Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

CONSTRUCTION OF IMPUTATION CELLS FOR THE CANADIAN LABOUR FORCE SURVEY

D. Haziza, C. Charbonnier, O. S. Y. Chow, J. F. Beaumont[1]

ABSTRACT

In large scale surveys, it is almost guaranteed that some level of nonresponse will occur. Generally, statistical agencies use imputation as a way to treat item nonresponse. A common preliminary to imputation is the formation of imputation cells. In this article, we study the formation of these cells using two methods. The first method is similar to that of Eltinge and Yansaneh (1997) in the case of weighting cells and the second is the method currently used in the Canadian Labour Force Survey. Using Labour Force data, simulation studies are performed to test the impact of the response rate, the response mechanism and constraints on the quality of the point estimator under both methods.

KEY WORDS: Imputation cells; hot-deck imputation; uniform response mechanism; ignorable response mechanism; nonignorable response mechanism.

1. INTRODUCTION

1.1 The problem

Despite the best efforts made by the survey staff to maximize response, it is almost certain that some degree of nonresponse will occur in large scale surveys. Essentially, survey statisticians distinguish between two types of nonresponse, total or unit nonresponse (when no information is collected on a sampled unit) and partial or item nonresponse (when the absence of information is limited only to some variables). Generally, weighting adjustment methods are used to compensate for unit nonresponse whereas imputation is used to compensate for item nonresponse. The main idea behind a weighting adjustment is to increase the sampling weights of the respondents in order to compensate for the nonrespondents, while imputation is a process where an "artificial value" is produced to replace a missing value. It is customary in both weighting and imputation to first classify respondents and nonrespondents into cells, formed on the basis of information recorded for all units in the sample. There are at least two reasons motivating the formation of cells instead of imputing directly the value resulting from the use of a regression model: (1) it is convenient when it is desired to impute more than one variable at a time and (2) it is more robust to model misspecification.

In practice, there are several methods used for the formation of imputation cells. Common methods include the following: (1) For stratified samples, one may use strata or groups of strata as imputation cells. This is common in business surveys. (2) If the parameters of interest are domain means or totals, and if the domains of interest are known at the planning stage, then one may use these domains as imputation cells. (3) Imputation cells may also be formed through combinations of simple auxiliary variables such as province, age, sex, etc. (4) Available decision tree software such as Knowledge Seeker (Kass, 1980) and CART (Steinberg and Colla, 1995) can also be used to form homogeneous imputation cells.

The objective of this paper is to investigate the formation of imputation cells using two methods. The first method (*method1*) is similar to the one proposed by Eltinge and Yansaneh (1997) in the case of weighting

[1] David Haziza, Cédric Charbonnier, Ophelia Chow and Jean-François Beaumont, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

cells. The second method (*method2*) is the one currently used in the Canadian Labour Force Survey (LFS). Throughout this article, random hot-deck imputation is used to compensate for nonresponse to item *y*. That is, a missing value for item *y* is replaced by the *y*-value of a unit selected randomly from the set of respondents to item *y*.

1.2 Brief overview of the Labour Force Survey

The Canadian Labour Force Survey (LFS) is a monthly panel survey that provides detailed information on employment characteristics such as the unemployment rate (at provincial and national levels), hours worked, job description and earnings. It also provides some information on the traits of the working population such as age, marital status, education etc. The LFS consists of a sample of approximately 53,000 households (approximately 130,000 respondents) in Canada. Each household remains in the survey for six consecutive months, and every month, one sixth of the sample is replaced. Within each household, every member, age 15 or older, is interviewed. Excluded from the survey are persons living in the Indian Reserves, full time members of the Canadian Armed force and inmates of institutions. These exclusions only account for less than 2 % of people aged 15 or older in Canada.

1.3 Simulation Data

We have performed two simulation studies to assess the performance of the two methods. From the LFS, we selected individuals who were asked the earnings questions in January 2001. This provided a data set with 11270 individuals with more than 100 variables.

Since questions on income are sensitive they often have a lower response rate than other variables. For this reason we have chosen *weekly earnings* as the variable of interest. For the simulation we considered this data set as the population from which nonresponse was generated. From the complete file the "true" population mean of *weekly earnings* was $554.97. This "true" value was compared to the imputed estimator under three types of response mechanism: uniform, ignorable and nonignorable (see section 2).

2. DEFINITIONS AND THEORETICAL RESULTS

Let U be a population of size N. The objective is to estimate the population mean $\bar{Y} = \frac{1}{N}\sum_U y_i$ of a variable of interest *y*. Assume that a sample s of size *n* is selected from U according to some design $p(s)$. Because of nonresponse to item *y*, only a subset s_r of *s* is observed. It is assumed that the response mechanism $q(s_r \mid s)$ is independent of *s*, that is, the conditional probability that s_r is observed, $q(s_r \mid s)$, is equal to $q(s_r)$. Let a_i be the response indicator equal to 1 if the selected sampled unit *i* is a respondent to item *y* and equal to 0 otherwise. Assume that the a_i's are independent Bernoulli (p_i) random variables.

Kalton and Kasprzyk (1986) showed that several imputation methods (mean imputation, hot-deck imputation, ratio imputation,...) may be represented using the following model

$$m: y_i = \mathbf{z}_i' \boldsymbol{\beta} + \varepsilon_i, \tag{1}$$

where $\mathbf{z}$ is a vector of auxiliary variables available for all units in the sample and $E_m(\varepsilon_i) = 0,\ E_m(\varepsilon_i\, \varepsilon_j) = 0,\ i \neq j,\ V_m(\varepsilon_i) = \sigma_i^2$. We assume that $\sigma_i^2 = \boldsymbol{\lambda}'\mathbf{z}_i$ for some constant vector $\boldsymbol{\lambda}$. Note that this does not severely restrict the range of imputation models.

An imputed estimator of $\bar{Y}$, denoted by $\bar{y}_I$, is given by

$$\bar{y}_I = \frac{1}{\sum_s w_i} \sum_s w_i \tilde{y}_i \tag{2}$$

where w_i is the survey weight attached to unit *i*, $\tilde{y}_i = \begin{cases} y_i & \text{if sampled unit } i \text{ is a respondent} \\ y_i^* & \text{if sampled unit } i \text{ is a nonrespondent} \end{cases}$

and y_i^* denotes the imputed value for missing value y_i. Model (1) can be used to justify the form of the imputed value for unit *i*, which is given by $y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_\mathbf{r}$, where $\hat{\mathbf{B}}_\mathbf{r} = \left(\sum_s w_i a_i \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / \sigma_i^2 \right)^{-1} \sum_s w_i a_i \mathbf{z}_\mathbf{i} y_i / \sigma_i^2$. Using the y_i^*'s, the asymptotic bias under the sampling design and the response mechanism is given by

$$\text{Bias}(\bar{y}_I) \approx \frac{1}{N} \sum_U \mathbf{z}_\mathbf{i}' (\mathbf{B}_p - \mathbf{B}), \tag{3}$$

where $\mathbf{B}_p = \left(\sum_U p_i \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / \sigma_i^2 \right)^{-1} \sum_U p_i \mathbf{z}_\mathbf{i} y_i / \sigma_i^2$ and $\mathbf{B} = \left(\sum_U \mathbf{z}_\mathbf{i} \mathbf{z}_\mathbf{i}' / \sigma_i^2 \right)^{-1} \sum_U \mathbf{z}_\mathbf{i} y_i / \sigma_i^2$. Note that $\mathbf{B}$ is an estimator of the model parameter $\boldsymbol{\beta}$, that is, $\mathbf{B}$ is an estimate that we would have obtained had the entire finite population been observed. Also, note from (3) that, the imputed estimator (2) is approximately unbiased if $\mathbf{B}_p - \mathbf{B} \approx 0$. Under model (1), $\mathbf{B}_p - \mathbf{B}$ is model-consistent for zero if the model covariance between the probability of response p_i and the model residual $(y_i - \mathbf{z}_\mathbf{i}' \boldsymbol{\beta})$ is zero for all *i* in the population. Under mean or hot-deck imputation, (3) reduces to

$$\text{Bias}(\bar{y}_I) \approx \frac{1}{N\bar{P}} \sum_U (p_i - \bar{P})(y_i - \bar{Y}), \tag{4}$$

where $\bar{P} = \frac{1}{N} \sum_U p_i$. From (4), we note that the imputed estimator (2) is approximately unbiased if the model covariance between the probability of response p_i and the population mean $\bar{Y}$ is zero for all *i* in the population or if the population covariance between variables p and y is zero. This requirement is clearly unrealistic in practice since it is likely that the response probability will depend on variables potentially correlated with the variable of interest *y*. For this reason, $\bar{y}_I$ is called an "unadjusted estimator". To reduce the bias, one often partitions the population into *C* imputation cells U_c, $\left(\bigcup_{c=1}^{C} U_c = U \right)$ which leads to a corresponding partitioning of the sample *s* into cells s_c, $\left(\bigcup_{c=1}^{C} s_c = s \right)$. Imputation is then performed independently within each cell, leading to an "adjusted" estimator given by

$$\bar{y}_{I,c} = \sum_{c=1}^{C} w_c' \bar{y}_c \tag{5}$$

where $w_c' = \frac{\sum_{s_c} w_i}{\sum_s w_i}$ is a measure of the relative importance of cell *c* and $\bar{y}_c = \frac{1}{\sum_{s_c} w_i} \sum_{s_c} w_i \tilde{y}_i$, is the imputed estimator for cell *c*, $c = 1, \ldots, C$. Under mean or hot-deck imputation, the bias of the adjusted estimator is given by

$$\text{Bias}(\bar{y}_{I,c}) \approx \frac{1}{N} \sum_{c=1}^{C} \bar{P}_c^{-1} \sum_{U_c} (p_i - \bar{P}_c)(y_i - \bar{Y}_c), \tag{6}$$

where $\overline{P}_c = \frac{1}{N_c}\sum_{U_c} p_i$, $\overline{Y}_c = \frac{1}{N_c}\sum_{U_c} y_i$ and N_c is the number of units in U_c. It follows that the bias in (6) is approximately equal to zero if the population covariance between variables p and y is approximately equal to zero for each cell. In practice, one attempts to accomplish this by constructing cells that are homogeneous in the response probabilities p_i or in the item y_i, or both.

Survey statisticians generally define three types of response mechanisms: uniform, ignorable and nonignorable. A uniform response mechanism is a mechanism for which $p_i = P(i \in s_r) = p \quad \forall i \in U$. Informally speaking, a response mechanism is ignorable with respect to a model if it depends on some auxiliary variables of the model but not on the model error. For example, a mechanism for which the probability of response depends on some auxiliary variables but not on the variable of interest; that is $p_i = P(i \in s_r) = p(\mathbf{z}_i)$ is ignorable with respect to model (1). In other words, if the response mechanism is independent of the model distribution then it is ignorable with respect to that model. An nonignorable response mechanism is a mechanism which is not ignorable with respect to some model. For example, if the probability of response depends on the variable of interest; that is, $p_i = P(i \in s_r) = p(\mathbf{z}_i, y_i)$, then the response mechanism is not ignorable with respect to model (1). Note that under a uniform response mechanism $p_i = p$, and (4) is equal to 0. Consequently, if it is believed that the response mechanism is uniform throughout the population, it is not necessary to form imputation cells to reduce the bias. However, it may be useful to form imputation cells in order to reduce the variance of the imputed estimator. Also, note that under nonignorable response, (5) will not generally be equal to 0 since in this case, the population covariance between p and y is likely not to be 0.

3. THE PROPOSED METHOD

3.1 Description of the method

In this section, we first describe *method1* proposed to form imputation cells. This method is similar to the one used by Eltinge and Yansaneh (1997) in the case of weighting cells. As noted in section 2, the goal is to partition the sample in such a way that, within cells, units (respondents and nonrespondents) are homogeneous with respect to one of the two criteria, response probabilities p_i or item values y_i. The steps for the formation of cells may then be described as follow:

1. Using auxiliary information available on all sampled units, predict the response probabilities p_i and the variable of interest y_i. Two scores, $\hat{p}_i$ and $\hat{y}_i$, are then available for all units in the sample (respondents and nonrespondents).
2. Choose one of the two criteria ($\hat{p}_i$ or $\hat{y}_i$) and, using an equal-quantile method or a classification algorithm, partition the sample into cells. In this paper, we use an equal-quantile method to partition the sample, that is the cell boundaries are determined by the j/k quantiles of the $\hat{p}_i$ or $\hat{y}_i$ populations, $j = 1,...,k-1$. We are currently investigating the use of classification algorithms.
3. Within each cell, perform random hot-deck imputation and compute the within-cell imputed estimator $\overline{y}_c$.
4. Combine these estimators as in (5).

Note that under *method1*, the cells are disjoint so that a donor may only be used in the cell in which it belongs.

3.1.1 Modelling the response probabilities and the item *weekly earnings*

The first step in forming the imputation cells is to predict the response probabilities p_i and the variable of interest y_i. The reader is referred to Chow (2001) for a detailed description of models building.

Since $p_i = P(a_i = 1)$ and a_i is a binary variable, prediction of p_i has been done using a logistic regression. The following variables were identified as important predictors:
province, age, student status, standard industry classification, year starting working, standard occupational classification, marital status, fixed/variying hours, owning/renting dwelling status, type of dwelling, tips/comissions job, household size, class of worker, sex and *labour force status*.

A linear regression model was used to predict the continuous variable of interest *weekly earnings*. The following variables were identified as important predictors:
full time/part time status, standard occupational classification, age, standard industry classification, year starting working, education level, student status, marital status, sex, firm size, permanent/temporary job status and *union status*.

3.1.2 Forming the cells

To illustrate *method1*, we generated nonresponse, according to the *ignorable1* response mechanism described in the Appendix. The overall response rate was fixed at 0.7 (which is slightly lower than the worst extreme observed in the LFS for *weekly earnings*). Using the selected predictors (see section 3.1.1), we obtained two scores ($\hat{p}_i$ and $\hat{y}_i$) for every unit in the sample. Then, using an equal-quantile method, we partitioned the sample into equal size cells.

Table 1 shows the estimates after performing random hot-deck imputation within each cell as well as the differences $\bar{y}_{I,c} - \bar{y}_{I,1}$ measuring the impact on the estimate of using c cells versus using 1 cell and $\bar{y}_{I,c} - \bar{y}_{I,c-1}$ measuring the impact on the estimate when using an additional cell. As the number of cells increases, the estimates quickly approach the true population mean. With 5 imputation cells, the estimated relative error is only 0.91% and only marginal improvement is obtained with a larger number of cells. Looking at the difference, $\bar{y}_{I,c} - \bar{y}_{I,1}$, measuring the impact of using c cells versus using 1 cell, again after 5 imputation cells the impact of using 6 or more cells stabilises around 70. Differences between using c cells and c-1 cells also show a rapid decrease in the gain from using one additional imputation cell.

Table 1: Estimates of *weekly earnings*

Number of cells	Estimate	Difference ($\bar{y}_{I,c} - \bar{y}_{I,1}$)	Difference ($\bar{y}_{I,c} - \bar{y}_{I,c-1}$)	Relative error
1	627.35	.	.	0.1304
2	581.85	-45.49	-45.49	0.0484
3	568.64	-58.71	-13.21	0.0246
4	562.71	-64.64	-5.92	0.0140
5	560.03	-67.32	-2.68	0.0091
6	558.43	-68.92	-1.59	0.0062
7	557.83	-69.52	-0.59	0.0052
8	557.10	-70.25	-0.73	0.0038
9	556.47	-70.87	-0.62	0.0027
10	555.85	-71.50	-0.62	0.0016

3.1.3 Within-cells diagnostics

Once the cells are formed, it may be useful to perform some diagnostics. A possible diagnostic is

the assessment of the within-cell bias. Consider the following estimators

$$\bar{y}_{w-c}^{(1)} = \frac{1}{\sum_{s_c} w_i \hat{p}_i^{-1}} \sum_{s_c} w_i \hat{p}_i^{-1} y_i \tag{7}$$

and

$$\bar{y}_{w-c}^{(2)} = \frac{1}{\sum_{s_c} w_i} \sum_{s_c} w_i \hat{y}_i \tag{8}$$

It is easily seen that the within-cell estimator (7) is approximately unbiased under the sampling design and the response mechanism for $\bar{Y}_c$ provided that $\hat{p}_i = p_i$ for all $i \in s_c$. Similarly, the within-cell estimator (8) is approximately design-model unbiased for $\bar{Y}_c$ provided that $E_m(\hat{y}_i) = E_m(y_i)$, where $E_m(.)$ denotes the expectation operator with respect to the model given in (1). These two estimators may then be used to measure the within-cell bias $\hat{B}_c = \bar{y}_c - \bar{y}_{w-c}^{(j)}$ $j = 1,2$. Large values of $\hat{B}_c$ may indicate a large within-cell bias in $\bar{y}_c$ but as Eltinge and Yansaneh (1997) pointed out, it may also indicate that the errors $\hat{p}_i - p_i$ (or $\hat{y}_i - y_i$) are large. Hence, caution should be applied when interpreting large values of $\hat{B}_c$. If however, one considers both diagnostics (7) and (8) simultaneously, large $\hat{B}_c$ in only the probability of response p_i or the variable of interest y_i is less indicative of bias than if both values are large. Regardless of the cause, a large value of $\hat{B}_c$ should be investigated.

Table 2 shows the within-cell estimates as well as the within-cell bias for up to four cells. Note that only estimator (7) was used for diagnostics purposes. As the number of cell increases the data within the cells becomes more homogeneous and therefore the within-cell bias decreases. We observe in Table 2, that in each case, the within-cell bias for the first cell is always larger than the others. This may indicate some problems in the modelling of the probability of response. The use of a more sophisticated method for forming the cells, e.g., clustering algorithms may also be useful to reduce the within-cell bias in the first cell.

Table 2: Within cell estimates and within-cell biases

Number of cells		$\bar{y}_c$	$\bar{y}_{w-c}$	$\hat{B}_c$
1	1	627.35	569.04	58.31
2	2.1	355.33	326.61	28.72
	2.2	808.38	795.07	13.30
3	3.1	254.62	235.51	19.10
	3.2	551.79	545.61	6.18
	3.3	899.34	890.38	8.95
4	4.1	199.17	184.05	15.18
	4.2	451.06	446.30	4.76
	4.3	646.84	643.19	3.64
	4.4	953.49	947.23	6.26

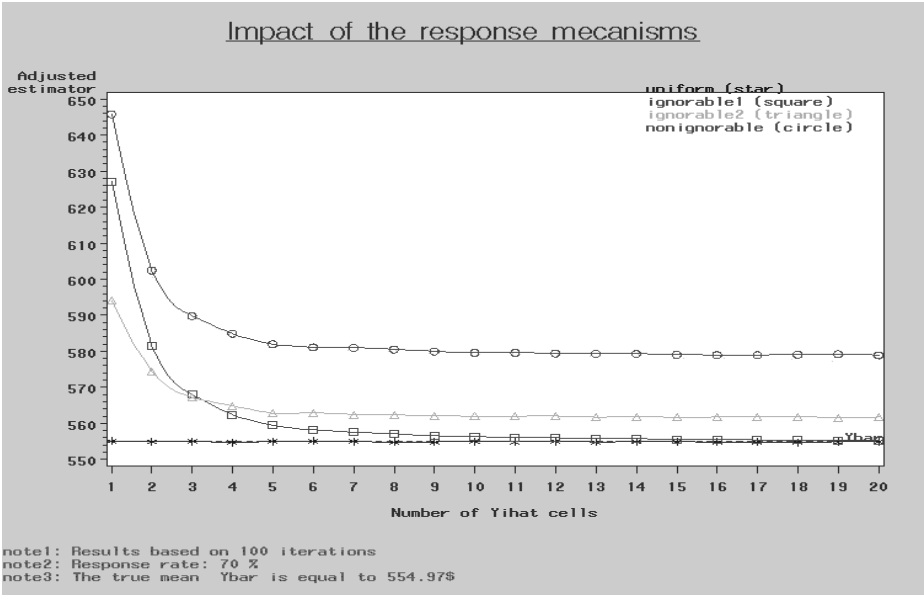
3.2 Simulation results

To test *method1*, we conducted a simulation study using the population described in section 1.3. We generated nonresponse $R = 100$ times according to the response mechanisms described in Appendix. For each iteration, the response rate was set at 70%. The response probabilities $\hat{p}_i$ and item values $\hat{y}_i$ were then

predicted and cells were formed using the equal-quantile method. In each cell, random hot-deck imputation was performed and the estimates of the population mean for *weekly earnings* were computed.

Recall that the true mean in the population was $554.97. Under uniform response, the estimator is approximately unbiased regardless of the number of cells, as expected (see Graph 1). Under the ignorable response mechanisms (*ignorable1* and *ignorable2*), we note that, as the number of cells increases, the bias approaches zero. In the nonignorable case, we see that, although the use of cells leads to an important reduction of the bias, the adjusted estimator remains biased, as expected (see section 2). Once again, note that most of the bias reduction is achieved with a relatively small number of cells (5 to 10). After this the bias becomes relatively stable and there is no further gain from increasing the number of cells. This result, which seems somehow surprising at first hand, has also been observed by Eltinge and Yansaneh (1997) in the case on weighting cells. We obtained similar results when the response probabilities were used as the criteria to form cells and also with a response rate of 0.5.

Graph 1



4. THE CURRENT METHOD

4.1 Description of the method

In this section, we briefly describe the current method (*method2*) used in the LFS for the formation of imputation cells. For a more detailed description of the method, the reader is referred to Lorenz (1996). Here, the cells are formed through combinations of categorical variables. First, one selects variables that are correlated with the variable of interest *weekly earnings* and then classifies these variables in order of importance, from the most significant to the least significant. For our data, the variables in descending importance are: *full time/part time status*, *standard occupational classification*, *age*, *standard industry classification*, *year starting working*, *education level*, *student status*, *marital status*, *sex*, *firm size*, *permanent/temporary job status* and *union status*.

To ensure stability of the point estimator, the following two constraints are specified:

1. The minimal number of donors in a given cell is fixed to k .
2. In a given cell, the number of donors must be greater than the number of recipients.

The imputation cells are defined by the different combinations of the variables listed above. Random Hot-deck imputation is then performed in all the cells satisfying the above constraints. However, since the initial number of cells is very large, it is likely that a large number of cells will not satisfy the constraints. In this case, the least significant variable (*union status*) is dropped from the list and the cells are defined by combinations of the remaining variables. Once again, random hot-deck imputation is performed in all cells that satisfy the constraints. If, there are still some cells that do not satisfy the constraints, the least significant variable (*permanent/temporary job)* is dropped from the list and so on. This process continues until each recipient has found a donor or the maximum number of variables has been dropped. In this case, no collapsing is allowed beyond *standard occupational classification*. Note that unlike *method1*, the cells are not disjoint; therefore, a donor may be used several times at different stages of the process.

4.2 Simulation study

We conducted a simulation study to test the impact of the response mechanisms and the constraints on the quality (bias) of the point estimator for the population mean. Nonresponse was generated $R = 300$ times according to the response mechanisms described in the Appendix. At each iteration, imputation cells were formed according to the method described in section 4.1. The overall response rate was set to 70%.

Table 3, 4 and 5 report the relative bias of the point estimator where the minimal number of donors k in each cell has been fixed to 3, 8 and 1 respectively. From the results it is clear that under uniform response, the point estimate is approximately unbiased (less than 0.5 %) regardless of k. Uniform response throughout the population is a strong assumption so the inference will be valid regardless of the model used to form the cells. For the *ignorable1* and *ignorable2* response mechanisms, we note that as the severity of the constraints increases (i.e., the minimal number of donors in each cell increases), the relative bias of the estimator increases; e.g., with $k = 1$ under ignorable1, the relative bias of the estimator is 0.6 % whereas with $k = 8$, the relative bias of the estimator is 9.02 %. Under nonignorable response, the point estimator is biased even with $k = 1$ (4.52 %), as expected, but the bias increase as k increases (11.46 % with $k = 8$). One explanation may be that as the constraints become more severe, it is increasingly difficult to have donors within a cell, so more variables will be dropped from the list. Consequently, some important predictors may be omitted, leading to biased estimates.

In conclusion, it seems that for all types of response mechanisms (except uniform response), the quality of the estimator is altered as the severity of the constraints increases. It is interesting to note that the quality of the estimator may also be sensitive to the order of the variables. Indeed, if one does not classify the variables properly, some important predictors may be omitted from the list in a very early stage of the process, which may lead to a substantial amount of bias. Also, the quality of the estimator may be sensitive to the response rate, since as the response rate decreases, the second constraint may be more difficult to satisfy, again leading to a substantial bias. Finally, one should also look at the sample size versus the number of cells created by the combination of the selected variables. If the sample size is substantially smaller than the number of cells, the impact of the constraints may be greater on the quality of the point estimate.

Table 3 Relative bias of the estimator for $k = 3$

Mechanism	**Estimate**	**Relative Bias (%)**	**95% CI for Relative Bias**
Uniform	552.64	-0.42	(-0.52; -0.32)
Ignorable 1	591.04	6.50	(6.10; 6.89)
Ignorable 2	561.28	1.14	(1.02; 1.25)
Nonignorable	600.04	8.12	(7.72; 8.52)

Table 4 Relative bias of the estimator for $k = 8$

Mechanism	Estimate	Relative Bias (%)	95% CI for Relative Bias
Uniform	553.58	-0.25	(-0.35;-0.15)
Ignorable 1	604.98	9.02	(8.70;9.31)
Ignorable 2	564.42	1.70	(1.55;1.85)
Nonignorable	618.57	11.46	(11.05;11.87)

Table 5 Relative bias of the estimator for $k = 1$

Mechanism	Estimate	Relative Bias (%)	95% CI for Relative Bias
Uniform	552.79	-0.39	(-0.49; -0.30)
Ignorable 1	558.27	0.60	(0.52; 0.67)
Ignorable 2	558.93	0.71	(0.64; 0.78)
Nonignorable	580.03	4.52	(4.45; 4.58)

REFERENCES

Chow, O. S. Y., (2001), "Model Building for construction of imputation cells in the Labour Force Survey", Internal report, Ottawa, Canada: Statistics Canada.

Eltinge, J. L., and Yansaneh, I. S. (1997), "Diagnostics for formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey", *Survey Methodology*, 23, pp. 33-40.

Kalton, G., and Kasprzyk, D. (1986), "The treatment of missing survey data", *Survey Methodology*, 12, pp. 1-16.

Kass, G. V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, 29, No 2, pp. 119-127.

Lorenz, P. (1996), "Head Office Hot Deck Imputation System Specifications", Internal Report, Ottawa, Canada: Statistics Canada.

Smith, P. J., Hoaglin, D. C., Battaglia, M. P., Rao, J. N. K., and Daniels, D. (2001), "Evaluation of Adjustment for Partial Nonresponse Bias, Applied to Provider nonresponse in the National Immunization Survey", paper presented at the Annual Meeting of the Statistical Society of Canada, Ottawa, Canada.

Steinberg, D., and Colla, P. (1995), "Tree-Structured Non-Parametric Data Analysis", San Diego, CA: Salford Systems.

APPENDIX

The response mechanisms are described below:

Uniform	p_i is constant, i.e. $\forall i, p_i = p$
Ignorable1	p_i is a function of $\hat{y}_i$ such that $p_i = \alpha + (1-\alpha)\exp(1-\lambda\hat{y}_i)$, $0 \le \alpha \le 1$
Ignorable2	p_i is a function of x_i (Age) such that $p_i = \alpha + (1-\alpha)\exp(1-\lambda x_i)$, $0 \le \alpha \le 1$
Nonignorable	p_i is a function of y_i (*Weekly Earnings*), such that $p_i = \alpha + (1-\alpha)\exp(1-\lambda y_i)$, $0 \le \alpha \le 1$

The constants α and λ are fixed to obtain the desired response rate.