

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Data Editing**  
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

**MODEL EXPLICIT ITEM IMPUTATION FOR DEMOGRAPHIC SURVEYS AND CENSUSES**

Submitted by the U.S. Bureau of the Census<sup>1</sup>

**Contributed paper**

**I. INTRODUCTION**

1. The paper presents a model-based item imputation methodology intended as an alternative to the hot deck substitution algorithms for demographic surveys and censuses. We develop an imputation methodology based on specific statistical assumptions. These assumptions, when satisfied, guarantee an imputation with desirable statistical properties. The imputations produced with our methodology indirectly enjoy the properties of consistency and efficiency, in the sense that they are closely associated with consistent and efficient estimators. We intend to develop our methodology in the bayesian framework. We feel it is a natural environment for the development of simulation processes. In this framework, we treat the parameters of a model as random variables. This makes it easier to quantify the error of our imputation procedure.

2. In subsection II.1, we begin the discussion with a review of the items subject to imputation in the short questionnaire for the 1998 dress rehearsal for Census 2000 in the U.S. We use this example throughout the paper to illustrate our procedures. In subsection II.2, we review the methodology used by the Bureau of the Census to impute items that were not reported in the short questionnaire.

3. In section III we present our model-based imputation procedure. Subsection III.1 introduces the model that will serve as the inferential engine for the item imputation and to quantify the errors intrinsic in the procedure. Section III.2 details the steps of the procedure. Subsection III.3 reviews the underlying bayesian theory that motivates our methodology.

4. In section IV we present results on the item imputation for the 1998 dress rehearsal for Census 2000. We evaluate the relative biases between the population counts based on the imputations from the model-based procedure with those based on the imputations from the traditional hot deck procedure. Subsection IV.1 presents the results for 28 demographic categories. Subsection IV.2 gives an in-depth analysis of the imputation of tenure for the "Black Households".

---

<sup>1</sup> Prepared by Yves Thibaudeau.

## II. BACKGROUND

### II.1 The 2000 Short Form

5. In March 1998 the Census Bureau conducted a dress rehearsal in preparation for the 2000 Census. Three sites were targeted for this operation: the city of Sacramento, California; a portion of the rural area of South-Carolina; and a Menomonee Indian reservation. The Sacramento test site is more diverse than the others in terms of race and Hispanic origin. We selected it to experiment with our item imputation procedure. Each housing unit in Sacramento was sent a census questionnaire requesting six demographic items for each occupant of the housing unit: tenure, race, Hispanic origin, sex and age. About 15 % of the housing units were asked additional questions. To demonstrate our methodology we focus on these six items. A mail reply is expected from the household in each housing unit. Units for whom no mail reply is received are the objects of a non-response follow-up operation, which, in 1988, included a sampling procedure. After the data collection operations are completed, there remains a substantial proportion (about 10%) of incomplete records corresponding to units who did not provide all the demographic items requested. These items must be imputed.

6. The paper focuses on imputation of the household items. We define four household items which unambiguously characterize each household. The first household item is tenure, i.e. the home ownership status of the household. The three other household items are defined through the householder. They are race of the householder, origin of the householder, and sex of the householder. There is exactly one householder per household. Therefore, these items are uniquely defined for each household. It is clear why tenure is considered a household item. There are operational reasons for treating race and origin as household variables. When race or origin are unreported for an entire household, the values for the race or origin of the householder are imputed through statistical procedures. Then, the values of the imputations carry to the other members of the household. This procedure does not account for mixed households, but is a reasonable approximation of the reality. Note that when race or origin are reported for at least one member of the household, then the imputation is deterministic, in the sense that the unreported items are substituted with reported items according to a predetermined hierarchy (1. substitute from the brother, 2. from the mother, etc.). The sex of the householder seldom needs to be imputed. We include it in the household variables because it interacts with them. We refer to Williams (1998) for a model-based imputation procedure to impute age and sex for household members other than the householder. In the next subsection we review the traditional imputation methodology for the population census.

### II.2 The Sequential Hot Deck

7. The Census Bureau uses a sequential hot deck (Kovar and Whitridge, 1995) to process the item-imputation for the decennial census. The sequential hot deck (SHD) is essentially a one-pass algorithm. Except for a preliminary initialization, the SHD processes each census record only once and imputes any unreported item on the spot by substituting the last recorded value for the item. In general, the records are sorted according to geographical proximity, and thus the last record in the census file usually corresponds to a nearby neighbor. With this simple scheme, only a few reported values for each item need to be kept in memory during processing. This approach is a carry-over from an era when each pass through the census file was very costly in time and resources. In the 1998 version, additional constraints are imposed on the SHD through the class variables (Treat, 1994). For example, race is a class variable for origin. Accordingly, when a household does not report the origin item, it is borrowed from the last household of the same race who reported the item. Subject-matter experts are responsible for the selection of the class variables.

8. Fay and Town (1998) suggest a rationale for relying on the nearest neighbors and the SHD. They mention the concept of local exchangeability, which is akin to assumptions in some non-parametric procedures. The exchangeability assumption certainly holds when the demographic composition of a population is locally homogeneous with respect to geography. For example, we can expect home ownership, or the lack thereof, to often identically characterize immediate neighbors. Indeed, owners tend to congregate, as do renters. Local homogeneity may also apply to items such as race and Hispanic origin. Although exchangeability remains a functional concept even in situations where local homogeneity may not

apply, in such situations the SHD imputes unreported items on the basis of partial information. It is conceivable that there is valuable information for the purpose of item imputation in records that do not correspond to physically close households whose items are unreported. The SHD generally ignores any such information. In addition, there are situations where the SHD ignores close neighbors. This happens when the class variable points to a household that is physically at some distance. For example, when imputing origin, a household of the same race must be retrieved to provide a substitute origin item. It could be that the nearest household of the same race is some distance away. In such a case, the SHD ignores information on the direct neighbors. In the next section we propose an item imputation procedure articulated on the basis of information retrieved at various levels of geography.

### III. MODEL-BASED ITEM IMPUTATION

#### III.1 A Model for Population Counts

9. We present our model-based imputation methodology as an alternative to the SHD for household item imputation. First, we attempt to give an intuitive motivation for our approach. The corner stone of the model-based approach is the specification of a statistical model designed to capture the trends shaping the data. We introduce a model representing a probabilistic population process at the level of a tract. A tract is a connected geography of approximately 1700 households. The model integrates a probabilistic structure conveying information for the imputation of unreported items of a given household. The information is brought from three levels of geography: the level of the household, the level of the neighbor and the level of the tract. The household provides its own fragmented information since, for the vast majority of households, some of the household items are reported. Neighbor and tract level information is always available. Tracts are large enough to provide a plentiful information. Neighbors who can serve as information providers are available. Such a neighbor may not be directly adjacent to the household of interest.

10. We establish the notation. Let  $N_{ijklmno}$  be the population count for the households with tenure  $i$ , race  $j$ , origin  $k$ , sex  $l$ , tenure of the neighbor  $m$ , race of the neighbor  $n$ , and origin of the neighbor  $o$ , for a specific tract. The notation for tenure is  $i = 1$  if the unit of the household is owned, and  $i = 2$  if it is rented. The notation for race is  $j = 1$  if the race of the householder is White,  $j = 2$  if the race is Black,  $j = 3$  if the race is Asian, and  $j = 4$  if the race is Other. The notation for origin is  $k = 1$  if the origin of the householder is non-Hispanic, and  $k = 2$  if the origin is Hispanic. The notation for sex is  $l = 1$  if the sex of the householder is male, and  $l = 2$  if the sex is female. The indices  $i, j, k, l$  delineate 32 categories or cells for the households. We further characterize the households in terms of the demographics of the neighbor, the household preceding the referenced household in the order of the census file. The tenure of the neighbor is represented by  $m = 1$  if the neighbor owns the neighboring housing unit, and  $m = 2$  if the neighbor rents. The race of the neighbor is given by  $n = 1$  if the neighbor (householder) is non-Black, and  $n = 2$  if the neighbor is Black. The origin of the neighbor is defined by  $o = 1$  if the neighbor is non-Hispanic, and  $o = 2$  if the neighbor is Hispanic. This notation distinguishes 256 types of households for each tract.

11. Assume that the total for the household population  $N$  is known for a given tract. Let  $p_{ijklmno}$  be the inclusion probability for type  $i, j, k, l, m, n, o$ , i.e. the probability that an arbitrary household in the tract is of that type. Then  $\{N_{ijklmno}\}$ , the population counts for the 256 household types, has a multinomial distribution with total count  $N$ , and probabilities  $\{p_{ijklmno}\}$ . The likelihood function is

$$L(\{N_{ijklmno}\}; \{p_{ijklmno}\}) = \prod_{ijklmno} (p_{ijklmno})^{N_{ijklmno}} \quad (1)$$

This likelihood involves 255 parameters. Initially, there is only one constraint imposed on the 256 inclusion probabilities. They must be greater than zero and must add up to one. Our goal is to construct a meaningful model based on a small number of key parameters. We therefore submit the inclusion probabilities to additional constraints expressing our knowledge on the behavior of the population of a tract. We choose the set of constraints known as the homogeneous association assumptions (Schafer 1997, p. 293). We set these

constraints in the model through the specification of a log-linear model:

$$\log(p_{ijklmno}) = m + T_i + R_j + H_k + S_l + t_m + r_n + h_o + (T * R)_{ij} + (T * H)_{ik} + (T * S)_{il} + (R * H)_{jk} + (R * S)_{jl} + (H * S)_{kl} + (T * t)_{im} + (R * r)_{jn} + (H * h)_{ko} \quad (2)$$

12. The parameters with one subscript represent the main effects corresponding to the items for the values identified by the subscript. The parameters with two subscripts represent the interaction effects between two items for the values identified by the subscripts. To prevent redundancies, we set to zero the sums of the parameters over one index ( $\sum_i T_i = 0$ ,  $\sum_i (T * R)_{ij} = 0$ , etc.). The architecture of the log-linear model defined in (2) is inspired by some of the same concepts that motivate the SHD, since the model includes interaction effects between the items of neighbouring households.

### III.2 A Model-Based Imputation Methodology

13. The model defined in (1) and (2) provides us with the tools to impute unreported items. Since some items may be unreported, we assume that  $N_{ijklmno}$  may not be available for some values of  $i, j, k, l, m, n, o$ . Let  $\Omega$  represent the reported information.  $\Omega = \{O_{ijklmno}\}$  is the set of observed counts and  $O_{ijklmno}$  is the count for observed household type  $i, j, k, l, m, n, o$ . In addition to the values given in 3.1, the indices  $i, j, k, l$  can take the value “\*” to indicate that the item was not reported. For example  $O_{ij*lmno}$  is the observed population count of households with unreported origin, and with tenure, race, sex identified by  $i, j, l$ , and tenure, race, origin of the neighbour identified by  $m, n, o$ . The model-based methodology for the imputation of unreported household items unfolds in two stages:

**First Stage:** For each tract, compute  $\hat{y}_{ijklmno}$  the maximum likelihood estimator (MLE) of the inclusion probabilities  $\{p_{ijklmno}\}$ , on the basis of  $\Omega$  and the model in (2).

**Second Stage:** For each configuration of reported and unreported items in the tract, compute  $\Psi_{ij*lmno}^k$  the MLE of the conditional inclusion probabilities. Then simulate a multinomial distribution based on  $\{\Psi_{ij*lmno}^k\}$  to impute the unreported items.

14. The first stage is carried through with the EM algorithm, which allows us to evaluate the MLE, based on  $\Omega$ . We illustrate the second stage in the situation where only origin is sometimes unreported. The strategy fully covers the general situation where any item may be unreported. To carry the second stage, we compute  $\{\Psi_{ij*lmno}^k\}$  the MLE of the conditional inclusion probabilities, from  $\hat{y}_{ijklmno}$  the MLE of the inclusion probabilities, according to the following formula.

$$\Psi_{ij*lmno}^k = \frac{y_{ijklmno}}{y_{ij1lmno} + y_{ij2lmno}} ; \quad k = 1, 2 \quad (3)$$

We impute Hispanic origin for each of the  $O_{ij*lmno}$  households by flipping a coin with probability  $\Psi_{ij*lmno}^1$  of a tail landing, and probability  $\Psi_{ij*lmno}^2$  of a head landing. If the actual landing is a tail, we substitute non-Hispanic for origin of the household, if the landing is a head, we substitute Hispanic for origin. The next section presents evaluation tools. We use these tools to quantify the performance of our item imputation procedure in terms of estimation bias and mean squared error.

### III.3 The Constrained Dirichlet Conjugate Family

15. In order to possess the necessary statistical tools for evaluating the performance of a model-based item imputation, we establish the procedure in the context of the bayesian framework. We identify a conjugate family of prior and posterior distributions that allows us to derive the distribution of the population counts induced by the model. The family of constrained dirichlet distributions is a natural conjugate family when the likelihood function is multinomial and constrained by a log-linear model. We recall the general form of the probability density function corresponding to an arbitrary member of the constrained dirichlet family. The density defines a probability distribution on the inclusion probabilities  $\{p_{ijklmno}\}$ , subject to the constraints given in (2).

$$f(\{p_{ijklmno}\} | \{a_{ijklmno}\}) = C(\{a_{ijklmno}\}) \prod_{ijklmno} (p_{ijklmno})^{a_{ijklmno}} \quad (4)$$

$C(\{a_{ijklmno}\})$  is the integration constant, and  $\{a_{ijklmno}\}$  are the characteristic hyperparameters for a member of the constrained dirichlet family. This family of densities is convenient because, if the prior density is in this family, then, after observing  $\{N_{ijklmno}\}$ , the posterior density is also in this family. The posterior density is defined by replacing  $a_{ijklmno}$  with  $a_{ijklmno} + N_{ijklmno}$  in (4).

16. In our situation, since some items are unreported, the population counts  $\{N_{ijklmno}\}$  are not available. However, we can produce a Markov chain Monte-Carlo (MCMC) to approximate the posterior distribution conditional on the reported data  $\Omega$ . Based on the model in (2) we can approximate the posterior distribution of  $\{p_{ijklmno}\}$  conditional on  $\Omega$ . Likewise, we can approximate the predictive distribution of  $\{N_{ijklmno}\}$ , conditional on  $\Omega$ . In the bayesian framework, the predictive distribution expresses our knowledge on  $\{N_{ijklmno}\}$ , after observing the data  $\Omega$ . It provides the tools to quantify the error involved in approximating  $\{N_{ijklmno}\}$  with a set of counts derived from an imputation procedure. In section IV we give results for an MCMC approximating the predictive distribution of the population counts for the 1998 dress rehearsal.

## IV. RESULTS FOR MODEL-BASED IMPUTATION AND THE SEQUENTIAL HOT DECK

17. In this section we present the item imputation results for the 1998 dress rehearsal for Census 2000, along with the results for two implementations of the procedure in subsection III.2, using the model (2). In addition, we give the approximations of two predictive distributions corresponding to the two implementations. We approximated each predictive distribution with 1800 cycles of BIPF. The BIPF algorithm simulates the constrained dirichlet distribution in (4) by successively generating unconstrained dirichlet distributions corresponding to joint marginal distributions of the distribution in (4). We refer to Schafer (1997) and Gelman et Al. (1995) for detailed descriptions of the BIPF algorithm.

### IV.1 Adjustment for the Dictionary of Hispanic Names

18. Our goal is to identify and evaluate the relative biases between our model-based imputation methodology and the SHD imputation methodology, in terms of population counts. To do so, we must control for relative biases between the implementations of the two types of methodologies that do not reflect the differences in imputation techniques. In 1998 the U.S. Census Bureau developed a dictionary of last names to determine origin when unreported. The dictionary is site based. In Sacramento, out of 3934 households with unreported origin, 2604 had origin imputed through the dictionary, meaning their last names had strong Hispanic, or strong non-Hispanic, affiliations. The remaining 1330 households had their origin imputed with the SHD. It is important for our analysis to adjust for the relative biases that are artifacts of the implementation of the dictionary. To achieve this, we computed two versions of the posterior and predictive distributions, corresponding to the two versions of our imputation procedure.

19. In the first version we ignore the imputations obtained through the dictionary and we assume that no information is available on the origin of the 3934 households who did not report it. Table 1 gives the predictive mean and predictive standard deviation corresponding to the predictive distribution of the population counts for 28 demographic categories under this assumption. For the second version, we respect the imputation of origin based on the dictionary. In this case, we assume that only the 1330 households who were not assigned origin through the dictionary did not report it. Table 1 gives the predictive means and standard deviations under this assumption. Table 1 also gives population counts obtained through two runs of our model-based imputation procedure, the first run ignoring the last name dictionary, and the second run respecting it. We give the results along with the reported counts for the 1998 dress rehearsal. These counts were obtained with the SHD and the dictionary.

20. In the next subsection, we identify the cases where the relative bias between model-based imputation and the SHD can be explained by the use of the last name dictionary. We then examine a situation where the relative bias can not be attributed to the methodological differences between the model-based procedure and the SHD.

## **IV.2 Imputation of Hispanic Origin with the Hispanic Last Name Dictionary**

21. We can evaluate the effect of the dictionary by computing the distance between the two predictive means for each category in Table 1. If the distance is large, more than two standard deviations, then we consider that the dictionary had an impact. The difference between the predictive means of the count of Hispanics, with and without the dictionary (21038.66 vs. 21144.86), is large ( $S.D. = 17.25$ ), thus the dictionary has a serious impact. Moreover, the second version of the predictive mean is close to the count of Hispanics for the dress rehearsal (21024), in terms of standard deviation (10.33). Therefore, the count of Hispanics reported for the dress rehearsal is consistent with the predictive mean if we accept the results from the dictionary. We observe that, after adjusting for the effect of the dictionary of last names, the counts of the dress rehearsal are consistent with the predictive distribution for most demographic categories involving specific values of origin. In general, Table 1 suggests that the dictionary may be deflecting a bias confounded with the value of the origin item. It appears that non-Hispanic households do not report origin in part because they are not Hispanic. Additional investigation is needed to confirm this hypothesis, but if this hypothesis holds, the dictionary methodology is an important break-through. We propose a functional model integrating the dictionary, in section V. In the next subsection, we look at categories where the results of the dress rehearsal are inconsistent with the predictive mean, even after adjusting for the effect of the dictionary.

## **IV.3 Imputation of Tenure for Black Households**

22. Table 1 reveals a serious discrepancy between the value of the predictive mean (7538.7) for the population count of “Black Owners”, and the count produced with the SHD (7661) in terms of the predictive standard deviation (20.2). In an attempt to explain this, we compare the rate of ownership for “Blacks” with the rate of ownership for their neighbor. The neighbor can be of any race. The rationale for this comparison is based on the principle behind the SHD: when tenure is unreported, the SHD borrows it from the neighbor. This technique is at the root of the discrepancy.

23. In Table 2 we observe that, according to the SHD, the rate of ownership for “Black Households” with imputed tenure (.436) is larger than the rate of ownership for the “Black Households” with reported tenure (.379). At the same time, the rate of ownership for “Blacks” with imputed tenure is close to that of their neighbors (.419). We conclude that the SHD reproduces the tenure of the neighbor, without adjusting for race. On the other hand, the predictive mean of the rate of ownership for “Blacks” with imputed tenure (.375) is in agreement with the rate of ownership for “Blacks” with reported tenure. The inconsistency between the rate of ownership for “Blacks” from the SHD and the predictive mean rate is serious. The SHD gave a rate of ownership more than five standard deviations (.0104) above the predictive mean rate for the “Blacks” with imputed tenure.

**Table 1. Sequential Hot Deck vs. Model-based Imputation and Two Predictive Distributions**

Population Count	SHD Imputation with Dictionary	Model-Based Imputation With Dictionary	Model-Based Imputation	Predictive Mean with Dictionary $E[N^D   \Omega]$	Predictive S. D. with Dictionary $\sqrt{V[N^D   \Omega]}$	Predictive Mean $E[N   \Omega]$	Predictive S. D. $\sqrt{V[N   \Omega]}$
All	138271	138271	138271	138271	0	138271	0
White	89032	88890	88876	88927.1	36.2	88930.3	35.4
Black	19962	19937	19979	19957.0	15.6	19953.9	16.1
Asian	17405	17423	17439	17426.6	14.8	17427.1	15.2
Other	11872	12021	11977	11960.3	35.2	11959.8	32.8
Non-His.	117247	117221	117127	117232.3	10.3	117126.1	17.2
Hispanic	21024	21050	21144	21038.7	10.3	21144.9	17.2
White N.-H.	79964	79936	79854	79934.8	15.6	79864.6	18.9
White His.	9068	8954	9022	8989.3	34.8	9065.6	35.8
Black N.-H.	19357	19328	19344	19345.1	10.6	19335.0	11.8
Black His.	605	609	635	611.9	12.5	618.9	14.1
Asian N.-H.	16887	16911	16901	16909.8	10.1	16895.4	11.1
Asian His.	518	512	538	516.8	11.5	531.6	12.6
Other N.-H.	1039	1046	1028	1042.7	3.6	1031.1	3.9
Other His.	10833	10975	10949	10917.6	34.3	10928.7	32.9
Owner	70054	70064	70056	70021.7	42.6	70022.8	43.1
Renter	68217	68207	68215	68249.3	42.6	68248.2	43.1
White Own.	47722	47776	47778	47770.8	40.6	47776.1	41.0
White Rent.	41310	41114	41098	41156.3	40.9	41154.1	42.1
Black Own.	7661	7576	7543	7540.7	20.9	7538.7	20.2
Black Rent.	12301	12361	12436	12416.3	21.8	12415.2	22.3
Asian Own.	9810	9848	9880	9874.1	19.0	9874.9	19.0
Asian Rent.	7595	7575	7559	7552.5	18.1	7552.2	19.0
Other Own.	4861	4864	4855	4836.2	26.7	4833.1	26.6
Other Rent.	7011	7157	7122	7124.1	28.8	7126.7	29.1
N.-H. Own.	60645	60676	60576	60615.4	38.7	60559.4	40.0
N.-H. Rent.	56602	56545	56551	56617.0	39.0	56566.7	40.2
His. Owner	9409	9388	9480	9406.3	20.3	9463.4	23.0
His. Renter	11615	11662	11664	11632.3	20.7	11681.4	23.3

24. This is an example of a situation where the SHD does not adjust for interactions between the household items, race and tenure in this case. Kovar and Whitridge (1995) caution against using the SHD when the sort order of the files generates a systematic bias. This seems to be what's happening here. Sorting on geography only does introduce a bias. A different sort order, or the introduction of additional class variables (race), could possibly solve the problem. Of course, the sort order is irrelevant with the model-based methodology. The model-based imputations are automatically adjusted to reflect the interaction between race and tenure.

**Table 2 - Rates of Ownership for Black Households in Sacramento**

Household Type	Number of Black Households	Ownership of the Neighbors	Ownership for the Households with Reported Tenure	Ownership for the Households with Imputed Tenure (SHD)	Predictive Mean Rate of Ownership	Predictive S.D. of the Rate of Ownership
Tenure Reported	18176	.428	.379	N /A	.378	.000345
Tenure Unreported	1786	.419	N /A	.436	.375	.0104

## V. FUTURE RESEARCH AND CONCLUSION

25. The model-based approach is very rich and we feel that we have barely scratched the surface. In the context of the decennial Census, there remain several avenues of investigation that appear promising. For instance, it is evident to us that the imputation of origin aided by the dictionary of Hispanic last names is worth investigating. What we have in mind is a slightly more elaborate version of the model given in (2). We can add terms on the RHS of (2). Let  $d_p$  be the (main) discrimination effect of the dictionary on the population, and let  $(H * d)_{k,p}$  be the interaction effect between the origin of the last name and the origin of the household. We have  $p = 1$ , if the last name is not Hispanic,  $p = 2$  if the name is Hispanic, and  $p = 3$  if the origin of the name cannot be determined.  $k$  denotes origin as before. The probabilistic nature of the model accounts for cases where the origin of the last name does not correspond to the origin of the household.

26. The most important feature of our procedure is the feedback through the predictive distribution. The predictive distribution provides a rigorous yardstick to evaluate to what extent some assumptions may be erroneous. We feel that, when the primary goal of the methodologist is to preserve multivariate dependencies, the model-based procedure is a more natural imputation device than the SHD. It may be that the process of designing a model is more conducive to an elicitation of the multivariate dependencies than writing the specifications for the SHD. It is our hope that the advantages of the model-based approach will motivate practitioners to use and implement models for imputation purposes.

## References

- Fay, R. E., and Town, M. K. (1998). "Variance Estimation for the 1998 Census Dress Rehearsal," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman & Hall.
- Kovar, J. G., and Whitridge, P. J. (1995). "Imputation of Business Survey Data," *Business Survey methods*, Cox, Binder, Chinnappa, Christianson, Colledge, Kott Ed., Wiley.

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

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 Census Bureau.

Williams, T. R. (1998). "Imputing Person Age for the 2000 Census short Form: A Model Based Approach", *Proceedings of the Section on Survey Research Methods*, American Statistical Association.