Potential Methodologies for Count Imputation for the Decennial Census

Richard A. Griffin, U.S. Census Bureau

Key Words: EM Algorithm, Multinomial Distribution, Log Linear Model, Spatial Models

1. Introduction[1]

Count imputation was used for Census 2000 for housing unit records lacking a status designation of occupied, vacant or nonexistent as well as for known occupied units with unknown population count. A "hot-deck" imputation methodology was used to determine donors to be used for donees requiring imputation.

Count imputation as implemented for Census 2000 was a deterministic method in that given the census data the imputed values are fixed. Alternative stochastic imputation methods which randomly select imputed values from a distribution could also be used. Since the results of Count Imputation effect many important uses of the Census such as allocation of congressional seats and revenue distribution, the Census Bureau is conducting research on imputation alternatives to the Census 2000 methodology. This paper presents results from simulations of alternative stochastic imputation methodologies using log linear models on Census 2000 data. These methodologies assume a multinomial distribution and take advantage of the monotone missing data pattern to produce explicit maximum likelihood estimates by the factored likelihood method for some log linear models. For other log linear models the EM Algorithm is utilized. Neighbor characteristics are used resulting in a spatial modeling application.

Section 2 provides an overview of the Census 2000 count imputation methodology. Section 3 documents a potential alternative count imputation methodology that involves the EM Algorithm, log linear models, and a simple spatial modeling application. Examples of such models have been simulated using Census 2000 data for the state of Delaware. Section 4 compares the results of these simulations with the Census 2000 count imputation results. The paper closes with a general summary and discussion of future research.

[1]This paper reports the results of research and analysis undertaken by Census bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

2. Overview of Census 2000 Count Imputation

2.1 Background

The purpose of this section is to provide an overview of how we implemented count imputation in Census 2000. This procedure is referred to as count imputation since we were concerned with only missing housing unit status and missing population count for occupied housing units. These missing data affect the population total. Other missing data such as missing demographic data were handled during the Characteristic Imputation procedure which occurred after Count Imputation was completed. Only the Count Imputation was required for the population numbers due December 31, 2000.

After many stages of census operations, such as mail list development, update/leave, and list/enumerate operations, postal verification check, new construction program, and other late adds operations, a final list of housing units existing on census day was established. At the end of follow-up activities and data capture processing, some census occupied housing unit records did not contain information on the number of persons or did not contain information on whether the census housing unit was occupied, vacant, or delete. The source of these omissions may have been from respondents not providing correct or timely information or from any unanticipated operational obstacle.

The Count Imputation was designed to fill in missing housing unit status and the number of persons for any occupied census housing unit without household size. The operation was done concurrently as part of the creation of the Hundred percent Census Unedited File (HCUF) on a flow basis by Local Census Office (LCO). Count Imputation was the last process to complete the HCUF. Other HCUF creation activities such as merging the Decennial Master Address File with the Decennial Response File for establishing census housing unit records were completed prior to the Count Imputation operation.

A subsequent operation, substitution, was used to

impute the 100% data (age, sex, race, Hispanic Origin, relationship) for units with a population count greater than zero imputed by Count Imputation. Substitution was also necessary for occupied housing units with a known population if all the person data were missing. The substitution edit replicated the person records from nearby fully enumerated households of the same size in their entirety. The selection of nearby households was independent of the selection of a donor household in Count Imputation. Thus the household used to impute the population count in Count Imputation could be different than the household used to impute the 100% data in substitution; although both were required to have the same population size.

2.2 Methodology

Under the assumption that housing unit status and number of persons living in a housing unit are more similar in a nearby neighborhood than a far away community, the nearest-neighbor hot deck method was used. This means that the data from the closest available neighbor was used to fill in the missing data. Geographical closeness of housing units was determined by sorting all housing units and group quarters within a tract by block number, street name, and house number. Based on the sorted sequence, backward and forward searches were conducted to find a donor for a unit with missing data. The unit with missing data was known as a donee. The nearest available unit meeting specified requirements (see Table below) in either direction was used as a donor to fill in the data for the donee. The donee took the donor's housing unit status and population size as its own.

The Count Imputation consisted of three distinct processes defined as:

- Household Size Imputation - The Census Bureau imputed a population count for a housing unit when Census Bureau records indicated that the housing unit was occupied, but did not show the number of individuals residing in the unit.

- Occupancy Imputation - When Census Bureau records indicated that a housing unit existed but not whether it was occupied or vacant, the Bureau imputed occupancy status (occupied or vacant), and then, if the unit was imputed to be occupied, the household size of the donor record was used.

- Status Imputation - When the Census Bureau's records had conflicting or insufficient information about whether an address represented a valid, non-duplicated housing unit, the Bureau first imputed for the status of the unit (occupied, vacant, delete), then, if occupied, the household size of the donor record was used.

The three types of imputation categories were subdivided into single housing units and multi-units to form six estimation categories. The nearest-neighbor hot deck was done separately for single units and multi-units within each of these three imputation categories. One restriction was that each donor could only be used once as a nearest donor for Count Imputation. Household Size Imputation was done first, followed by Occupancy Imputation and then Status Imputation.

3. Alternative Count Imputation Methodology

The methodology described in this example has been formulated using Little and Rubin (1987). We want to initially use three dimensional log linear models since the theory is much easier to understand than for higher dimensional models and there are many more possible models as the number of variables increases. Note that for four dimensional tables there are 113 different hierarchical loglinear models all of which include the main effect terms. This number grows extremely rapidly, for 10 variables the number is 3,475,978. As a practical matter we may only want to consider two way interactions no matter how many variables are in the model.

Consider a census tract for which Table 4 is a cross-classification of households that do not require count imputation by Neighbor Population (N; vacant, 1, 2, 3, 4, 5, or 6+ persons, or nonexistent housing unit), Structure (S: Single unit address or multi-unit address), and Housing Unit Population (P; same categories as for N). A nonexistent housing unit is an address listing which has been deleted from the files by census operations such as a business address or demolished structure. Let x_{jkl} be the count of housing units that do not require count imputation in the jth category of P, kth category of N and lth category of S. This notation is demonstrated in a few of the cells of Table 4. All cells will have a count which could sometimes be 0.

Table 2 is a cross-classification of household that do require count imputation by S and N. Let r_{kl} be the count of housing units that require count imputation in the kth category of N and the lth category of S.

Thus we have the simplest form of a monotone missing data pattern. Complete response is obtained for variables S and N, and response to P is only obtained for a subset of the respondents to S and N. For some loglinear models, a monotone pattern allows us to obtain maximum likelihood estimates of the cell probabilities by using a factorization of the likelihood. For other loglinear models we will use the Expectation-Maximization (EM) algorithm.

Since all housing units have a complete response for N and S, Table 3 is a cross-classification of all households by S and N. Let s_{kl} be the count of all housing units in the kth category of N and the lth category of S.

For a 3-way contingency table with cell probabilities π_{jkl} we consider three loglinear models. Only models which include an interaction term for P and N as well as for P and S are considered. We would not use N or S for our modeling to predict P if they did not have a significant interaction with P.

First we have the saturated model denoted {PNS}

$$\ln \pi_{jkl} = \alpha + \alpha_j^{(P)} + \alpha_k^{(N)} + \alpha_l^{(S)} + \alpha_{jk}^{(PN)} + \alpha_{jl}^{(PS)} + \alpha_{kl}^{(NS)} + \alpha_{jkl}^{(PNS)} \quad (1)$$

The terms $\alpha_j^{(P)}$, $\alpha_k^{(N)}$, $\alpha_l^{(S)}$ are called the main effects of P, N, and S respectively.

The terms $\alpha_{jk}^{(PN)}$, $\alpha_{jl}^{(PS)}$, $\alpha_{kl}^{(NS)}$ are called two-way associations between P and N, P and S, and N and S, respectively.

The term $\alpha_{jkl}^{(PNS)}$ is called the three-way association between P, N, and S.

The second model is the Partial Association Model denoted {PN, PS, NS}

$$\ln \pi_{jkl} = \alpha + \alpha_j^{(P)} + \alpha_k^{(N)} + \alpha_l^{(S)} + \alpha_{jk}^{(PN)} + \alpha_{jl}^{(PS)} + \alpha_{kl}^{(NS)} \quad (2)$$

the third model is the Conditional Independence Model denoted {PN, PS} for which at each level of P, S and N are independent.

$$\ln \pi_{jkl} = \alpha + \alpha_j^{(P)} + \alpha_k^{(N)} + \alpha_l^{(S)} + \alpha_{jk}^{(PN)} + \alpha_{jl}^{(PS)} \quad (3)$$

Note that since for the fully classified housing units, each unit is also someone's neighbor it is unlikely that there would be a PS interaction and no NS interaction. However, this third model is included for illustration.

The likelihood for the combined data factors into a term for the distribution of NS involving all cases and a term for the distribution of P given NS involving only the completely classified cases. These two distributions involve distinct parameters for the models {PNS} and {PN, PS, NS}. These models do not need the EM Algorithm, the factored likelihood method can be used, although the maximum likelihood estimation for the distribution of P given NS requires an iterative procedure such as raking for model {PN, PS, NS} . The model {PN, PS} requires the EM algorithm.

3.1 Maximum Likelihood Estimates (MLE) for Loglinear Models

x_{jkl} denotes the observation in Table 4 for P=j, N=k, and S=l. For example, x_{342} is the observed count for P=2 persons (the third P value), N=3 persons (the fourth N value), and S=multi (the second S value).For model {PNS}, the MLEs are the observed x_{jkl} values The MLEs estimates for model {PN, PS, NS} are obtained by a 3 dimensional raking using the three sets of 2 variable marginal controls (x_{jk+} , x_{j+l} , x_{+kl}). This is the only three variable model that does not have explicit MLEs.

For model {PN, PS} at each level of P, N and S are independent. The MLEs are as follows:

$$\hat{m}_{jkl} = \frac{x_{jk+}\, x_{j+l}}{x_{j++}}$$

3.2 Models that can use Factored Likelihood

For these models the following equation is used:

$$Pr(P=j,N=k,S=l)=Pr(N=k \wedge S=l)Pr(P=j|N=k,S=l) \quad (4)$$

In all cases $Pr(N=k \wedge S=l)$ is obtained from Table 3. To compute $Pr(P=j \mid N=k, S=l)$ first do ML on Table 4 under the particular model to get the expected counts for the completely classified cases. Then $Pr(P=j \mid N=k, S=l)$ is computed using these counts. Next equation (4) is used to obtain the complete joint distribution. These probabilities are used to allocate the partially classified cases from Table 2 to the appropriate cell of the completely classified table. Thus the final totals in each cell of the table are obtained.

This will be illustrated for model {PNS}. For model {PN,PS,NS} the MLEs were computed using 2 iterations of the 3 dimensional raking.

For example $\hat{m}_{342}$ is the MLE for P = 2 persons (the 3rd value of j), N = 3 persons (the 4th value of k), and S = multi (the second value of l) using Table 2.

$\hat{m}_{342} = x_{342}$. Then

$Pr(P=2\ persons|N=3\ persons,S=multi) = \frac{\hat{m}_{342}}{\hat{m}_{.42}} = \frac{x_{342}}{x_{.42}}$

from Table 3 $Pr(N=3\ persons \wedge S=multi) = \frac{s_{42}}{s_{..}}$

and from equation (4)

$$Pr(P=3,N=4,S=2)=Pr(N=4 \wedge S=2)Pr(P=3|N=4,S=2) = (\frac{s_{42}}{s_{..}})(\frac{x_{342}}{x_{.42}}) = p_{342}$$

These probabilities are used for assigning the partially classified cases from Table 2 to cells of the complete classification.

For example looking at the cell with P = 2 persons, N = 3 persons, S = multi, we observed x_{342} cases from Table 4. There were r_{42} partially classified cases with N = 3 persons and S = multi (Table 2) so using the above probabilities the desired count for P = 2 persons, N = 3 persons, and j = multi is

$$x^1_{342} = x_{342} + r_{42} * \frac{p_{342}}{\sum_{j=1}^{8} p_{j42}}$$

The x^1_{jkl} counts make up the full table to be used for imputation.

3.3 Models that cannot use Factored Likelihood - EM Algorithm Necessary

Models that do not include a term for the distribution of NS cannot be estimated using the factored likelihood method. This makes sense since if a NS term is not in the model then $Pr(N=k \wedge S=l)$ is not to be estimated under that model. These models are estimated by first doing MLE under the model on the fully classified cases to obtain starting values. If any of these starting values are zero it is necessary to use a small positive number in that cell so that the E step of the EM Algorithm allocates missing data to that cell with a non-zero probability. Note that the SAS program that performs the EM Algorithm takes care of these by spreading the count of missing cases that could fall in a group of cells evenly over all possible cells. Using the starting values from the MLE under the model the E step and the M step are done iteratively to convergence. For this illustration one iterations is done for model {PN,PS}.

For example $\hat{m}_{441}$ is the MLE for P = 3 persons (the 4th value of j), N = 3 persons (the 4th value of k), and S = single (the first value of l) using Table 4 and model {PN,PS}.

$$\hat{m}_{441} = \frac{x_{44+}\ x_{4+1}}{x_{4++}}$$

For the first E step, we want to use the counts from MLE using Table 4 to allocate missing data cases to the appropriate cell. Let $p_{jkl} = \frac{\hat{m}_{jkl}}{\hat{m}_{...}}$ be the cell probabilities formed using these MLEs. The E step finds the conditional expectation of the missing data given the observed data and "current" or latest MLEs. For example looking at the cell with P = 3 persons, N = 3 persons, S = single, we observed x_{441} cases from Table 4. There were r_{41} partially classified cases from Table 2 with N = 3 persons and S = single so using the probabilities estimated by MLE the desired count for P = 3 persons, N = 3 persons, and S = single is

$$x^1_{441} = x_{441} + r_{41} * \frac{p_{441}}{\sum_{j=1}^{8} p_{j41}}$$

The first M step is to perform MLE under the {PN,NS} model on a table of x^1_{jkl} counts obtained by allocating the partially classified cases to the x_{jkl} counts from Table 4 in this manner.

For example $\hat{m}^1_{441}$ is the MLE for P = 3 persons (the 4th value of j), N = 3 persons (the 4th value of k), and S = single (the first value of l).

$$\hat{m}^1_{441} = \frac{x^1_{44+}\, x^1_{4+1}}{x^1_{4++}}$$

The $\hat{m}^1_{jkl}$ values make up the full table after this first M step. This completes the first iteration of the EM Algorithm. For the second iteration we repeat the process using this table in place of Table 4. This process is continued to convergence resulting in a Table of counts to be used for imputation.

4. Results

The methodology described in Section 3 was implemented for all three models on each of the census tracts in Delaware. For each model, the final table of counts were compared with the same counts obtained from the final Census 2000 files which include count imputation as described in section 2. Note that we did not actually perform imputation for the alternative methodology from section 3. Imputation would involve using the final estimated counts to obtained desired probabilities and imputing each case via an independent draw from the appropriate Bernoulli distribution. Using these final counts effectively assumes no use of the three imputation categories shown in Table 1 and uses the expected results of the Bernoulli imputation. It would be possible to use these imputation categories to adjust the imputation probabilities based on allowing only a subset of the counts. For example for Household size imputation only allow the imputation of an occupied unit.

The for each of the 169 census tracts in Delaware, for each of the three models for each potential final housing unit category (vacant, 1, 2, 3, 4, 5, 6+, or nonexistent) the difference between the census 2000 imputation and the alternative imputation count for that category was determined.
Results are given in Table 5.

5. Summary and Discussion

While these differences do not appear to be very large, their effect on data uses such as congressional apportionment is unknown. This simulation was performed to gain knowledge of how to use alternatives to the traditional hot deck imputation. The independent variables used here are examples of those that could be used. We plan on exploring many possible independent variables that may be good predictors of housing unit population count. We also will explore log linear models with more than 3 variables. Since determining the ML estimates can get very complicated for models with more than 3 variables we would plan on developing a general SAS program that uses the EM algorithm for any given model. The EM algorithm will converge to the MLE estimates. Thus, if the factored likelihood method can be used but instead the EM algorithm is used, you still get the MLE proportions to be used for imputation. For more than 3 dimensions we do not want to look at each possible model and determine if the EM algorithm is necessary.

Finally it is important to note that for this paper no assumption is made about the correct or true imputation. We do not know which, if any, of the counts examined are correct so we do not know which imputation methodology is best. For our future research we plan on creating a “truth deck” so that alternative imputation methodologies can be compared with a goal of determining which is best.

References

Little, R., and Rubin D., Statistical Analysis with Missing Data, John Wiley & Sons, 1987

TABLE 1
Summary of Count Imputation Categories:

Type of Imputation	Estimation Category	Donees	Donors
Household size	1a. Single Units 1b. Multi-units	Occupied units with unknown population count.	Occupied units with a population count from enumerator completed forms.
Occupancy	2a. Single Units 2b. Multi-units	Units known to exist (either occupied or vacant).	Occupied or vacant units from enumerator completed forms.
Status	3a. Single Units 3b. Multi-units	Units for which we knew nothing.	Occupied, vacant, or delete units from enumerator completed forms.

TABLE 2 PARTIALLY CLASSIFIED TABLE

	Neighbor Pop. (N)							
Structure (S)	Vac.	1	2	3	4	5	6+	NE
Single	r_{11}	r_{21}	r_{31}	r_{41}	r_{51}	r_{61}	r_{71}	r_{81}
Multi	r_{12}	r_{22}	r_{32}	r_{42}	r_{52}	r_{62}	r_{72}	r_{82}

TABLE 3 COMPLETE DATA FOR NxS

	Neighbor Pop. (N)							
Structure (S)	Vac.	1	2	3	4	5	6+	NE
Single	s_{11}	s_{21}	s_{31}	s_{41}	s_{51}	s_{61}	s_{71}	s_{81}
Multi	s_{12}	s_{22}	s_{32}	s_{42}	s_{52}	s_{62}	s_{72}	s_{82}

TABLE 4 COMPLETELY CLASSIFIED TABLE
Note

		Neighbor Pop. (N)							
Unit Pop. (P)	Structure (S)	Vac.	1	2	3	4	5	6+	NE
Vac.	Single	x_{111}	x_{121}						
	Multi	x_{112}	x_{122}						
1	Single								
	Multi								
2	Single								
	Multi								
3	Single								
	Multi								
4	Single								
	Multi								
5	Single								
	Multi								
6+	Single								
	Multi								
NE	Single							x_{871}	x_{881}
	Multi							x_{872}	x_{882}

Table 5 Count Difference Distribution for Census 2000 minus Alternative.
Over all 169 tracts, each cell has 1. Mean, 2. Standard Deviation, 3. Minimum, 4. Maximum

Category	Model {P,N,S}	Model {PN,PS,NS}	Model {PN,PS}
Vacant	1. 0.342 2. 4.103 3. -25.793 4. 31.248	1. 0.283 2. 3.973 3. -24.573 4. 31.927	1. 0.177 2. 4.061 3. -24.573 4. 32.007
1 person household	1. 0.156 2. 3.019 3. -11.004 4. 23.536	1. 0.201 2. 2.994 3. -11.693 4. 22.691	1. 0.184 2. 2.904 3. -10.252 4. 22.553
2 person household	1. -0.287 2. 2.449 3. -12.927 4. 6.565	1. -0.310 2. 2.483 3. -13.558 4. 7.273	1. -0.219 2. 2.549 3. -13.567 4. 8.225
3 person household	1. -0.123 2. 2.077 3. -17.089 4. 7.965	1. - 0.30 2. 1.603 3. -8.187 4. 8.078	1. 0.020 2. 1.558 3. -7.869 4. 8.165
4 person household	1. -0.006 2. 1.392 3. -5.699 4. 7.046	1. -0.0219 2. 1.446 3. -6.853 4. 7.075	1. 0.011 2. 1.408 3. -6.668 4. 7.077
5 person household	1. 0.067 2. 0.757 3. -3.463 4. 3.505	1. 0.045 2. 0.761 3. -3.429 4. 2.496	1. 0.058 2. 0.787 3. -3.435 4. 3.584
6+ person household	1. 0.080 2. 0.705 3. -3.282 4. 4.428	1. 0.069 2. 0.698 3. -3.728 4. 4.317	1. 0.073 2. 0.700 3. -3.667 4. 4.329
Nonexistent	1. -0.231 2. 2.332 3. -14.647 4. 12.403	1. -0.237 2. 2.485 3. -14.195 4. 15.552	1. -0.303 2. 2.335 3. -13.943 4. 10.872