for $A(f)$ and another predictor $\hat{S}(\boldsymbol{s})$, which we term the "MLE" of the expected value of $S(\boldsymbol{s})$, which is the actual count of householders in state $\boldsymbol{s}$ at the tract level. An error measure for $\hat{S}(\boldsymbol{s})$ will be useful in section 7 to evaluate the imputation results

obtained with $A(f)$ relative to those with the HSHD. We compute $\hat{\boldsymbol{P}}$ by maximizing (3), in terms of $\boldsymbol{Q}$, with the EM algorithm. Because of the factorization described in section 4, this maximum also yields $\hat{\boldsymbol{P}}$.

To derive measures of error in predicting $S(\boldsymbol{s})$ for a given $\boldsymbol{s}$, consider all the triples of the form $(\boldsymbol{t}, \mathrm{Z}, \boldsymbol{v})$ in (1) that are observed in the sample (*i.e*, the tract) for which it is possible, but due to item non-response it is not known, that one or more householders corresponding to such a triple are in state $\boldsymbol{s}$. Let $\Lambda(\boldsymbol{s})$ be the number of such triples. We index these triples with $\lambda = 1, \ldots, \Lambda(\boldsymbol{s})$. Let $\delta(\lambda)$ be the number of householders corresponding to triple λ, and let $\rho_\lambda(\boldsymbol{s})$ be the probability that such a householder is indeed in state $\boldsymbol{s}$, where $\rho_\lambda(\boldsymbol{s})$ is derived from $\boldsymbol{P}$. Let $\Delta(\boldsymbol{s}, \lambda)$ be the unknown number of householders who are indeed in state $\boldsymbol{s}$ among the $\delta(\lambda)$ candidates. Based on our model we have $S(\boldsymbol{s}) = S_{\mathrm{obs}}(\boldsymbol{s}) + \sum_{\lambda=1}^{\Lambda(\boldsymbol{s})} \Delta(\boldsymbol{s}, \lambda)$, where $S_{\mathrm{obs}}(\boldsymbol{s})$ is the number of householders who reported being in state $\boldsymbol{s}$ and $\Delta(\boldsymbol{s}, \lambda)$ is Binomial $(\delta(\lambda), \rho_\lambda(\boldsymbol{s}))$. Furthermore, let $\hat{S}(\boldsymbol{s}) = S_{\mathrm{obs}}(\boldsymbol{s}) + \sum_{\lambda=1}^{\Lambda(\boldsymbol{s})} \delta(\lambda) \hat{\rho}_\lambda(\boldsymbol{s})$, where $\hat{\rho}_\lambda(\boldsymbol{s})$ is the MLE of $\rho_\lambda(\boldsymbol{s})$. If we treat the λ's as independent predictors, like in a regression situation, and since $\hat{\boldsymbol{P}}$ is asymptotically normal with mean $\boldsymbol{P}$, we have the following large sample approximation for the MSE of $\hat{S}(\boldsymbol{s})$ in predicting $S(\boldsymbol{s})$.

$$E\left[\left(\sum_{\lambda=1}^{\Lambda(\boldsymbol{s})} \delta(\lambda)\,\hat{\rho}_\lambda(\boldsymbol{s}) - \Delta(\boldsymbol{s}, \lambda)\right)^2 \,\middle|\, \boldsymbol{P}\right]$$

$$\approx \mathrm{V}\left(\sum_{\lambda=1}^{\Lambda(\boldsymbol{s})} \delta(\lambda)\hat{\rho}_\lambda(\boldsymbol{s}) \,\middle|\, \boldsymbol{P}\right) + \mathrm{V}\left(\sum_{\lambda=1}^{\Lambda(\boldsymbol{s})} \Delta(\boldsymbol{s}, \lambda) \,\middle|\, \boldsymbol{P}\right). \qquad (4)$$

Let V_p and V_ε be the first and second variances on the RHS of (4). Gelman and Rubin (1991), Larsen (1996), and Schafer (1997, page 324) introduce data augmentation Bayesian iterative proportional fitting (DABIPF) to simulate posterior and predictive distributions associated with log-linear models with data missing at random. We can use DABIPF to approximate model-consistent estimators for V_p and $V_\varepsilon + V_p$ through simulations of the posterior distribution of $\sum_{\lambda=1}^{\Lambda(\boldsymbol{s})} \delta(\lambda)\,\rho_\lambda(\boldsymbol{s})$ and the predictive distribution of $S(\boldsymbol{s})$ respectively. Furthermore, we approximate the MSE of the demographic counts obtained imputing with $A(f)$ by adding another V_ε to $V_\varepsilon + V_p$ in (4) to account for the additional noise of the "dice roll" involved in $A(f)$.

6. Modeling and Sensitivity Analysis

6.1 A Conditional Independence Model for Sacramento

Using the notation of section 3, the householder variables in Ψ are race, origin, tenure, and sex. The categories for race are White, Black, Asian, and Other. For origin they are Hispanic and non-Hispanic. For tenure they are owner and renter. For sex they are male and female. The neighbor variables in Ξ are race, origin, and tenure. The categories for race of the neighbor are Black and non-Black. The categories for origin and tenure of the neighbor are the same as for the householder. We design Θ in (3), by selecting interactions between the variables in Ψ and Ξ. To ensure equivalence between (2) and (3), we select the interactions of all orders between the variables in Ξ. We attempt to maintain through the imputations the correlation between successive householders in f in terms of each item in Ξ. Thus we include each interaction associating an item in Ξ to the corresponding item in Ψ. We complete the model by selecting consistency associations: We include the six interactions representing the associations involving a pair of items in Ψ. The resulting contingency table has 256 cells, and the log-linear model has thirty free parameters.

This model leads to a conditional independence transition structure. For example, conditional on the race of the closest complete neighbor, the race of the householder is independent of the tenure of the closest complete neighbor. Conditional independence allows us to combine neighbor information obtained from multiple neighbors to produce a synthetic closest complete neighbor. This approach ensures that we can use all the information available from the closest neighbor, even if he is not complete. With this approach, the correlation structure among the items of the householder is maintained whenever only one item per householder is imputed. In Sacramento, among 138,271 householders, approximately 0.1% did not report sex, 3.5% did not report race, 2.9% did not report origin, and 7.6% did not report tenure. Furthermore, race and origin are missing jointly for 0.49% of the householders, race and tenure 0.48%, origin and tenure 0.69%. Given these low rates of jointly missing items, we expect our model to do well.

6.2 Sensitivity Analysis and Evaluation

In section 7 we use the standard error of the predictive distribution of $S(\boldsymbol{s})$ to approximate $\sqrt{V_\varepsilon + V_p}$, the error of $\hat{S}(\boldsymbol{s})$ in predicting $S(\boldsymbol{s})$, as derived in (4), and we assume asymptotic normality of $\hat{S}(\boldsymbol{s}) - S(\boldsymbol{s})$. The accuracy of this approximation depends on the accuracy of the approximation of the distribution of the MLE $\hat{\boldsymbol{P}}$ with the posterior distribution of $\boldsymbol{P}$. This later approximation is accurate asymptotically when the model holds, but we still need to verify the extent to which this asymptotic result is applicable when the sample is finite. To do so we examine the sensitivity of the posterior distribution of $\boldsymbol{P}$ under prior changes. A low sensitivity implies that the posterior distribution of $\boldsymbol{P}$ is a good approximation of the distribution of $\hat{\boldsymbol{P}}$. We focus on the posterior distribution for the conditional probability that origin is Hispanic, conditional on each race. An increase of 0.1 in the value of α, the prior parameter of the constrained Dirichlet family (Schafer 1997, page 346), which is the natural family for (3), is equivalent to observing three additional Hispanics and three additional Non-Hispanics of each race. Table 3 gives the posterior

modes and standard deviations (SD) of the posterior density of the conditional probability that origin is Hispanic given each race, for four choices of α, for a specific tract X. Figure 1 shows the posterior of the conditional probability given race is White. This posterior is stable under prior disturbances and we expect it to give a good approximation for the distribution of the corresponding MLE. On the other hand, Figure 2, which shows the posterior of the conditional probability given race is Black, displays a high sensitivity, suggesting that our proposed asymptotic approximation is less accurate in this case. This is not surprising in light of the facts that, for Blacks, the MLE of the conditional probability is close to 0 and the domain (race) size is smaller (among the 1,583 householders in tract X, there are 1,087 Whites, 179 Blacks, 56 Asians, 172 Others, while 89 did not report race). In the next section we focus on cases where the conditional probabilities are not near 0 or 1, and the size of the domain is large. We retain the choice $\alpha = 0.01$ for the prior, which is approximately Jeffrey's prior on the marginal conditional probabilities that define the model. It is beyond the scope of the paper to address the difficulties when the domain is small and/or the MLE is near 0/1.

Table 3
MLE, Posterior Mode (approximate), and Standard Deviation for the Conditional Probabilities of Origin Being Hispanic given Race for Four Choices of Prior Distribution

Race	MLE	Mode $\alpha = 0.01$	S.D. $\alpha = 0.01$	Mode $\alpha = 0.1$	S.D. $\alpha = 0.1$	Mode $\alpha = 0.5$	S.D. $\alpha = 0.5$	Mode $\alpha = 1$	S.D. $\alpha = 1$
White	0.1784	0.178	0.01195	0.184	0.01247	0.180	0.01219	0.188	0.01186
Black	0.07428	0.069	0.02272	0.081	0.02330	0.120	0.02428	0.160	0.02782
Asian	0.09113	0.105	0.04086	0.108	0.04550	0.195	0.04881	0.276	0.04952
Other	0.9662	0.966	0.01171	0.964	0.01347	0.950	0.01495	0.930	0.01666

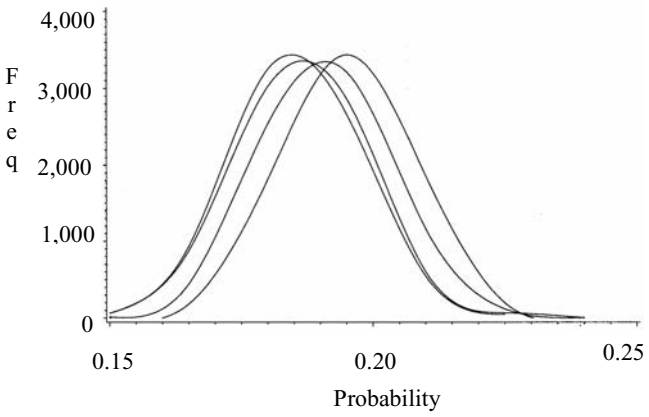


Figure 1. Posterior Distribution Prob. Origin is Hispanic – White Householder.

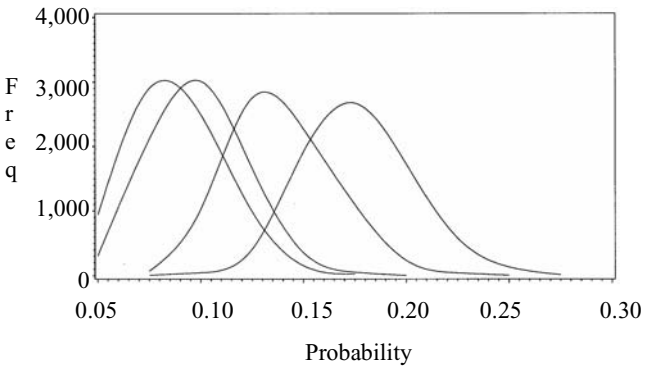


Figure 2. Posterior Distribution Prob. Origin is Hispanic – Black Householder.

7. Results for the Sacramento Dress Rehearsal

Table 4 gives count estimates at the level of Sacramento derived with $A(f)$ based on the model of section 6.1 fitted for each of the 102 tracts, as well as count estimates obtained with the HSHD. Table 4 also gives error measurements based on a sequence of 2000 DABIPF iterations with 2000 burn-in iterations, for each of the 102 tracts in Sacramento (see appendix A for convergence), serving to approximate $\sqrt{V_\varepsilon + V_p}$ derived from (4). We call $\sqrt{V_\varepsilon + V_p}$ the prediction error of the MLE. We estimate $\sqrt{V_\varepsilon}$ separately by "rolling dice" loaded with the MLE. We call $\sqrt{V_\varepsilon}$ the model residual error. We use $\sqrt{2V_\varepsilon + V_p}$, which we call the total imputation error, to express the error of $A(f)$ in estimating the true count. If we assume $\hat{S}(\boldsymbol{s})$ is positively correlated with the HSHD, the prediction error of the MLE can be used as an upper bound for the standard error of the distance between the count estimates corresponding to the MLE and the HSHD. For the Black owners, this distance is severely incompatible with the hypothesis that the MLE and the HSHD have the same expectation. This is no surprise in light of the results of section 2.

Interestingly, the results of table 4 can serve to improve the performance of the HSHD. Since tenure is unreported twice as often as race, our results for the Black owners suggest improving the HSHD by including race as a class variable for the imputation of tenure with the HSHD. Table 5 shows results obtained with this re-engineered HSHD, and exchangeability of tenure between nearest neighbors based on this new post-stratification is more plausible than for the original scheme.

Table 4
Population Counts and Uncertainty Measures for Sacramento

	Imputed Count with HSHD	Imputed Count with Model	MLE of the Expected Count	Model Residual Error	Prediction Error of the MLE	Total Imputation Error
All	138,271	138,271	138,271.0	0.0	0.0	0.0
White	89,032	88,914	88,927.7	31.5	35.2	47.2
Black	19,962	19,943	19,952.9	14.9	16.5	22.3
Asian	17,405	17,421	17,426.2	14.0	14.9	20.5
Other	11,872	11,993	11,964.1	29.8	33.5	44.8
Hispanic	21,024	21,050	21,038.1	10.3	10.6	14.7
Non-Hispanic	117,247	117,221	117,232.8	10.3	10.6	14.7
Owner	70,054	70,022	70,026.3	42.8	43.3	60.9
Renter	68,217	68,249	68,244.7	42.8	43.3	60.9
White Hispanic	9,068	8,972	8,991.1	29.9	33.6	45.0
White Non-Hispanic	79,964	79,942	79,936.6	15.4	15.7	22.0
Black Hispanic	605	612	608.6	11.0	12.6	16.7
Black Non-Hispanic	19,357	19,331	19,344.3	10.8	10.7	15.2
Asian Hispanic	518	515	516.5	10.0	11.5	15.2
Asian Non-Hispanic	16,887	16,906	16,909.7	10.4	10.3	14.6
Other Hispanic	10,833	10,951	10,921.9	29.7	33.3	44.6
Other Non-Hispanic	1,039	1,042	1,042.3	3.5	3.4	4.9
White Owner	47,722	47,767	47,770.5	37.8	41.3	56.0
White Renter	41,310	41,147	41,157.3	39.0	41.4	56.9
Black Owner	7,661	7,538	7,542.3	19.6	20.7	28.5
Black Renter	12,301	12,405	12,410.6	21.1	22.5	30.8
Asian Owner	9,810	9,853	9,872.8	18.4	18.6	26.1
Asian Renter	7,595	7,568	7,553.4	18.2	18.8	26.1
Other Owner	4,861	4,864	4,840.7	24.4	28.2	37.3
Other Renter	7,011	7,129	7,123.4	25.4	28.6	38.2
Hispanic Owner	9,409	9,434	9,402.2	19.5	20.9	28.6
Hispanic Renter	11,615	11,616	11,629.9	20.1	21.4	29.4
Non-Hispanic Owner	60,645	60,588	60,618.0	38.9	39.4	55.4
Non-Hispanic Renter	56,602	56,633	56,614.8	38.7	39.6	55.4

Table 5
HSHD with Race as an Additional Class Variable

	Imputed Count With HSHD	Imputed Count With HSHD Re-Engineered With Race as a Class Variable	Imputed Count With Model	MLE Of The Expected Count	Prediction Error Of the MLE
White Owner	47,722	47,687	47,767	47,770.5	41.3
Black Owner	7,661	7,573	7,538	7,542.3	20.7
Asian Owner	9,810	9,851	9,853	9,872.8	18.6
Other Owner	4,861	4,840	4,864	4,840.7	28.2
Owner	70,054	69,951	70,022	70,026.3	43.3

8. Conclusion

In section 2 we have shown that the HSHD may fail to retrieve exchangeable householders, producing a bias relative to a situation where exchangeability holds. As more evidence that $A(f)$ partly corrects this relative bias, we compare the observed and the imputed cross-product ratios (Bishop, Fienberg and Holland 1975, page 14) between two races (Black, White) and the two tenures. We look at the cross product ratio involving:

1. Only observed householders.
2. Householders with tenure imputed with the HSHD.
3. Householders with tenure imputed with $A(f)$.

There are 73 tracts where all these cross-product ratios can be measured. 2. The HSHD produces cross-product ratios smaller than those observed for 53 tracts. $A(f)$ displays more symmetry as it produces cross-product ratios smaller than observed only for 43 tracts. A sign test confirms that $A(f)$ (p = 0.064) is more in sync with the observations than the HSHD ($p = 0.0001$).

In general, we expect the HSHD to give good count estimates when the householders tend to geographically coalesce by domain of study. But difficulties arise in a situation where domains of study exhibiting substantial distributional dissimilarities are geographically integrated. In such a situation, implementing the HSHD requires accurate parsing of the class variables. Frugality is tantamount when specifying class variables, but at the same time the price to pay for omitting a crucial variable can be substantial. Thus the designer of the HSHD has little room for error. By contrast, although model misspecification certainly remains a danger, the user of $A(f)$ has more freedom to posit several domain partitions without impeding on the ability of $A(f)$ to adjust the imputations for the local geographical character, based on information from the closest complete neighbor. $A(f)$ will be useful to impute categorical measurements when the impact of the relevant domain partitions on the measurements is not known a priori, and some of the relevant domains may define small subpopulations dispersed within the entire population. Then, based on policy considerations, $A(f)$ can be applied directly, or to help parse the class variables of the HSHD, as we did in section 7.

A referee notes that a comparison with a procedure based on an unbiased sample, building on the method of Zanutto and Zaslavsky (1995a, b), would be a defining test for $A(f)$. This procedure would require collecting information from the item non-respondents on a scale sufficiently large to ensure bias detection, and we should take advantage of any such opportunity to perform a test of this type. Unfortunately, because of limited resources, samples containing this information are seldom collected. Nevertheless, we are hopeful that the analysis of the returns from Census 2000 aided with procedural information can provide new insights on the reliability of $A(f)$.

Acknowledgements

The author is indebted to William Winkler for his guidance. The author is grateful to two referees for their discernment, to Eric Slud, Don Malec and Joseph Schafer for essential discussions, and to Andrew Gelman and Don Rubin for providing their unpublished paper. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

Appendix A – Convergence of DABIPF

We ran two chains of 8,000 iterations each, with over-dispersed starting points, for the case $\alpha = 0.01$, for tract *X*. We computed $\sqrt{\hat{R}}$ (Gelman and Rubin 1992) for $\boldsymbol{Q}(\boldsymbol{s}, \boldsymbol{t})$ in (3), for sequences of 1,000, 2,000, and 4,000 iterations, after burn-in lags of 1,000, 2,000, and 4,000 iterations respectively. After 2,000 iterations, with 2,000 burn-in iterations, we observed that $\sqrt{\hat{R}} \leq 1.010$ in all studied cases, including those in table 3. We think this level of accuracy is acceptable for approximating modes and variances.

References

Agresti, A. (1990). *Categorical Data Analysis*. Wiley-Intersience.

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. and Holland, P.W. (1975). *Discrete Multivariate Analysis*. MIT Press.

Chen, J., and Shao, J. (1997). Biases and variances of survey estimators based on nearest neighbor imputation. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 365-369.

Chen, J., and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16, 2.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition. Wiley.

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.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 227-232.

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.

Fay, R.E., and Town, M.K. (1998). Variance estimation for the 1998 census dress rehearsal. *Proceedings for the Section on Survey Research Methods*, American Statistical Association, 605- 610.

Gelman, A., and Rubin, D.B. (1991). Simulating the Posterior Distribution of Loglinear Contingency Table Models. Unpublished Technical Report, Harvard University.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 4.

Kostanich, D.L. (1999). DSSD Census 2000 Dress Rehearsal Memorandum Series #A, US Bureau of the Census.

Kovar, J.G., and Whitridge, P.J. (1995). Imputation of Business Survey Data. *Business Survey Methods*, (Eds. Cox, D., Binder, Chinnappa, Christianson, M. Colledge and Kott). Wiley.

Larsen, M.D. (1996). *Bayesian Approaches to Finite Mixture Models.* Doctoral Dissertation, Department of Statistics, Harvard University.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Meng, X.M. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 4.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 4.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 434.

Sande, I.G. (1981). Imputation in surveys: coping with reality. *Survey Methodology*, 7, 21-43.

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

Thibaudeau, Y. (1988). *Approximating the Moments of a Multimodal Posterior Distribution with the Method of Laplace*. Doctoral Dissertation, Department of Statistics, Carnegie Mellon University.

Treat, J.B. (1994). *Summary of the 1990 Census Imputation Procedures for the 100% Population and Housing Items*. DSSD REX Memorandum Series BB-11, US Bureau of the Census.

Wilson, E.B. (1998). Communication to Dan E. Philip. Housing and Household Economics Statistics Division, US Bureau of the Census.

Zanutto, E., and Zaslavsky, A.M. (1995a). A model for imputing nonsample households with sampled nonresponse follow-up. *Proceedings for 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 follow-up. *Proceedings of the Annual Research Conference*, U.S. Department of Commerce, Bureau of the Census. 673