GOVERNMENTS DIVISION REPORT SERIES
(Research Report #2009-4)

Evaluation of Alternative Imputation Methods for the Public Libraries Survey (PLS)

Irene Brown
Terri L. Craig

U.S. Census Bureau
Washington, DC 20233

CITATION: Brown, Irene, Terri L Craig. 2009. Evaluation of Alternative Imputation Methods for the Public Libraries Survey (PLS). Governments Division Report Series, Research Report #2009-4

Report Completed: September 11, 2009
Report Issued: October 2, 2009

Evaluation of Alternative Imputation Methods for the Public Libraries Survey (PLS)

Irene Brown[1], Terri L Craig[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
The Public Libraries Survey (PLS) imputes for missing data. Although the unit response rate is very high, there is still item nonresponse that requires imputation. PLS uses a variety of imputation methods depending on the item that is missing data. The purpose of this study is to evaluate current imputation methods along with some new alternatives. The imputation methods being evaluated are ratio, cell mean, adjusted cell mean, hot-deck, and combinations thereof. Another objective is to minimize the number of imputation methods being used. The study also compares current methods of determining the imputation cell boundaries to new methods using the population variable.

Keywords: Imputation, Cell boundaries

1. Background

The Public Libraries Survey (PLS) provides a national census of public libraries and their public service outlets. The U.S. Census Bureau collects the information for the Institute of Museum and Library Services (IMLS) through the Library Statistics Working Group. (http://harvester.census.gov/imls/publib.asp) Prior to survey year 2007, the survey was administered by the National Center for Education Statistics (NCES). PLS is designed as a universe survey, with approximately 9,100 libraries in the 50 states, the District of Columbia, and the outlying areas of Puerto Rico, Guam, the Northern Marianas, and the Virgin Islands. Unit response rates are high, generally around 98 percent, whereas response to some items on the questionnaire can be below 85 percent at the state level. Based on research done in 1997, PLS began imputing for missing data for libraries in the 50 states and the District of Columbia, but not the outlying areas. The imputed data are included in the tabulation of state and national totals (macro-data release). Some micro-data releases do not include imputed data and others do. When it is, the value is flagged as imputed.

The current imputation methods generally follow the recommendation of the research done in 1997 (Hilton, 1997). The definition of the imputation cells changed after the research was complete from stratification by state and population size to stratification by OBE (Office of Business Economics, now known as the Bureau of Economic Analysis) region code and size of population served. The number of population size cells varies within each OBE region code, along with the boundaries. The cell boundaries are found based on natural breaks in the distribution. An attempt is made to maintain an equal number of libraries within each imputation cell. Once the boundaries are set, if the unit response rate of an imputation cell is under 75 percent or the imputation cell contains fewer than 15 libraries, that cell is combined with an adjacent cell. The imputation

method used depends on the item being imputed. There are approximately eight methods currently being used. The method most used is a growth rate applied to prior year data. If prior year data are not available an adjusted cell mean is generally used. The use of a sum of details set equal to total is applied when appropriate. A few items use a ratio imputation model, when a highly correlated item is available. Where other imputation methods used are cell mean, cell median, direct substitution of prior year data, and hot-deck procedures.

The first part of the study evaluated the defining of imputation cells within the OBE regions. The second part evaluates the current imputation methods along with alternative methods with the purpose of reducing the number of imputation methods being used. We used survey year 2006 data along with prior year data going back to survey year 2002, except for new items introduced in 2006, which used survey year 2007 and 2006 data.

2. Current and Alternative Methods

2.1 Defining of imputation cell boundaries

Currently, PLS determines imputation cells by OBE region code and size of population served. This study compared three methods to define cell boundaries for the size of population served. They are the current method, the cumulative root frequency method proposed by Dalenius and Hodges (1959), and a geometric method proposed by Gunning and Horgan (2004).

First Method: The current method generally applies small annual judgmental tweaks to the cell boundaries determined in 1997. The "tweaks" occur mostly when a cell unit response rate is less than 75 percent. The number of population size cells within the OBE region generally stays the same. We used the imputation cell assignments from survey year 2006.

Second Method: The Dalenius and Hodges cumulative root frequency method (Dalenius and Hodges, 1959) defines boundaries of a continuous variable for a number of cells (L) as follows:

1. Group the continuous variable into K classes, the ranges being equal within the classes;
2. Determine the frequency in each class, f_i (i=1, 2, …K);
3. Calculate the square root of the frequencies in each class, $\sqrt{f_i}$;
4. Cumulate the square root of the frequencies, $cumf = \sum_{i=1}^{K} \sqrt{f_i}$;
5. Divide the sum of the square root by the number of strata/cells, $Q = cumf / L$
6. Take the upper boundaries of each stratum/cell to be the values of the population size variable corresponding to *Q, 2Q, … (L-1) Q, LQ.*

Third Method: The geometric method (Gunning and Horgan, 2004) defines boundaries for L cells as follows:

1. Find the minimum and maximum values of the variable;

2. Calculate the common ratio, $r = (\max/\min)^{1/L}$;
3. Take the boundaries of each stratum to be the values corresponding to the terms in the geometric progression: $a, ar, ar^2, ..., ar^L$, where a = the minimum.

The geometric method is designed to obtain homogenous cells for positively skewed populations. The size of population served variable along with many of the items collected on the PLS questionnaire are positively skewed.

2.2 Imputation Methods

In survey year 2008, a total of 51 items will be subjected to imputation. PLS currently employs approximately eight imputation methods, of which not all were researched in 1997. One goal for this study is to reduce the number of methods being used. The study evaluated nine imputation methods, which are defined below, for each item. The ninth method is the only one not currently being used by PLS. The study also expanded current methods that use prior year data from one or two years prior to include going back three or four years for prior data, i.e., data from time period *t*-2, *t*-3, or *t*-4 may be substituted for data from time period *t*-1 as necessary.

Method 1: Prior year data with a cell mean growth rate:

$$y_{h,i,t} = y_{i,t-1} * \left(\frac{\sum_{j=1}^{n_{hr}} y_{h,j,t} \Big/ y_{h,j,t-1}}{n_{hR}} \right)$$

The ratio is the mean growth rate of respondents in an imputation cell *h*, n_{hR} denotes total number of respondents R in a cell, i represents the i^{th} nonrespondent in imputation cell *h*, *j* represents the j^{th} respondent in imputation cell *h*, and *t* is the current year (note that other prior period data for the same unit *i* may by substituted for data from time period *t* -1 as necessary). This imputation method assumes the underlying model:

$$y_{h,i} = \beta y_{h,i,t-1} + \varepsilon_{h,j} \,.\; \varepsilon_{h,i} \sim N(0, \sigma^2 y_{h,i,t-1}^2)$$

Under this model, the $\hat{\beta}$ for the imputation cell given above is a B.L.U.E..

Method 2: Prior year data with a hot-deck growth rate (The hot-deck procedure uses the growth rate of a respondent that is next in order to the nonrespondent when ordered by size of population served within the same OBE region.):

$$y_{h,i,t} = y_{h,i,t-1} * \left(\frac{y_{h,j,t}}{y_{h,j,t-1}} \right)$$

where *h* represents an imputation cell, *j* is the next respondent after the i^{th} nonrespondent in the imputation cell, when ordered by population of legal service area, *t*-1 is the prior year (*t*-2, *t*-3, and *t*-4 data may be used if *t*-1 is not available).

Method 3: An adjusted cell mean:

$$y_{h,i} = \bar{y}_{h,R} * \frac{x_{h,i}}{\bar{x}_h}$$

where *h* represents an imputation cell, *R* denotes the group of respondents in an imputation cell, *x* is population of legal service area (available for all units in the survey).

Method 4: The cell mean:

$$y_{h,i} = \bar{y}_{h,R}$$

where *h* represents an imputation cell, *R* denotes the group of respondents in a cell.

Method 5: Prior year ratio to another item (i.e., historic imputation or auxiliary trend imputation):

$$y_{i,t} = x_{i,t} * \frac{y_{i,t-1}}{x_{i,t-1}}$$

where *x* is an item that is highly correlated with y, *t* is the current year, *t*-1 is the prior year (*t*-2, *t*-3, and *t*-4 data may be used if *t*-1 is not available). This imputation method requires available data for item *x* from current and prior time periods and available data for item *y* from the prior time period.

Method 6: Cell median ratio with another item:

$$y_{h,i} = x_{h,i} * \left(\frac{y}{x}\right)_{h,median,R}$$

where *h* represents an imputation cell, *x* is an item that is highly correlated with *y*, the median ratio is calculated from all respondents R in the cell.

Method 7: Prior year data with no growth rate (direct substitution):

$$y_{h,i,t} = y_{h,i,t-1}$$

where *h* represents an imputation cell, *t* is the current year, *t*-1 is the prior year (*t*-2, *t*-3, and *t*-4 data may be used if *t*-1 is not available).

Method 8: Cell median:

$$y_{h,i} = y_{h,Median,R}$$

where *Median* is calculated from all respondents R in a cell.

Method 9: Sequential hot-deck (The hot-deck procedure uses the data of a respondent that is next in order to the nonrespondent when ordered by size of population served within the same OBE region.):

$$y_{h,i} = y_{h,j},$$

where *h* represents an imputation cell, *j* is the next respondent in imputation cell *h* when ordered by population of legal service area.

3. Empirical Analysis

3.1 Imputation cell boundaries

PLS prefers imputation cells to be homogenous. To make valid comparisons between the methods of defining cell boundaries for the size of population served, the same number of cells within each OBE region as found in survey year 2006 were used. The method with the lowest stratified coefficient of variation (CV) of the studied population would, in theory, produce more homogenous imputation cells then the other methods. For each imputation cell development method, the CV for population characteristic Y (population of legal service area) was calculated as:

$$CV_{Method}(\bar{Y}) = \frac{\sqrt{\sum_{h=1}^{L}\left(N_h^2 / N^2\right) V_h(\bar{Y}_h)}}{\bar{Y}}$$

where *h* is an imputation cell, *L* is the number of imputation cells, and N and N_h are the total population and cell population sizes, respectively. The imputation cell mean is calculated as:

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} Y_i}{N_h}.$$

The variance within an imputation cell is calculated as:

$$V_h(\bar{Y}_h) = \frac{\sum_{i=1}^{N_h} (Y_i - \bar{Y}_h)^2}{N_h - 1}$$

The overall mean is calculated as:

$$\bar{Y} = \frac{\sum_{i}^{N} Y_i}{N}.$$

Imputation cells must also contain enough respondents to adequately impute for the nonrespondents in a cell. PLS requires imputation cells to have a unit response rate of at least 75 percent and at least 15 respondent libraries. If not, the cell is collapsed with adjacent cells till the requirements are met. These requirements were followed in evaluating the three imputation cell boundary methods.

3.2 Imputation Methods

The selection of the best imputation method for each item depends on the uses of the data. PLS uses data that has been imputed at the individual level to obtain state and national level totals (macro-data release). When the goal is to produce estimates of totals, the “best” imputation method minimizes the difference between the true total and a total with imputed data. However, there are also concerns about individual imputations

(micro-data release). To address this, the "best" method would minimize the difference between the individual imputed values and actual reported values. Although there is an attempt to balance both aspects of data quality, macro-data considerations outweigh micro-data for imputed values.

To evaluate the nine imputation methods, we first found the missingness pattern for each item. From a file of complete respondents, we then simulated missing data based on the item's missingness pattern. This process was repeated independently 500 times. To mimic the production process that will occur in survey year 2008, we defined the imputation cells by OBE region and size of population served using the method found best from those described in Section 3.1. For each of the missing values, an imputed value for each of the nine methods described in Section 2.2 was created. A national level total was calculated from the file of complete respondents for each item (the true total). For each of the 500 simulations and nine imputation methods, a national level total that included the imputed values was also calculated.

To look at the difference between the true total and the simulated totals with imputed data, average bias (AB) for each imputation method was calculated as

$$AB = \frac{\sum_{s=1}^{500} (\hat{\theta}_s - \theta)}{500}$$

and mean square error (MSE) as

$$MSE = \frac{\sum_{s=1}^{500} (\hat{\theta}_s - \theta)^2}{500}$$

where θ is the "true" total from the data of complete responses and $\hat{\theta}_s$ is the total with imputed data from the s^{th} simulation.

When macro-data release is the primary consideration, the imputation method with an average bias closest to zero and lowest mean square error (MSE) would be considered the "best". The average bias will contain both negative and positive differences that could cancel. So it is possible to have an unbiased method that yields a high MSE. In this case, we prefer the method with the lowest MSE.

To evaluate the difference between actual values and imputed values, the mean absolute error (MAE), as defined by Nordholt (1998) was calculated as

$$MAE = \frac{\sum_{i=1}^{n} | \hat{y}_i - y_i |}{n}$$

and the mean squared deviation (MSD) as

$$MSD = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}$$

where $\hat{y}_i$ is the imputed value for an observation, y_i is the reported value for an observation, and n is the number of imputed observations.

When micro-data preservation is the primary consideration, the imputation method with mean imputation error closest to zero and smallest mean squared deviation is the "best". When the macro-level (average bias and MSE) "best" method conflicts with the micro-level (mean absolute error and mean squared deviation), we selected the "best" macro-level method (i.e. lowest MSE).

4. Results

4.1 Imputation Cell Boundaries

We obtained imputation cells by OBE region and size of population served using the three methods described in Section 2.1. These initial results showed that the geometric method had the lowest CV (see Table 1).

Table 1: CV Associated with Imputation Cell Boundary Method

			Geometric Method
CV	890.33	709.13	170.22

However, for the geometric method, there were a number of the size cells within each OBE region that did not meet the criteria of 15 libraries and 75 percent unit response rate. The current and cumulative root frequency methods had all cells meet the criteria. (See Table 2 for an example of two OBE regions)

Table 2: Number of Libraries and Unit Response Rates by Imputation Cells

						Geometric Method	
							Response Rate
01	1	31	96.77%	31	96.77%	1	100.00%
01	2	46	100.00%	38	100.00%	30	96.67%
01	3	81	96.30%	54	96.30%	135	97.78%
01	4	53	100.00%	88	98.86%	299	97.66%
01	5	84	96.43%	101	97.03%	358	95.53%
01	6	100	98.00%	128	97.66%	328	93.90%
01	7	93	95.70%	201	96.52%	124	83.87%
01	8	97	98.97%	293	95.22%	15	86.67%
01	9	713	91.87%	364	88.74%	8	62.50%
04	1	36	100.00%	24	100.00%	10	100.00%
04	2	36	100.00%	25	100.00%	22	100.00%
04	3	44	100.00%	27	100.00%	37	100.00%
04	4	134	100.00%	43	100.00%	100	100.00%
04	5	74	98.65%	60	100.00%	199	99.50%
04	6	80	97.50%	98	98.98%	296	97.30%
04	7	217	97.70%	125	98.40%	496	95.56%
04	8	165	97.58%	152	98.03%	346	88.15%
04	9	165	96.36%	487	97.13%	113	65.49%
04	10	679	85.57%	589	83.70%	11	54.55%

To follow PLS procedure, we examined combining adjacent cells to meet the cell size and response rate criteria. We first increased the numbers of cells within the OBE regions and then applied the geometric method. The resulting cells were collapsed to achieve at least 15 libraries and 75 percent unit response rate. To compare to the other methods, we used the number of cells from the geometric method after collapsing to apply the cumulative root frequency method and found natural breaks in the data for the current method. This time the cumulative root frequency method had the lowest CV (see Table 3) and all methods met the requirement of at least 15 libraries and 75 percent unit response rate per cell (See Table 4).

Table 3: CV Associated with Imputation Cell Boundary Method After Collapsing Cells

			Geometric Method
CV	887.72	636.90	1650.17

Table 4: Number of Libraries and Unit Response Rates by Imputation Cells and Cell Boundary Method After collapsing Cells

						Geometric Method	
							Response Rate
01	1	31	96.77%	25	100.00%	53	98.11%
01	2	46	100.00%	30	96.67%	48	95.83%
01	3	81	96.30%	33	100.00%	105	99.05%
01	4	53	100.00%	64	96.88%	133	97.74%
01	5	84	96.43%	69	97.10%	146	95.89%
01	6	100	98.00%	91	97.80%	158	97.47%
01	7	93	95.70%	102	97.06%	173	94.22%
01	8	97	98.97%	118	96.61%	130	95.38%
01	9	237	94.51%	199	95.48%	157	93.63%
01	10	205	94.15%	203	96.06%	124	88.71%
01	11	271	87.82%	364	88.74%	71	78.87%
04	1	36	100.00%	18	100.00%	39	100.00%
04	2	36	100.00%	19	100.00%	32	100.00%
04	3	44	100.00%	22	100.00%	38	100.00%
04	4	63	100.00%	21	100.00%	46	100.00%
04	5	71	100.00%	34	100.00%	45	100.00%
04	6	74	98.65%	44	100.00%	85	98.82%
04	7	80	97.50%	57	100.00%	83	100.00%
04	8	96	98.96%	62	98.39%	84	96.43%
04	9	121	96.69%	125	98.40%	127	96.85%
04	10	91	98.90%	152	98.03%	158	98.73%
04	11	112	96.43%	149	97.99%	190	95.79%
04	12	148	95.27%	338	96.75%	198	95.45%
04	13	658	85.41%	589	83.70%	505	82.38%

4.2 Imputation Methods

In survey year 2008, there will be 51 items available for imputation. The breakout of Databases into State Databases, Local Databases, and Other Databases, along with

Number of Registered Borrowers were new items in survey year 2006 and are set to begin being imputed in survey year 2008. For these four items, we used survey year 2007 and 2006 data to test the imputation methods. Since most of the remaining 47 items had been collected for all studied years, we used data from survey years 2002 through 2006 (in reverse chronological order) to test the imputation methods for these items. Also, we used the new cumulative root frequency method from Section 3.1 to define the imputation cells. This was done to insure getting the "best" imputation methods based on the new procedures that will be implemented in survey year 2008.

For PLS, a library is considered a unit response if at least three of the five key items have a valid response. Unit response rates for the five years were around 97 percent. Item response rates varied by collection year. In survey year 2006, the item response rates ranged from 90 percent to 100 percent, whereas in survey year 2003, the rates ranged from 63 percent to 100 percent. Three items - number of Central libraries, Branch libraries, and Bookmobiles - had 100 percent response for all five years, making it unnecessary to research imputation methods for these items.

We first evaluated the imputation methods under the (somewhat unrealistic) assumption that a method could always be performed, i.e. prior year data and data for a highly correlated item were always reported. Table 5 shows the percentage of "best" methods for each evaluation criterion. Although Method 4 (Cell mean) yields the estimates with the lowest Average Bias, it also generally produced a much larger MSE than the other methods. Method 9 (Sequential hot-deck) was never selected as "best" by any of the criteria. For the remaining three criteria, Method 5 (Ratio to another item) had best performance in terms of MSE, MAE, and MSD, trailed by Method 7 (Direct substitution).

Table 5: Percentage of "Best" Methods by Evaluation Criteria.

				Mean Squared Deviation
1: Mean growth rate	0	12	4	12
2: Hot deck growth rate	6	0	0	2
3: Adjusted cell mean	16	6	0	2
4: Cell mean	39	4	0	0
5: Ratio to another item	27	49	45	43
6: Median ratio to another item	2	6	6	12
7: Direct substitution of prior year	10	23	43	27
8: Cell median	0	0	2	2
9: Sequential hot-deck	0	0	0	0

As mentioned above, the results presented in Table 5 are based on an unrealistic assumption because prior year or other item data are not always available. Consequently,

we next evaluated the imputation methods by inducing the full missingness pattern. (i.e., where prior years data and other items data are missing.) For this evaluation, Methods 1, 2, 5, 6, and 7 required a "backup." Methods 3, 4, 8, and 9 do not need "backups", so we decided to choose one from these four methods as a backup to the other five methods. Table 6 presents percentages of "best" methods of the four considered methods by evaluation criterion obtained using the more realistic simulated data. Based on these results, we decided to use Method 3 as the backup method in the next part of the evaluation.

Table 6: Percentage of Backup Methods picked "Best" by Evaluation Criteria

				Mean Squared Deviation
3: Adjusted cell mean	31	82	70	80
4: Cell mean	67	18	4	14
8: Cell median	0	0	26	6
9: Sequential hot-deck	2	0	0	0

Finally, we calculated the new imputes based on the total missingness pattern. Table 7 presents the percentage of methods picked "best" by each criterion when method 3 was used as the back up method.

Table 7: Percentage of Methods picked "Best" by Evaluation Criteria with a back up of adjusted cell mean.

				Mean Squared Deviation
1: Mean growth rate	14	27	12	20
2: Hot deck growth rate	4	0	0	0
3: Adjusted cell mean	21	8	0	2
4: Cell mean	25	2	0	4
5: Ratio to another item	10	8	2	4
6: Median ratio to another item	10	14	10	20
7: Direct substitution of prior year	8	41	59	50
8: Cell median	0	0	17	0
9: Sequential hot-deck	8	0	0	0

Once again, for average bias Methods 3 (Adjusted cell mean) and 4 (Cell mean) had the best performance, but exhibited much poorer performance than the other methods for the

remaining criteria. Method 7 (Direct substitution) had the best performance in terms of MSA, MAE, and MSD, with Method 1 (Mean growth rate) also demonstrating good performance on these criteria (although not as strong as Method 7).

Lastly, we looked at the difference of using data from three or four years back instead of just two years for Methods 1, 2, 5, and 7 (methods that use prior year data). We calculated two separate imputes based on the total missingness pattern, one going back two years and another going back four years. Method 3 (Adjusted cell mean) was used as the backup if no prior year data were available. Generally, going back four years performed better than going back two years for MSE, MAE, and MSD. Table 8 gives the percentage of items by method and criteria and shows that going back four years was better than going back only 2 years.

Table 8: Percentage where 4 years of prior data was better than 2 years of prior data

				Mean Squared Deviation
1: Mean growth rate	12	50	76	68
2: Hot deck growth rate	24	44	74	53
5: Ratio to another item	41	53	82	71
7: Direct substitution of prior year	12	41	94	88

5. Conclusion

For PLS, the current method of determining imputation cell boundaries for the size of population can be burdensome because of the lack of automation. This problem is addressed by using the cumulative root frequency method (which can be automated) to determine the imputation cell boundaries. Moreover, this method has demonstrated statistical (design) advantages, achieving the lowest CV of the considered methods. This provides evidence that the populations within imputation cells are homogeneous, whereas the between-cell populations are heterogeneous.

Since the cumulative root frequency method will be implemented in survey year 2008, we used it in the development of the imputation cells for testing the nine imputation methods. When imputation methods can always be used, Method 5 had the superior performance. However, this strong performance was not seen when we induced a more realistic nonresponse pattern into the data. When the full missingness pattern was taken into account, Methods 7 (Direct substitution) and 1 (Mean growth rate) had the best performance when combined with Method 3 (Adjusted cell mean) as the backup. These results are similar to the results found in 1997; a mean growth rate applied to prior year data with a cell mean as the backup when prior year data were not available gave the "best" imputes for most items.

In an effort to reduce the number of methods being used, we evaluated only one backup to each method. For example, we did not test if using Method 5 (Ratio with another item), and then Method 7 (Direct substitution), and then Method 3 (Adjusted cell mean) would result in a lower Average Bias, MSE, Mean Absolute Error, or Mean Squared

Deviation. Generally, there was an improvement in imputes when going back three or four years instead of two for prior year data.

Since imputed data are generally only released in macro-data, the methods chosen for each item will be decided by the average bias and MSE criteria. Our recommendation is to use either Method 1 (Mean growth rate) or Method 7 (direct substitution) with method 3 (adjusted cell mean) as the backup method for most items. We also suggest that going back four years for prior year data for all methods will be an improvement over the current practice of going to the backup method after only looking for two years of prior data. For the capital revenue and expense items and the database items that are volatile (i.e., change a lot from year to year), we recommend using Method 3 (adjusted cell mean).

Acknowledgments

We acknowledge the excellent review and comments on this paper from Rita Petroni and Katherine J. Thompson from the U.S. Census Bureau. We also acknowledge the assistance given from the Public Libraries Survey staff at the U.S. Census Bureau for their providing data and help in understanding the survey.

References

1. Cochran, W.G. (1977). *Sampling Techniques, 3rd edition*. New York: Wiley.

2. Dalenius, T., and Hodges, J.L. (1959). "Minimum Variance Stratification." *Journal of the American Statistical Association*, **54**, pp. 88-101.

3. Gunning, P. and Horgan, J.M. (2004). "A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations", Retrieved on September 26, 2008 from http://www.statcan.ca/english/ads/12-001-XIE/12-001-XIE20040027749.pdf

4. Hilton, J. (1997). "Public Libraries Survey Imputation Research". Internal Memo, U.S. Census Bureau.

5. Nordholt, E.S. (1998). "Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistics Review*, **66**, pp. 157-180.