Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

WORKING PAPER NO. BSMD-87-002

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-87-002

MÉTHODOLOGIE

A COMPARISON OF DIFFERENT IMPUTATION TECHNIQUES FOR QUANTITATIVE DATA

by

Marcel Bureau
Sylvie Michaud
Madhavi Sistla

BSMD

November 3, 1986

* This is a preliminary version. Do not quote without author's permission. Comments are welcome.

1. PURPOSE OF THE STUDY

The purpose of this project is to compare, using a simulation study, the quality of estimators associated with various imputation methods. The comparison was performed by calculating the bias, variance and mean square error of each of the estimators studied. In addition, a calculation of the correlation coefficient between each variable after imputation is used to check whether a given imputation method maintains the correlational structure among variables.

N.B. Madhavi Sistla, a student in the COSEP program, performed the simulation and imputation. For further details, see her report entitled:"A COMPARISON OF DIFFERENT IMPUTATION TECHNIQUES", BSMD, August 1986. The broad outlines of that report are reproduced here in order to present the entire study in a single document.

2. SIMULATION METHODOLOGY

The population studied was artificially generated; it consisted of five variables (X_1, ..., X_5) and 1000 records. Each variable is a linear combination of independent exponential random variables with mean 1.0. The variables are defined as follows:

$$X_1 = 1.0 * E_1,$$

$$X_2 = 0.75 * E_1 + 0.661438 * E_2,$$

$$X_3 = 0.25 * E_1 + 0.321278 * E_2 + 0.913392 * E_3,$$

$$X_4 = (0.20 * E_1 - 0.075593 * E_2 + 0.519257 * E_3 + 0.827440 * E_4),$$

$$X_5 = (0.40 * E_1 + 0.302372 * E_2 + 0.495797 * E_3 + 0.586642 * E_4 + 0.398257 * E_5),$$

where the E_i, (i=1, ..., 5), are exponential variables with mean 1.0. The constants were chosen to ensure that the correlations among the variables $X_1, \ldots, X_5$ approximately match a given structure. The correlations fall between 0.2 and 0.8: (See the first line of Table F).

2.2 Creation of non-response

Three non-response models are considered. For each model, 500 records exhibit total response (donors) and 500, partial non-response (candidates). Among the candidates, non-response is generated randomly and independently for the variables X_1, X_3 and X_5. The non-response rate for these variables for the candidate subpopulation is set at 50%, 75% and 100% respectively. Thus, the overall response rates are 25%, 37.5%, and 50% respectively. Variables X_2 and X_4 are always present.

The three models differ in how they determine the population of donors and candidates. The first model (model A) depicts the case in which respondents and non-respondents share the same distribution (that is, there is no response bias); 500 records are randomly chosen from the 1000 records to form the candidate subpopulation. The remaining 500 records constitute the donor subpopulation. The second and third models depict cases in which the distributions of the donors and candidates are different. In the second model (model B), 125 candidates are chosen from the records whose X_5 value is above the median value for X_5, and 375 are chosen from those whose value is below the median value for X_5. The remaining 500 records form the donor subpopulation. In the third model (model C), 125 candidates are chosen from the records whose X_5 value is below the median value for X_5, and 375 are chosen from those whose X_5 value is above the median value for X_5. The remaining 500 records form the donor subpopulation. Thus, models B and C correspond to situations where X_5 is biased up and down respectively. (That is, an estimate which was based on responses only and did not adjust weights would be biased in this way).

2.3 Sampling and replications

Each of the three models studied consists of a population of 1000 records; 500 donors and 500 candidates. In each of the three populations, a simple random

sample of 500 records is selected. Every imputation method is then applied to each variable displaying non-response and estimates are calculated. In each model, the sample selection, imputation and estimation process is replicated 25 times.

3. IMPUTATION TECHNIQUES STUDIED

3.1 Introduction

The imputation methods studied may be grouped under the two following categories. First, methods which make use of information supplied by the respondents and, second, methods involving the nearest neighbour technique. Briefly, the nearest neighbour technique consists of finding, for a given candidate, the closest donor (according to a defined distance function). To do so, the data values are transformed to be uniformly distributed in the range (0,1) to remove scaling effects due to a large range of variables in the data set. The nearest neighbour technique must make use of matching variables. A matching variable is defined as a variable which is present in both the donor and candidate records, and is related to the variable being imputed. In this study, the MINMAX distance function (L_∞ norm) is used. The distance between a candidate and a donor is therefore determined as follows:

$$D(X_j,Y) = \max_i |x_{ij} - y_i|,$$

where Y is the candidate, X_j is the jth donor and i is a subscript designating matching variables. Note that the nearest neighbour technique consists of selecting the minimum distance D (MINMAX) over all donors j.

3.2 Estimators studied

If Y is the missing variable (X_1, X_3 or X_5),
Z is the auxiliary variable (X_2, X_4),
c is the subscript designating a candidate,
d is the subscript designating a donor,
r is the subscript designating a respondent[1],

[1] The difference between a donor and a respondent is that a respondent is defined separately for each variable while a donor is defined for the entire record. The donor is therefore a record for which there is a response for every variable.

n_r is the number of respondents for variable Y,

n is the number of donors,

N is the size of the population,

(i) is a subscript referring to the ith nearest neighbour,

then $$\bar{Y}_r = 1/n_r \sum_{i=1}^{n_r} y_i ,$$

$$\bar{Z}_r = 1/n_r \sum_{i=1}^{n_r} z_i ,$$

$$\bar{Z} = 1/N \sum_{i=1}^{N} z_i .$$

The following imputation methods were studied:

1. Imputation using the respondents' mean

$$Y_c = \bar{Y}_r ,$$

2. and 3. Imputation of the respondents' mean, with ratio adjustment

$$Y_c = \bar{Y}_r \cdot Z_c/\bar{Z}_r ,$$

4. Imputation of the value of the nearest neighbour

$$Y_c = Y_{d(1)} ,$$

5. Imputation of the mean of the values of the five nearest neighbours

$$Y_c = 1/5 \sum_{i=1}^{5} Y_{d(i)} ,$$

6. and 7. Imputation of the value of the nearest neighbour, with ratio adjustment

$$Y_c = Y_{d(1)} \cdot Z_c/Z_{d(1)} ,$$

8. and 9. Imputation of the respondents' mean, adjusted by a weighting factor

$$Y_c = \bar{Y}_r \cdot \bar{Z}/\bar{Z}_r.$$

Note that the three methods in which an adjustment is made using an auxiliary variable are first applied using the auxiliary variable with the highest correlation (estimators 2, 6 and 8), then using the auxiliary variable with the lowest correlation (estimators 3, 7 and 9). Table 1 displays the relative ranking of the correlations.

TABLE 1

Variable requiring imputation	:	X_1	X_3	X_5
Auxiliary variable with highest correlation	:	X_4	X_2	X_2
Auxiliary variable with lowest correlation	:	X_2	X_4	X_4

Note also that estimators 8 and 9 are essentially ratio estimators. The same value is imputed for all candidates. Similarly, the same value is imputed for all candidates with estimator 1.

The estimates of the mean, obtained after the imputation will be of the form

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n_r} y_{di} + \sum_{i=n_r+1}^{N} y_{ci}}{N}.$$

If $\hat{\bar{Y}}$ is the estimate of the mean,

$\bar{Y}$ is the true population mean,

m is the number of replicates (it is 25 for this study).

The variance, bias and MSE (mean square error) of the estimates can be expressed as:

$$\text{variance} = V(\hat{\bar{Y}}) = \frac{\sum_{i=1}^{m} [\hat{\bar{Y}}_i - (\sum_{i=1}^{m} \frac{\hat{\bar{Y}}_i}{m})]^2}{m-1}$$

$$\text{bias} = \quad B(\hat{\bar{Y}}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{\bar{Y}}_i - \bar{Y})$$

$$\text{MSE} = \quad V(\hat{\bar{Y}}) + B(\hat{\bar{Y}})^2$$

For each of the 25 replicates in the simulation, an estimate of the variance can be obtained as:

i) $$V(\hat{\bar{Y}}) = \frac{(N-n_r)}{N} * \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{(y_{ri} - \bar{y}_r)^2}{n_r - 1}$$

or,

ii) $$V(\hat{\bar{Y}}) = \frac{(N-n_r)}{N} * \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{(y_{ri} - \hat{R}\, x_{ri})^2}{n_r - 1} ,$$

for the ratio estimators.

An average of these variances over the 25 replicates can be compared to the variance of the imputed estimates.

4. PRESENTATION OF RESULTS AND DISCUSSION

4.1 Analysis of Variance

To compare the imputation methods against each other, it was decided to perform an analysis of variance (ANOVA) to test the hypothesis of equal means. An ANOVA was performed separately for each of the three variables imputed and for each of the three models studied in the simulation. The unit of observation was the mean value of each variable, after imputation. Recall that the classic ANOVA test requires three basic assumptions. First, there must be normality of observations. That is, the means associated with the 25 replications for a given imputation method must be normally distributed. A Shapiro-Wilk normality test was therefore carried out on each of the three imputed variables, for each of the nine methods and for the three models studied. The results of these tests showed, in general, that the normality hypothesis is valid. In fact, for the 81 cases tested, the normality hypothesis was rejected in only 3, at a 5%

significance level. This is about the number that would be expected. Second, for each imputation method, the variance of the observations must be constant (the variance of the 25 means). A Hartley variance homogeneity test was performed and, at the 5% significance level, equality of variances was confirmed in every case but one. Finally, if it is assumed that the various imputation experiments are independent, then it can be accepted that all the assumptions required for the ANOVA are satisfied. That is, it is assumed that the outcome of one imputation method is not related to the outcome of another, on the same data set. While it is unlikely that this assumption is completely true, it will be regarded as such.

The actual ANOVA can now be performed. The purpose is to test the null hypothesis (H_0) that the means obtained for each of the nine imputation methods are equal. For cases in which the null hypothesis is rejected, the methods which are significantly different can be determined from a contrast analysis using Tukey's T method.

Using a 5% level of significance, it is found that, for model A, H_0 cannot be rejected for variables X_1 and X_3. The conclusion is, therefore, that there are no significant differences between the various imputation methods. On the other hand, still using a 5% level of significance, the hypothesis that the methods are equal is rejected for variable X_5. At the 1% level of significance, however, H_0, is not rejected. The contrast analysis shows that method 9 produces significantly higher estimates than method 5 at the 5% level of significance.

At the 1% level of significance, H_0 is strongly rejected, for all variables, in the cases of models B and C. In other words, there is always a significant difference between the imputation methods. Note that the results corresponding to models B and C are similar but not exactly identical. This is to be expected due to the asymmetrical distributions of the variables. The missings in X_5 are concentrated at the low end in model B, but would be more spread out in the upper tail in model C. From the contrast analysis, it appears that the imputation methods which use the nearest neighbour technique tend to be grouped while methods using respondents show much greater dispersion; see tables A and B. However, it should be noted that, despite this apparent homogeneity, there are instances in which methods using the nearest neighbour technique do produce estimates which differ significantly among themselves.

From the results obtained, it appears that the methods which make use of an auxiliary variable to make a ratio adjustment are equivalent regardless of whether or not the auxiliary variable used is highly correlated with the variable to be imputed. In fact, the contrast analysis shows that, in most cases, methods 2 and 3, 6 and 7, and 8 and 9 are not significantly different. When they are significantly different, it was observed that the methods using the most highly correlated auxiliary variable produce estimates closer to the true population mean.

Among methods based on the nearest neighbour technique, it can be noted that the nearest neighbour alone is sufficient. Based on the results obtained, estimator 4 is never significantly different from the other methods which make use of the nearest neighbour technique. On the other hand, all methods using the nearest neighbour technique are not equivalent. Significant differences were sometimes observed in models B and C, when comparing estimators 5, 6 and 7.

When models B and C are considered, method 1, which consists of imputing the mean of the respondents, is observed to be always significantly different from the other methods and the estimates differ considerably from the true population mean. Consequently, estimator 1 will no longer be included in the discussion.

Continuing with models B and C, when variables X_3 and X_5 are considered, methods 2 and 3 produce estimates which are furthest from the true population mean. In addition, methods 3 and 7, 3 and 9, 2 and 6, and 2 and 8 are always significantly different in these situations; methods 8 and 9 always produce the best estimate of the true population mean. When the non-response rate is very high (for variable X_5) method 8 is significantly different from all the others. The same holds true for method 2.

Consequently, it appears that when models B and C are under consideration, those methods which use respondents produce estimates which are significantly different from those using the nearest neighbour technique. This difference increases as the non-response rate rises. The methods were observed to form four distinct groups (1), (2, 3), (4, 5, 6, 7) and (8, 9). Finally, recall that for model A, all methods produce similar results.

4.2 Bias, variance and mean square error

In order to compare the quality of the imputation methods under study, bias, variance and mean square error were calculated for all estimators, all variables and for the three models studied. The results are shown in tables C, D and E. A Student-t test can be performed to determine which estimator has a bias significantly different than zero. A Hartley test can be used to check for equality of variances of the estimators under study. Finally, an ANOVA of mean square errors will provide an indication of the variability of the various methods for each case considered.

(i) Bias

Tests of hypotheses were performed to test whether the mean values of the variables after imputation were significantly different than the true value. A t-test was used to perform the significance test. If the null hypothesis was rejected, the imputation estimator was regarded as biased. In general, based on the tests conducted, it was observed that the bias increases as the non-response rate increases.

Let us first consider model A. For variables X_1 and X_3, methods which make use of respondents tend to be less biased than those which use the nearest neighbour method. In fact, methods 1, 2 and 9 are unbiased and 3 and 8 have a relatively low bias, while all the methods using nearest neighbours are biased. In general, underestimation of the true population mean was observed. When the non-response rate is high (variable X_5), only methods 5 and 9 are biased.

For models B and C, method 1 is always biased. When variable X_1 is considered, it was observed that methods using nearest neighbours are always unbiased while those making use of respondents may be unbiased but tend to be more biased than the others. It should be noted that method 8 is always unbiased. As for variables X_3 and X_5 , methods 2 and 3 are always biased. Methods 8 and 9 are unbiased for variable X_3, while methods 4 and 5 are biased. When variable X_5 is considered, methods 8 and 9 are the only ones with little or no bias.

It is interesting that in model A (which has no non-response bias), X_1 is generally estimated worse than X_5, even though X_5 has more imputed values. The opposite, and expected, result occurs in models B and C. As far as bias is concerned, the estimators generally tend to form four groups. In fact, they can be grouped, in increasing order of bias, as follows: (8, 9), (4, 5, 6, 7), (2, 3) (1).

(ii) Variance

Recall that the Hartley test carried out for the variance analysis indicates that the variances of the imputation methods may be considered equal in all but one case. The variance of method 7 is significantly greater, in fact, than the others for variable X_5 of model B.

Let us first consider model A. For variables X_1 and X_3, it is observed that the methods using respondents tend to exhibit lower variance than those using the nearest neighbour method. For models B and C and for variable X_1, variances of the estimators which make use of nearest neighbours tend to be slightly higher than the others. Similarly, for variables X_3 and X_5, methods using respondents tend to have lower variance than those using nearest neighbours.

(iii) Mean square error

Since, in general, the equal variances hypothesis is not rejected, the difference between the mean square errors of the estimators studied is essentially a function of bias only. This explains why bias and mean square errors are observed to follow the same trends; see table E.

Let us first consider model A. For variables X_1 and X_3, it is not possible to conclude from the contrast analysis of mean square errors associated with each imputation method that there is a significant difference between the methods. However, it was noted that, as with bias, the mean square errors associated with the methods using respondents tended to be lower than those associated with methods using the nearest neighbour method. On the other hand, for variable X_5, method 9 was observed to have a significantly higher mean square error than the others.

Let us next consider models B and C. Mean square errors for method 1 are always significantly higher than the others. When method 1 is excluded, an ANOVA of the mean square errors shows that, for variable X_1, all methods are similar when model B is considered. Method 6 is observed to have the greatest mean square error and method 8 the smallest. It should also be noted that the mean square error of method 8 is significantly lower than that of method 6 when model C is considered. For variable X_3, methods 2 and 3 have a relatively high mean square error as compared to the others; for X_5, their mean square error is significantly higher than the others. When variable X_3 is studied, every method, with the exception of 1, 2 and 3, can be considered equivalent. Nonetheless, it is observed that methods 8 and 9 tend to have a lower mean square error than those using nearest neighbours. For variable X_5, methods 8 and 9 were observed to maintain a rather low mean square error. For model B, the mean square error of estimator 8 is significantly lower than that for estimator 6.

(iv) Conclusions

When model A is considered, methods 1, 2 and 9 are recommended for variables X_1 and X_3. Method 1 appears to be preferable for X_1, and 2 and 9 for X_3. For variable X_5, all imputation methods are recommended with the exception of 5 and 9.

Let us next consider models B and C. Method 1, which consists of imputing the mean of the respondents, is not recommended. For variable X_1, all imputation methods except method 1 are recommended. Note, however, that methods 2 and 3 may be biased and that methods making use of nearest neighbours tend to have a slightly higher variance than the others. Finally, method 8 performs well in these situations. As for variable X_3, methods 6, 7, 8 and 9 are recommended. Note, however, that methods 6 and 7 may be biased. In such cases, methods 8 and 9 are superior. For variable X_5, only methods 8 and 9 are recommended.

4.3 Correlation structure among variables

In the original population, the variables $X_1, \ldots, X_5$ are defined according to a certain predefined correlation structure. What is the situation after imputation? In order to see whether the correlational structure among the variables is maintained after imputation, Pearson's correlation coefficient was calculated for every possible pair of variables on the data sets after imputation. The calculation was performed for all nine imputation methods studied and for each model; see tables F, G and H. Fisher's transformation was then applied to each of the observations. This transformation stabilizes variance and makes the distribution of correlations approximately normal. A Student-t test was performed to determine if there is a significant difference between the correlations obtained after imputation and the original correlations. [A Student-t test was also performed directly on the correlations, that is, without the Fisher transformation, by assuming that the mean correlation is normally distributed and by using the variance of the mean over all replications. The conclusions obtained were identical in both cases].

In terms of correlation, the estimators associated with methods using respondents were much less effective than those associated with methods using the nearest neighbour method. In fact, the hypothesis that the correlation structure remains the same after imputation is always rejected for methods 1, 2, 3, 8 and 9. Methods using the nearest neighbour method retain this structure much better. It was also observed that the higher the non-response rate, the poorer the correlations are maintained.

Let us first look at the case of correlations between two variables that are not imputed. Recall that in this study, only variables X_2 and X_4 were not imputed. It was observed from the results obtained that the correlations r_{24} calculated after imputation are not identical with the starting correlations. These differences are due to sampling. However, it can be concluded from a Student-t test performed for r_{24} that these correlations are maintained, at the 1% level. However, at the 5% level, the r_{24} correlation is not retained for model C.

Second, let us consider the case of correlations between a variable which is imputed and one which is not. These correlations are r_{12}, r_{23}, r_{25}, r_{14}, r_{34}

and r_{45}. These correlations are relatively well preserved for methods using the nearest neighbour technique. Note that, for these methods, r_{14} is always maintained for models A and B. The dispersion of the calculated correlations is greatest for r_{45} and r_{25}. X_5 is the variable requiring the most imputation. The true value ρ_{45} is the highest, while ρ_{25} is also high relative to others. It is difficult to interpret the cause-effect relationships here.

Third, let us consider the case of correlations between two imputed variables. These correlations are r_{13}, r_{15} and r_{35}. These correlations are quite well preserved for methods using the nearest neighbour technique in model A.

Finally, it is observed that methods 1, 8 and 9 produce estimates of correlation coefficients which, for all practical purposes, are identical. Recall that methods 1, 8 and 9 consist of imputation using the respondents' mean (with adjustment ratio in 8 and 9). Let us also note that, for these three methods, the same value is imputed for each replication. In addition, these three methods always produce an underestimate.

Methods 2 and 3 tend to reduce the correlation between the imputed variables and those which are used for adjustment in these methods. In other words, for method 2, correlations r_{12}, r_{34} and r_{54} were observed to decrease while correlations r_{14}, r_{23} and r_{25} increase. The contrary holds true for method 3. Most of the time, methods 6 and 7 underestimate the correlation coefficient. Finally, method 4 appears to be most effective for maintaining correlations, and method 1, the worst.

5. CONCLUSIONS AND RECOMMENDATIONS

In summary, we can conclude that when donors and candidates are identically distributed and when the non-response rate is low or moderate, imputation methods using respondents are recommended for estimating means. When the non-response rate is high, imputation methods using the nearest neighbour technique improve so that in this case, all methods are recommended.

Let us now consider the case in which donors and candidates are not identically distributed. The method which consists of imputing the respondents' mean is not

recommended, regardless of the non-response rate. We note that all other imputation methods are recommended if the non-response rate is low. When the non-response rate is moderate, only those methods which consist of imputing the value of a nearest neighbour adjusted by an auxiliary variable and the methods consisting of imputing the respondents' mean adjusted by a weighting factor are recommended. Only those methods consisting of imputing using the respondents' mean adjusted by a weighting factor are recommended when the non-response rate is high and estimation of univariate means and totals only are desired.

As for the correlational structure among the variables, the conclusion is that methods using the nearest neighbour technique are to be especially recommended. For these methods, the correlation between two imputed variables is generally retained.

Based on the results obtained in this study, it does not appear that the choice of the auxiliary variable used in the ratio adjustment methods affects the results. In fact, the estimators using an auxiliary variable highly correlated with the variable requiring imputation do not give different results from those which use a less correlated auxiliary variable. It appears that as the non-response rate increases, the effectiveness of the estimators falls. However more tests are necessary using a different mix of correlational structure and non-response level in order to determine the extent that this observation holds.

Table I provides a quick overview of the various properties of each of the estimators, for the dataset used. It is important to note that the conclusions of this study apply only to similar datasets. The sensitivity of the results to such design considerations as the underlying distributions of the variables, correlational structure, non-response rates, population size and sample size has not been determined. Also, no edit constraints have been placed on the variables. A major problem with such empirical studies is that certain parameters must be fixed in order to keep the size of the study within manageable bounds. Therefore, one should generalize the conclusions with some caution.

TABLE A

TUKEY'S STUDENTIZED RANGE TEST

MODEL B

Variable X_1			Variable X_3			Variable X_5			
groups		estimators	groups		estimators	groups			estimators
	A	1		A	1			A	1
	B	3		B	6			B	2
	B			B					
C	B	8		B	2			C	3
C	B			B				C	
C	B	4		B	3		D	C	6
C	B			B			D	C	
C	B	5	C	B	7	E	D	C	5
C	B	true mean	C	B		E	D		
C	B	7	C	B	5	E	D		4
C	B		C	B		E			
C	B	6	C	B	4	E			9
C	B		C		true mean	E			
C	B	2	C		9	E			7
C			C						true mean
C		9	C		8		F		8

Estimators with the same letter are not significantly different.

TABLE B

TUKEY'S STUDENTIZED RANGE TEST

MODEL C

Variable X_1			Variable X_3			Variable X_5		
groups		estimators	groups		estimators	groups		estimators
	A	6		A	9		A	8
	A			A	true mean			true mean
B	A	4		A	8		B	9
B	A			A			B	
B	A	7		A	6	C	B	6
B	A			A		C	B	
B	A	5		A	7	C	B	7
B	A	true mean		A		C		
B	A	8	B	A	4	C	D	4
B	A		B				D	
B	A	9	B	C	5		D	5
B				C				
B		3		C	2		E	3
B				C			E	
B		2		C	3		E	2
		1		D	1		F	1

Estimators with the same letter are not significantly different.

TABLE C

BIAS

($\times 10^{-2}$)

ESTIMATORS	Variable X_1			Variable X_3			Variable X_5		
	Non-response Model			Non-response Model			Non-response Model		
	A	B	C	A	B	C	A	B	C
1	-1,40*	4,73**	-5,63**	-1,18	16,03**	-16,61**	-0,11	38,95**	-39,35**
2	-1,29	-0,40	-1,91*	-0,57	4,83**	-6,80**	1,08	20,24**	-23,90**
3	-2,07**	1,25	-1,76*	-1,90*	4,49**	-7,65**	-0,43	12,67**	-22,13**
4	-2,81**	0,08	1,33	-2,12*	3,20**	-2,06*	-0,75	7,06**	-7,89**
5	-3,11**	0,06	0,86	-2,47**	3,44**	-5,26**	-2,02*	8,61**	-11,38**
6	-2,74**	-0,11	1,78	-1,65	5,00**	-0,45	0,68	11,44**	-4,44**
7	-2,97**	-0,08	0,87	-2,50*	3,53**	-0,49	-0,82	4,84**	-5,46**
8	-2,28**	0,28	-0,11	-2,30**	-0,58	-0,34	-0,66	-4,66**	3,99**
9	-1,23	-1,71*	-0,36	-0,15	-0,12	1,40	2,43*	6,29**	-1,36

Note: (1) The presented bias values are the mean of the 25 replicates.

(2) X_1 has the lowest non-response rate, and X_5 has the highest non-response rate.

(3) * implies significant at 5% level.

** implies significant at 1% level.

TABLE D

ESTIMATED MEAN VALUES AFTER IMPUTATION

Imputation Estimator	Variable X_1 True Value = 1.032			Variable X_3 True Value = 1.483			Variable X_5 True Value = 2.170		
	Non-response Model			Non-response Model			Non-response Model		
	A	B	C	A	B	C	A	B	C
1	1.018*	1.079**	0.975**	1.471	1.644**	1.317**	2.169	2.560**	1.777**
2	1.019	1.028	1.012*	1.478	1.532**	1.415**	2.181	2.373**	1.931**
3	1.011**	1.044	1.014*	1.464*	1.528**	1.407**	2.166	2.297**	1.949**
4	1.004**	1.032	1.045	1.462*	1.515**	1.463*	2.163	2.241**	2.091**
5	1.000**	1.032	1.040	1.459**	1.518**	1.431**	2.150*	2.256**	2.056**
6	1.004**	1.030	1.049	1.467	1.533**	1.479	2.177	2.285**	2.126**
7	1.002**	1.031	1.040	1.458*	1.519**	1.478	2.162	2.219**	2.116**
8	1.009**	1.034	1.030	1.460**	1.477	1.480	2.164	2.124**	2.210**
9	1.019	1.014*	1.028	1.482	1.482	1.497	2.194*	2.233**	2.157

Note: (1) The presented mean values are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has lowest non-response rate, and
X_5 has the highest non-response rate.

(3) * implies significant difference from true value at 5% level.
** implies significant difference from true value at 1% level.

TABLE E

ESTIMATED MEAN SQUARE ERRORS

($\times 10^{-3}$)

Imputation Estimator	Variable X_1 Non-response Model			Variable X_3 Non-response Model			Variable X_5 Non-response Model		
	A	B	C	A	B	C	A	B	C
1	1.33	3.66	4.43	1.84	27.70	29.17	1.54	153.78	157.69
2	1.39	1.05	1.68	1.41	4.19	5.85	1.72	42.73	59.58
3	1.51	1.16	1.44	1.81	3.87	7.18	1.27	17.81	51.02
4	2.60	1.08	2.09	2.48	3.87	2.34	1.55	6.94	8.17
5	2.16	1.02	1.83	2.24	3.08	3.80	1.88	9.27	14.89
6	2.69	1.20	2.47	2.25	5.74	1.96	1.74	16.49	4.05
7	2.64	1.06	1.71	2.84	4.43	2.98	1.97	8.03	4.92
8	1.65	0.93	1.11	1.99	1.83	1.58	1.83	3.85	3.92
9	1.44	1.26	1.40	1.52	2.02	1.58	3.79	6.57	4.21

Note: (1) The variance component of MSE is calculated as the variance of theestimates from each of the 25 replicates, of the entire dataset after imputation. The bias component is calculated as the mean bias of the 25 replicates, of the entire dataset after imputation. For each replicate, the bias is calculated as the difference between the estimated value and the true value.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.

TABLE F

ESTIMATED CORRELATION COEFFICIENTS AFTER IMPUTATION

Non-Response Model A

Imputation Estimator	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
True Value	.763	.290	.244	.431	.452	.196	.570	.530	.671	.811
1	.625**	.211**	.207**	.333**	.370**	.195	.402**	.428**	.578**	.556**
2	.614**	.377**	.396**	.456**	.722**	.195	.831**	.373**	.834**	.412**
3	.782**	.250**	.230*	.290**	.349**	.195	.342**	.712**	.786**	.906**
4	.726**	.303	.238	.443	.496**	.195	.590**	.517	.667	.766**
5	.760	.342**	.248	.477**	.506**	.195	.599**	.576**	.727**	.794**
6	.722**	.282	.254	.405**	.498**	.195	.592**	.493**	.674	.730**
7	.729**	.299	.239	.427	.484**	.195	.563	.534	.676	.798*
8	.625**	.210**	.207**	.333**	.370**	.195	.402**	.429**	.578**	.556**
9	.625**	.211**	.207**	.333**	.370**	.195	.402**	.428**	.578**	.556**

Note: (1) The presented correlation coefficients are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.
X_2 and X_4 have no non-response.

(3) * implies significant at 5% level.
** implies significant at 1% level.

TABLE G

ESTIMATED CORRELATION COEFFICIENTS AFTER IMPUTATION

Non-Response Model B

Imputation Estimator	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
True Value	.763	.290	.244	.431	.452	.196	.570	.530	.671	.811
1	.679**	.214**	176**	.316**	.359**	.199	.370**	.417**	.537**	.585**
2	.706**	.358**	.325**	.458**	.603**	.199	.744**	.437**	.730**	.472**
3	.794**	.278*	.236	.316**	.387**	.199	.372**	.659**	.765**	.862**
4	.763	.290	.243	.415*	.447	.199	.545**	.545*	.660	.782**
5	.787**	.334**	.249	.453**	.478**	.199	.566	.566**	.702**	.816
6	.750**	.257**	.250	.344**	.413**	.199	.477**	.513	.659	.684**
7	.767	.273*	.246	.346**	.416**	.199	.425**	.531	.621*	.662**
8	.679**	.226**	.175**	.338**	.381**	.199	.413**	.440**	.575**	.647**
9	.680**	.228**	.177**	.338**	.381**	.199	.409**	.440**	.572**	.641**

Note: (1) The presented correlation coefficients are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.
X_2 and X_4 have no non-response.

(3) * implies significant at 5% level.
** implies significant at 1% level.

TABLE H

ESTIMATED CORRELATION COEFFICIENTS AFTER IMPUTATION

Non-Response Model C

Imputation Estimator	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
True Value	.763	.290	.244	.431	.452	.196	.570	.530	.671	.811
1	.638**	.187**	.182**	.264**	.343**	.183*	.328**	.387**	.550**	.476**
2	.631**	.417**	.380**	.468**	.721**	.183*	.842**	.357**	.812**	.354**
3	.802**	.265**	.212**	.264**	.325**	.183*	.281**	.758**	.813**	.926**
4	.754	.306	.215**	.415	.440	.183*	.555*	.516	.687*	.768**
5	.793**	.376**	.224*	.487**	.491**	.183*	.577	.591**	.756**	.792**
6	.745**	.290	.234	.350**	.448	.183*	.551*	.506*	.670	.713**
7	.763	.301	.220**	.398**	.424**	.183*	.523**	.547	.726**	.804
8	.643**	.208**	.188**	.309**	.363**	.183*	.391**	.406**	.587**	.544**
9	.643**	.211**	.187**	.305**	.368**	.183*	.387**	.411**	.590**	.540**

Note: (1) The presented correlation coefficients are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.
X_2 and X_4 have no non-response.

(3) * implies significant at 5% level.
** implies significant at 1% level.

TABLE G

ESTIMATED CORRELATION COEFFICIENTS AFTER IMPUTATION

Non-Response Model B

Imputation Estimator	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
True Value	.763	.290	.244	.431	.452	.196	.570	.530	.671	.811
1	.679**	.214**	176**	.316**	.359**	.199	.370**	.417**	.537**	.585**
2	.706**	.358**	.325**	.458**	.603**	.199	.744**	.437**	.730**	.472**
3	.794**	.278*	.236	.316**	.387**	.199	.372**	.659**	.765**	.862**
4	.763	.290	.243	.415*	.447	.199	.545**	.545*	.660	.782**
5	.787**	.334**	.249	.453**	.478**	.199	.566	.566**	.702**	.816
6	.750**	.257**	.250	.344**	.413**	.199	.477**	.513	.659	.684**
7	.767	.273*	.246	.346**	.416**	.199	.425**	.531	.621*	.662**
8	.679**	.226**	.175**	.338**	.381**	.199	.413**	.440**	.575**	.647**
9	.680**	.228**	.177**	.338**	.381**	.199	.409**	.440**	.572**	.641**

Note: (1) The presented correlation coefficients are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.
X_2 and X_4 have no non-response.

(3) * implies significant at 5% level.
** implies significant at 1% level.

TABLE H

ESTIMATED CORRELATION COEFFICIENTS AFTER IMPUTATION

Non-Response Model C

Imputation Estimator	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
True Value	.763	.290	.244	.431	.452	.196	.570	.530	.671	.811
1	.638**	.187**	.182**	.264**	.343**	.183*	.328**	.387**	.550**	.476**
2	.631**	.417**	.380**	.468**	.721**	.183*	.842**	.357**	.812**	.354**
3	.802**	.265**	.212**	.264**	.325**	.183*	.281**	.758**	.813**	.926**
4	.754	.306	.215**	.415	.440	.183*	.555*	.516	.687*	.768**
5	.793**	.376**	.224*	.487**	.491**	.183*	.577	.591**	.756**	.792**
6	.745**	.290	.234	.350**	.448	.183*	.551*	.506*	.670	.713**
7	.763	.301	.220**	.398**	.424**	.183*	.523**	.547	.726**	.804
8	.643**	.208**	.188**	.309**	.363**	.183*	.391**	.406**	.587**	.544**
9	.643**	.211**	.187**	.305**	.368**	.183*	.387**	.411**	.590**	.540**

Note: (1) The presented correlation coefficients are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.
X_2 and X_4 have no non-response.

(3) * implies significant at 5% level.
** implies significant at 1% level.

TABLE A

TUKEY'S STUDENTIZED RANGE TEST

MODEL B

Variable X_1			Variable X_3			Variable X_5			
groups		estimators	groups		estimators	groups			estimators
	A	1		A	1			A	1
	B	3		B	6			B	2
	B			B					
C	B	8		B	2			C	3
C	B			B				C	
C	B	4		B	3		D	C	6
C	B			B			D	C	
C	B	5	C	B	7	E	D	C	5
C	B	true mean	C	B		E	D		
C	B	7	C	B	5	E	D		4
C	B		C	B		E			
C	B	6	C	B	4	E			9
C	B		C		true mean	E			
C	B	2	C		9	E			7
C			C						true mean
C		9	C		8		F		8

Estimators with the same letter are not significantly different.

TABLE B

TUKEY'S STUDENTIZED RANGE TEST

MODEL C

Variable X_1			Variable X_3			Variable X_5		
groups		estimators	groups		estimators	groups		estimators
	A	6		A	9		A	8
	A			A	true mean			true mean
B	A	4		A	8		B	9
B	A			A			B	
B	A	7		A	6	C	B	6
B	A			A		C	B	
B	A	5		A	7	C	B	7
B	A	true mean		A		C		
B	A	8	B	A	4	C	D	4
B	A		B				D	
B	A	9	B	C	5		D	5
B				C				
B		3		C	2		E	3
B				C			E	
B		2		C	3		E	2
		1		D	1		F	1

Estimators with the same letter are not significantly different.

TABLE C

BIAS

($\times 10^{-2}$)

ESTIMATORS	Variable X_1			Variable X_3			Variable X_5		
	Non-response Model			Non-response Model			Non-response Model		
	A	B	C	A	B	C	A	B	C
1	-1,40*	4,73**	-5,63**	-1,18	16,03**	-16,61**	-0,11	38,95**	-39,35**
2	-1,29	-0,40	-1,91*	-0,57	4,83**	-6,80**	1,08	20,24**	-23,90**
3	-2,07**	1,25	-1,76*	-1,90*	4,49**	-7,65**	-0,43	12,67**	-22,13**
4	-2,81**	0,08	1,33	-2,12*	3,20**	-2,06*	-0,75	7,06**	-7,89**
5	-3,11**	0,06	0,86	-2,47**	3,44**	-5,26**	-2,02*	8,61**	-11,38**
6	-2,74**	-0,11	1,78	-1,65	5,00**	-0,45	0,68	11,44**	-4,44**
7	-2,97**	-0,08	0,87	-2,50*	3,53**	-0,49	-0,82	4,84**	-5,46**
8	-2,28**	0,28	-0,11	-2,30**	-0,58	-0,34	-0,66	-4,66**	3,99**
9	-1,23	-1,71*	-0,36	-0,15	-0,12	1,40	2,43*	6,29**	-1,36

Note: (1) The presented bias values are the mean of the 25 replicates.

(2) X_1 has the lowest non-response rate, and X_5 has the highest non-response rate.

(3) * implies significant at 5% level.

** implies significant at 1% level.

TABLE D

ESTIMATED MEAN VALUES AFTER IMPUTATION

Imputation Estimator	Variable X_1 True Value = 1.032			Variable X_3 True Value = 1.483			Variable X_5 True Value = 2.170		
	Non-response Model			Non-response Model			Non-response Model		
	A	B	C	A	B	C	A	B	C
1	1.018*	1.079**	0.975**	1.471	1.644**	1.317**	2.169	2.560**	1.777**
2	1.019	1.028	1.012*	1.478	1.532**	1.415**	2.181	2.373**	1.931**
3	1.011**	1.044	1.014*	1.464*	1.528**	1.407**	2.166	2.297**	1.949**
4	1.004**	1.032	1.045	1.462*	1.515**	1.463*	2.163	2.241**	2.091**
5	1.000**	1.032	1.040	1.459**	1.518**	1.431**	2.150*	2.256**	2.056**
6	1.004**	1.030	1.049	1.467	1.533**	1.479	2.177	2.285**	2.126**
7	1.002**	1.031	1.040	1.458*	1.519**	1.478	2.162	2.219**	2.116**
8	1.009**	1.034	1.030	1.460**	1.477	1.480	2.164	2.124**	2.210**
9	1.019	1.014*	1.028	1.482	1.482	1.497	2.194*	2.233**	2.157

Note: (1) The presented mean values are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has lowest non-response rate, and
X_5 has the highest non-response rate.

(3) * implies significant difference from true value at 5% level.
** implies significant difference from true value at 1% level.

TABLE E

ESTIMATED MEAN SQUARE ERRORS

($\times 10^{-3}$)

Imputation Estimator	Variable X_1 Non-response Model			Variable X_3 Non-response Model			Variable X_5 Non-response Model		
	A	B	C	A	B	C	A	B	C
1	1.33	3.66	4.43	1.84	27.70	29.17	1.54	153.78	157.69
2	1.39	1.05	1.68	1.41	4.19	5.85	1.72	42.73	59.58
3	1.51	1.16	1.44	1.81	3.87	7.18	1.27	17.81	51.02
4	2.60	1.08	2.09	2.48	3.87	2.34	1.55	6.94	8.17
5	2.16	1.02	1.83	2.24	3.08	3.80	1.88	9.27	14.89
6	2.69	1.20	2.47	2.25	5.74	1.96	1.74	16.49	4.05
7	2.64	1.06	1.71	2.84	4.43	2.98	1.97	8.03	4.92
8	1.65	0.93	1.11	1.99	1.83	1.58	1.83	3.85	3.92
9	1.44	1.26	1.40	1.52	2.02	1.58	3.79	6.57	4.21

Note: (1) The variance component of MSE is calculated as the variance of theestimates from each of the 25 replicates, of the entire dataset after imputation. The bias component is calculated as the mean bias of the 25 replicates, of the entire dataset after imputation. For each replicate, the bias is calculated as the difference between the estimated value and the true value.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.

TABLE F

ESTIMATED CORRELATION COEFFICIENTS AFTER IMPUTATION

Non-Response Model A

Imputation Estimator	r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
True Value	.763	.290	.244	.431	.452	.196	.570	.530	.671	.811
1	.625**	.211**	.207**	.333**	.370**	.195	.402**	.428**	.578**	.556**
2	.614**	.377**	.396**	.456**	.722**	.195	.831**	.373**	.834**	.412**
3	.782**	.250**	.230*	.290**	.349**	.195	.342**	.712**	.786**	.906**
4	.726**	.303	.238	.443	.496**	.195	.590**	.517	.667	.766**
5	.760	.342**	.248	.477**	.506**	.195	.599**	.576**	.727**	.794**
6	.722**	.282	.254	.405**	.498**	.195	.592**	.493**	.674	.730**
7	.729**	.299	.239	.427	.484**	.195	.563	.534	.676	.798*
8	.625**	.210**	.207**	.333**	.370**	.195	.402**	.429**	.578**	.556**
9	.625**	.211**	.207**	.333**	.370**	.195	.402**	.428**	.578**	.556**

Note: (1) The presented correlation coefficients are the means of the 25 replicates, of the entire dataset after imputation.

(2) X_1 has the lowest non-response rate, and
X_5 has the highest non-response rate.
X_2 and X_4 have no non-response.

(3) * implies significant at 5% level.
** implies significant at 1% level.

Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

WORKING PAPER NO. BSMD-87-002

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-87-002

MÉTHODOLOGIE

COMPARAISON DE DIFFÉRENTES MÉTHODES D'IMPUTATION POUR DONNÉES QUANTITATIVES

par

Marcel Bureau
Sylvie Michaud
Madhavi Sistla

DMEE

3 novembre 1986

* Il s'agit d'une version préliminaire. Ne pas citer sans la permission des auteurs. Nous aimerions recevoir vos commentaires.

1. OBJECTIFS DE L'ÉTUDE

Le but du présent projet est de comparer, à l'aide d'une étude par simulation, la qualité des estimateurs associés à différentes méthodes d'imputation. Cette comparaison est faite en calculant le biais, la variance et l'erreur quadratique moyenne de chacun des estimateurs considérés. De plus le calcul après imputation du coéfficient de corrélation entre chaque variable permet de vérifier si une méthode d'imputation donnée préserve la structure de corrélation entre les variables.

N.B. La simulation et l'imputation ont été exécutées par Madhavi Sistla, une étudiante du programme COSEP. Pour plus de détails, voir son rapport intitulé: "A comparison of different imputation techniques", BSMD, Août 1986. Les grandes lignes de ce document sont rapportées ici dans le but de présenter un seul rapport couvrant entièrement cette étude.

2. MÉTHODOLOGIE DE LA SIMULATION

La population considérée est générée artificiellement. Elle comporte cinq variables ($X_1, \ldots, X_5$) et elle est de taille 1000. Chaque variable est une combinaison linéaire de variables exponentielles indépendantes de moyenne 1,0. Les variables sont définies de la façon suivante:

$$X_1 = 1{,}0 * E_1 \quad ,$$

$$X_2 = 0{,}75 * E_1 + 0{,}661438 * E_2 \quad ,$$

$$X_3 = 0{,}25 * E_1 + 0{,}321278 * E_2 + 0{,}913392 * E_3 \quad ,$$

$$X_4 = (0{,}20 * E_1 - 0{,}075593 * E_2 + 0{,}519257 * E_3 + 0{,}827440 * E_4) \quad ,$$

$$X_5 = (0{,}40 * E_1 + 0{,}302372 * E_2 + 0{,}495797 * E_3 + 0{,}586642 * E_4 + 0{,}398257 * E_5) \quad ,$$

où les E_i, $(i=1, \ldots, 5)$, sont des variables exponentielles de moyenne 1,0. Les constantes ont été choisies de façon à assurer que les corrélations entre les variables $X_1, \ldots, X_5$ suivent approximativement une structure donnée. Les corrélations se situent entre 0,2 et 0,8: voir la première ligne du tableau F.

2.2 Création de la non-réponse

Trois modèles de non-réponse sont considérés. Pour chacun de ces modèles, 500 unités ne présentent aucune non-réponse (les donneurs) et 500 unités ont de la non-réponse partielle (les candidats). Parmi les candidats, on génère aléatoirement et de façon indépendante de la non-réponse pour les variables X_1, X_3 et X_5. Le taux de non-réponse de ces variables pour la sous-population des candidats est fixé à 50%, 75% et 100% respectivement. Par conséquent, le taux de non-réponse global est de 25%, 37.5% et 50% respectivement. Les variables X_2 et X_4 sont toujours présentes.

Les trois modèles sont différents par leur façon de déterminer la population de donneurs et de candidats. Le premier modèle (modèle A) présente le cas où les répondants et les non-répondants ont la même distribution (c'est-à-dire qu'il n'y a pas de biais dû à la non-réponse); 500 unités sont choisies aléatoirement parmi 1000 unités pour former la sous-population de candidats. Les 500 unités qui restent constituent la sous-population de donneurs. Le second et le troisième modèles présentent les cas où la distribution des donneurs et des candidats diffèrent. Dans le second modèle (modèle B) 125 candidats sont choisis parmi les unités qui ont une valeur de X_5 supérieure à la médiane de X_5 et 375, parmi celles qui ont une valeur de X_5 inférieure à la médiane de X_5. Les 500 unités qui restent constituent la sous-population de donneurs. Et inversement dans le troisième modèle (modèle C) 125 candidats sont choisis parmi les unités ayant une valeur

de X_5 inférieure à la médiane de X_5 et 375, parmi celles ayant une valeur de X_5 supérieure à la médiane de X_5. Les 500 unités qui restent constituent la sous-population de donneurs. Les modèles B et C représentent les cas où la variable X_5 est biaisée vers le haut et vers le bas respectivement. C'est-à-dire qu'une estimation basée sur les répondants seulement, qui ne serait pas pondérée, serait biaisée vers le haut et vers le bas respectivement.

2.3 Échantillonnage et répétitions

Chacun des trois modèles considérés donne une population de 1000 unités; 500 donneurs et 500 candidats. Dans chacune des trois populations, un échantillon aléatoire simple de 500 unités est sélectionné. Toutes les méthodes d'imputation sont alors appliquées sur chacune des variables ayant de la non-réponse et des estimés sont calculés. Le processus de sélection d'un échantillon, d'imputation et d'estimation est répété 25 fois pour chaque modèle.

3. MÉTHODES D'IMPUTATION CONSIDÉRÉES

3.1 Préliminaires

Les méthodes d'imputation considérées peuvent être classées dans un des deux groupes suivants. Premièrement les méthodes utilisant l'information fournie par les répondants et deuxièmement, les méthodes faisant appel à la technique du plus proche voisin. Brièvement, la technique de plus proche voisin consiste à trouver, pour un candidat donné, le donneur qui lui ressemble le plus (en terme de fonction de distance). Pour ce faire une transformation uniforme a été appliquée sur les données de façon à ce quelles soient distribuées entre 0 et 1. Cette transformation permet d'éliminer l'effet d'amplitude dû à la grande dispersion des données. La technique du plus proche voisin requiert l'utilisation de variables d'appariement.

Une variable d'appariement est définie comme étant une variable qui est correlée avec la variable nécessitant imputation et qui est présente à la fois pour le donneur et pour le candidat. Pour cette étude, la fonction de distance minimax (L_∞ norme) est utilisée. La distance entre un candidat et un donneur est donc déterminée de la façon suivante:

$$D(X_j, Y) = \max_i |x_{ij} - y_i| \quad ,$$

où Y est le candidat, X_j est le j^e donneur et i est un indice désignant les variables d'appariement. Il faut noter que la technique du plus proche voisin consiste à prendre le minimum, sur l'ensemble de donneurs j , de la distance D (minimax).

3.2 Estimateurs considérés

Notons Y la variable nécessitant imputation (X_1, X_3 ou X_5),
Z la variable auxiliaire (X_2, X_4),
c l'indice désignant un candidat,
d l'indice désignant un donneur,
r l'indice désignant un répondant(1),
n_r le nombre de répondants pour la variable Y,
n le nombre de donneurs,
N la taille de la population,
(i) l'indice désignant le i^e plus proche voisin,

et soit
$$\bar{Y}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} y_i \quad ,$$

---

(1) La différence entre un donneur et un répondant est que le répondant est défini séparément pour chaque variable tandis qu'un donneur est défini pour l'unité toute entière. Le donneur est donc une unité pour laquelle il y a une réponse pour toutes les variables.

:

$$\bar{Z}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} z_i \quad ,$$

$$\bar{Z} = \frac{1}{N} \sum_{i=1}^{N} z_i \quad .$$

Les techniques d'imputation suivantes ont été étudiées:

1 Imputation par la moyenne des répondants

$$Y_c = \bar{Y}_r \quad ,$$

2 et 3 Imputation de la moyenne des répondants, ajustée par le ratio

$$Y_c = \bar{Y}_r \cdot \frac{Z_c}{\bar{Z}_r} \quad ,$$

4 Imputation de la valeur du plus proche voisin

$$Y_c = Y_{d(1)} \quad ,$$

5 Imputation de la moyenne des valeurs des cinq plus proches voisins

$$Y_c = \frac{1}{5} \sum_{i=1}^{5} Y_{d(i)} \quad ,$$

6 et 7 Imputation de la valeur du plus proche voisin, ajustée par le ratio.

$$Y_c = Y_{d(1)} \cdot \frac{Z_c}{Z_{d(1)}} ;$$

8 et 9 Imputation de la moyenne des répondants, ajustée par pondération

$$Y_c = \bar{Y}_r \cdot \frac{\bar{Z}}{\bar{Z}_r} \quad .$$

Il faut noter que les trois méthodes consistant à faire un ajustement à l'aide d'une variable auxiliaire sont appliquées une première fois avec

la variable auxiliaire la plus corrélée (estimateurs 2 , 6 et 8) et ensuite avec la variable auxiliaire la moins corrélée (estimateurs 3 , 7 et 9).

Variable nécessitant imputation:	X_1	X_3	X_5
Variable auxiliaire la plus corrélée:	X_4	X_2	X_2
Variable auxiliaire la moins corrélée:	X_2	X_4	X_4

Notons que les estimateurs 8 et 9 sont des estimateurs par le quotien. La même valeur est imputée pour tous les candidats. De même, pour l'estimateur 1, la même valeur est imputée pour tous les candidats.

Les estimations de la moyenne après imputation sont obtenues de la façon suivante:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n_r} y_{di} + \sum_{i=n_r+1}^{N} y_{ci}}{N} .$$

Soit, $\hat{\bar{Y}}$ l'estimation de la moyenne,
$\bar{Y}$ la moyenne de la population,
m le nombre de répétitions (25 pour cette étude).

La variance, le biais et l'erreur quadratique moyenne de chacun des estimateurs peuvent être obtenus de la façon suivante:

$$\text{Variance} = V(\hat{\bar{Y}}) = \frac{\sum_{i=1}^{m} [\hat{\bar{Y}}_i - (\sum_{i=1}^{m} \frac{\hat{\bar{Y}}_i}{m})]^2}{m-1}$$

$$\text{Biais} = B(\hat{\bar{Y}}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{\bar{Y}}_i - \bar{Y})$$

$$\text{erreur quadratique moyenne} = V(\hat{\bar{Y}}) + B(\hat{\bar{Y}})^2$$

Pour chacune des 25 répétitions de la simulation, une estimation de la variance est obtenue de la façon suivante:

$$\text{i)} \quad V(\hat{\bar{Y}}) = \frac{(N-n_r)}{N} * \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{(y_{ri} - \bar{y}_r)^2}{n_r - 1}$$

ou

$$\text{ii)} \quad V(\hat{\bar{Y}}) = \frac{(N-n_r)}{N} * \frac{1}{n_r} \sum_{i=1}^{n_r} \frac{(y_{ri} - \hat{R}\, x_{ri})^2}{n_r - 1},$$

pour les estimateurs par la quotient.

La moyenne de ces variances peut être comparée à la variance des estimations.

4. PRÉSENTATION ET DISCUSSION DES RÉSULTATS

4.1 Analyse de la variance

Pour comparer les méthodes d'imputation entre elles il a été décidé de tester l'hypothèse d'égalité des moyennes à l'aide d'une analyse de la variance. L'analyse de la variance est faite séparément pour chacune des trois variables ayant été imputées et pour chacun des trois modèles considérés dans la simulation. L'unité d'observation est la moyenne de la variable imputée pour les candidats et les donneurs. Rappelons que le test d'analyse de la variance classique requiert trois présupposés de base. Premièrement on doit avoir normalité des observations; c'est-à-dire qu'on doit s'assurer que les moyennes associées aux 25 répétitions pour une méthode d'imputation donnée sont normalement distribuées. Pour ce faire, un test de normalité de Shapiro-Wilk a été effectué sur chacune des trois variables imputées, pour chacune des 9 méthodes et pour les trois modèles considérées.

Ces tests permettent de conclure que l'hypothèse de normalité est en général respectée. En effet sur 81 cas vérifiés on rejette l'hypothèse de normalité dans 3 cas seulement en utilisant un niveau de signification de 5%. Ceci correspond au nombre de rejets auquel on pouvait s'attendre. Deuxièmement, on doit s'assurer que la variance des observations pour chaque méthode d'imputation est constante (la variance des 25 moyennes). Un test d'égalité de variances de Hartley a été effectué et, au niveau de signification de 5%, on peut conclure à l'égalité des variances dans tous les cas sauf un. Finalement, si on assume l'indépendance entre les différentes méthodes d'imputation, on peut admettre que tous les présupposés requis pour l'analyse de la variance sont respectés. C'est-à-dire qu'on suppose que, pour un même ensemble de données, les résultats provenant d'une méthode d'imputation donnée sont indépendants de ceux d'une autre méthode. Notons cependant qu'il est peu probable que ce présupposé soit tout à fait vrai.

On peut maintenant procéder à l'analyse de la variance proprement dite. On veut tester l'hypothèse H_o: il y a égalité entre les moyennes obtenues pour chacune des 9 méthodes d'imputation. Dans le cas où l'hypothèse H_o est rejetée une analyse des contrastes simples selon la méthode T de Tukey permet de déterminer qu'elles méthodes sont significativement différentes.

Si on utilise un niveau de signification de 5% on constate que pour le modèle A on ne peut rejeter H_o pour les variables X_1 et X_3 et on doit conclure qu'il n'y a pas de différences significatives entre les différentes méthodes d'imputation. Par contre l'hypothèse de l'égalité entre les méthodes est rejetée au niveau de signification de 5% dans le cas de la variable X_5. Cependant H_o n'est pas rejetée au niveau de 1%. L'étude des contrastes permet de voir que la méthode 9 produit une estimation significativement supérieure à la méthode 5 au niveau de signification de 5%.

technique du plus proche voisin ne sont pas équivalentes. Des différences significatives ont parfois été observées dans les modèles B et C, lorsque l'on compare les estimateurs 5, 6 et 7.

Lorsque les modèles B et C sont considérés on constate que la méthode 1 consistant à imputer la moyenne des répondants est toujours significativement différente des autres méthodes et l'estimé s'éloigne considérablement de la vraie moyenne de la population. Par conséquent, à partir de maintenant, l'estimateur 1 ne fera plus partie de la discussion.

Pour les modèles B et C toujours, lorsque les variables X_3 et X_5 sont considérées les méthodes 2 et 3 produisent des estimations qui sont les plus éloignées de la vraie moyenne de la population. De plus les méthodes 3 et 7, 3 et 9, 2 et 6 et 2 et 8 sont toujours significativement différentes dans ces situations; les méthodes 8 et 9 étant toujours celles qui estiment le mieux la vraie moyenne de la population. Lorsque le taux de non-réponse devient très élevé la méthode 8 est significativement différente de toutes les autres, pour la variable X_5. Il en est de même pour la méthode 2.

Par conséquent, il semble que lorsque les modèles B et C sont envisagées les méthodes utilisant les répondants produisent des estimations sensiblement différentes de celles utilisant la technique du plus proche voisin. Cette différence augmente à mesure que le taux de non-réponse s'élève. Il a été constaté que les méthodes ont tendance à former 4 groupes distincts (1), (2,3), (4,5,6,7) et (8,9). Finalement rappelons que pour le modèle A toutes les méthodes sont équivalentes.

4.2 Biais, variance et erreur quadratique moyenne

Dans le but de comparer la qualité des méthodes d'imputation considérées, le biais, la variance et l'erreur quadratique moyenne ont été calculés pour tout les estimateurs, pour toutes les variables et

pour les trois modèles considérés. Les résultats sont présentés dans les tableaux C, D et E. Un test t de Student permet de savoir quel estimateur possède un biais significativement différent de zéro. Un test de Hartley permet de vérifier l'égalité des variances des estimateurs considérés. Finalement, une analyse de la variance des erreurs quadratiques moyennes permet d'avoir une idée de la variabilité des différentes méthodes pour chacun des cas considérés.

(i) Biais

Puisque les observations peuvent être considérées comme étant normalement distribuées alors les biais sont aussi normalement distribués. L'égalité des variances pour chacune des méthodes permet de conclure que la variance des biais est aussi constante. Par conséquent, il semble que les présupposés nécessaire à un test t de Student sont respectés. De façon générale, à partir des tests effectués, il a été observé que plus le taux de non-réponse augmente plus le biais des estimateurs augmente.

Considérons premièrement le modèle A. Lorsque les variables X_1 et X_3 sont envisagées, les méthodes utilisant les répondants ont tendance à avoir un biais plus faible que les méthodes faisant appel à la technique du plus proche voisin. En effet, les méthodes 1, 2 et 9 sont non biaisées et 3 et 8 ont un biais relativement faible tandis que les méthodes utilisant les plus proches voisins sont toutes biaisées. En général, on remarque qu'il y a sous-estimation de la vraie moyenne de la population. Lorsque le taux de non-réponse est élevé, (variable X_5), seules les méthodes 5 et 9 sont biaisées.

Pour les modèles B et C la méthode 1 est toujours biaisée. Lorsqu'on considère la variable X_1 on remarque que les méthodes qui utilisent les plus proches voisins sont toujours sans biais tandis que celles qui font appel aux répondants peuvent être non

biaisées mais ont tendance à avoir un biais plus élevé que les autres. Il est à noter que la méthode 8 est toujours sans biais. Pour ce qui est des variables X_3 et X_5 on remarque que les méthodes 2 et 3 sont toujours biaisées. Pour la variable X_3, les méthodes 8 et 9 sont sans biais alors que 4 et 5 sont biaisées. Lorsqu'on considère la variable X_5 on constate que seules les méthodes 8 et 9 sont sans biais ou ont un biais faible.

Il est intéressant de remarquer que pour le modèle A (pour lequel il n'y a pas de biais dû à la non-réponse, la variable X_1 est en général moins bien estimée que la variable X_5, malgré que X_5 ait été imputée plus souvent. Dans le cas des modèles B et C, c'est le contraire. En effet, comme on pouvait s'y attendre, la variable X_5 est moins bien estimée que la variable X_1. En général on constate que, pour ce qui est du biais, les estimateurs ont tendance à former 4 groupes. En effet on peut classer les estimateurs par ordre croissant, en terme de biais, de la façon suivante: (8,9), (4,5,6,7), (2,3), (1).

(ii) Variance

Rappelons que le test de Hartley effectué pour l'analyse de la variance indique que les variances des méthodes d'imputation peuvent être considérées comme égales dans tous les cas sauf un. En effet la variance de la méthode 7 est significativement supérieure aux autres pour la variable X_5 du modèle B.

Considérons premièrement le modèle A. Lorsque les variables X_1 et X_3 sont envisagées on remarque que les méthodes utilisant les répondants ont tendance à avoir une variance plus faible que celles utilisant la technique du plus proche voisin. Pour les modèle B et C et pour la variable X_1, les variances des estimateurs faisant appel aux plus proches voisins ont tendance a être légèrement plus élevées que les autres. De même, pour les

variables X_3 et X_5, on remarque que les méthodes utilisant les répondants ont tendance à avoir une variance plus faible que les méthodes utilisant les plus proches voisins.

(iii) Erreur quadratique moyenne

Puisque l'hypothèse de l'égalité des variances n'est, en général, pas rejetée la différence entre les erreurs quadratiques moyennes des estimateurs considérés ne dépend essentiellement que du biais. C'est pourquoi on constate que les biais et les erreurs quadratiques moyennes suivent les mêmes tendances; voir tableau E.

Considérons premièrement le modèle A. Pour les variables X_1 et X_3, l'analyse des contrastes des erreurs quadratiques moyennes associées à chacune des méthodes d'imputation ne permet pas de conclure qu'il y a une différence significative entre les méthodes. Cependant, on remarque que, comme pour les biais, les erreurs quadratiques moyennes associées aux méthodes faisant appel aux répondants ont tendance à être plus faibles que celles associées aux méthodes utilisant la technique du plus proche voisin. Par contre, pour la variable X_5, on constate que la méthode 9 a une erreur quadratique moyenne significativement supérieure aux autres.

Deuxièmement considérons les modèle B et C. Les erreurs quadratiques moyennes de la méthode 1 sont toujours significativement supérieures aux autres. En excluant la méthode 1 et pour la variable X_1, l'analyse de la variance sur les erreurs quadratiques moyennes montre que toutes les méthodes sont semblables lorsqu'on considère le modèle B. On constate que la méthode 6 possède la plus grande erreur quadratique moyenne et la méthode 8 la plus petite. Il faut noter également que l'erreur quadratique moyenne de 8 devient significativement

inférieure à celle de la méthode 6 lorsque le modèle C est considéré. Pour la variable X_3, les méthodes 2 et 3 ont une erreur quadratique moyenne relativement élevée par rapport aux autres; pour X_5 elles ont une erreur quadratique moyenne significativement supérieure aux autres. Lorsque la variable X_3 est considérée on constate que toutes les méthodes, sauf 1,2 et 3, peuvent être considérées comme équivalentes. On note tout de même que les méthodes 8 et 9 ont tendance à avoir une erreur quadratique moyenne plus petite que celles utilisant les plus proches voisins. Pour la variable X_5 on remarque que les méthodes 8 et 9 maintiennent une erreur quadratique moyenne assez faible. Pour le modèle B on constate que l'erreur quadratique moyenne de l'estimateur 8 est significativement inférieure à celle de 6.

(iv) Conclusions

Lorsque le modèle A est envisagé on constate que les méthodes 1, 2 et 9 sont à recommander pour les variables X_1 et X_3. La méthode 1 semble préférable pour X_1 et 2 et 9 pour X_3. Pour la variable X_5, toutes les méthodes d'imputation sont recommandées sauf 5 et 9.

Deuxièmement considérons les modèles B et C. La méthode 1 consistant à imputer la moyenne des répondants n'est jamais recommandée. Pour la variable X_1, toutes les méthodes d'imputation sont recommandées sauf la méthode 1. Notons cependant que les méthodes 2 et 3 peuvent être biaisées et que les méthodes recourant aux plus proches voisins ont tendance à avoir une variance légèrement plus élevée que les autres. Finalement la méthode 8 est très bonne dans ces situations. Pour ce qui est de la variable X_3 il semble que les méthodes 6, 7, 8 et 9 sont recommandables. Il faut noter que les méthodes 6 et 7 peuvent être biaisées. Les méthodes 8 et 9 sont meilleures

dans ce cas. Si on considère la variable X_5, seules les méthodes 8 et 9 sont recommandées.

4.3 Structure de corrélation entre les variables

Dans la population originale les variables $X_1, \ldots, X_5$ possèdent une certaine structure de corrélation. Qu'en est-il après imputation? Dans le but de savoir si la structure de corrélation entre les variables est conservée après imputation le coéfficient de corrélation de Pearson a été calculé pour toutes les paires de variables possibles. Ce calcul a été effectué pour chacune des 9 méthodes d'imputation considérées et pour chaque modèle; voir tableaux F, G et H. La transformation de Fisher a été appliquée sur chacune des observations. Cette transformation permet de stabiliser la variance et rend la distribution des corrélations approximativement normale. Un test t de Student a été effectué pour savoir s'il y a une différence significative entre les corrélations obtenues après imputation et les corrélations originales. [Un test t de Student a également été effectué directement sur les corrélations, c'est-à-dire sans transformation de Fisher, en assumant que la moyenne des corrélations est normalement distribuée et en utilisant la variance de la moyenne sur 25 répétitions. Les conclusions obtenues sont identiques].

En terme de corrélation les estimateurs associés aux méthodes utilisant les répondants sont beaucoup moins efficaces que ceux associés aux méthodes utilisant la technique du plus proche voisin. En effet, l'hypothèse du maintient de la structure de corrélation après imputation est toujours rejetée pour les méthodes 1, 2, 3, 8 et 9. Les méthodes utilisant la technique du plus proche voisin maintiennent beaucoup mieux cette structure. On remarque également que plus le taux de non-réponse augmente, moins les corrélations sont conservées.

Premièrement regardons le cas des corrélations entre deux variables n'ayant pas été imputées. Rappelons que dans cette étude seules les

variables X_2 et X_4 n'ont pas été imputées. D'après les résultats obtenus on constate que les corrélations r_{24} calculées après imputation ne sont pas identiques aux corrélations originales. Ces différences sont dues à l'échantillonnage. Cependant le test t de Student effectuer pour r_{24} permet de conclure que ces corrélations sont conservées, au niveau 1%. Par contre, au niveau 5%, la corrélation r_{24} n'est pas conservée pour le niveau C.

Deuxièmement considérons le cas des corrélations entre une variable imputée et une variable non imputée. Il s'agit des corrélations r_{12}, r_{23}, r_{25}, r_{14}, r_{34} et r_{45}. Ces corrélations sont assez bien conservées pour les méthodes utilisant la technique du plus proche voisin. Notons que, pour ces méthodes, r_{14} est toujours conservée pour les modèles A et B. Notons également que les coéfficients de corrélations calculés r_{45} et r_{25} possèdent la plus grande dispersion. Rappelons que la variable X_5 est celle qui a été imputée le plus souvent. Notons finalement que les coéfficients de corrélation dans la population ρ_{45} et ρ_{25} sont très élevés par rapport aux autres. Peut-être y a-t-il une relation de cause à effet.

Troisièmement considérons le cas des corrélations entre deux variables ayant été imputées. Il s'agit des corrélations r_{13}, r_{15} et r_{35}. On remarque que ces corrélations sont très bien conservées pour les méthodes utilisant la technique du plus proche voisin dans le modèle A.

Finalement, on constate que les méthodes 1, 8 et 9 produisent des estimations des coéfficients de corrélation qui sont à toute fin pratique identiques. Rappelons que les méthodes 1, 8 et 9 consistent à faire l'imputation à l'aide de la moyenne des répondants (ajustée par le ratio pour 8 et 9). Mentionnons également que, pour ces trois méthodes, la même valeur est toujours imputée pour chacune des répétitions. De plus ces trois méthodes produisent toujours une sous estimation.

Les méthodes 2 et 3 cnt tendance à faire diminuer la corrélation entre les variables qui sont imputées et celles qui servent d'ajustement pour ces méthodes. C'est-à-dire que pour la méthode 2 on constate que les corrélations r_{12}, r_{34} et r_{54} diminuent tandis que les corrélations r_{14}, r_{23} et r_{25} augmentent. Pour la méthode 3, c'est le contraire. Les méthodes 6 et 7 produisent, la plupart du temps, une sous estimation du coéfficient de corrélation. Et enfin la méthode 4 semble la plus efficace pour maintenir les corrélations et la méthode 1, la pire.

5. CONCLUSIONS ET RECOMMANDATIONS

En résumé on peut conclure que lorsque la distribution des donneurs et des candidats est identique et que le taux de non-réponse est faible ou modéré les méthodes d'imputation utilisant les répondants sont recommandées pour estimer des moyennes. Lorsque le taux de non-réponse est élevé les méthodes d'imputation utilisant la technique du plus proche voisin s'améliorent de sorte que dans ce cas toutes les méthodes sont recommandées.

Considérons maintenant le cas où la distribution des donneurs et des candidats est différente. La méthode consistant à imputer la moyenne des répondants n'est pas recommandée quel que soit le taux de non-réponse. On remarque que toutes les autres méthodes d'imputation sont recommandées si le taux de non-réponse est faible. Lorsque le taux de non-réponse est modéré, seules les méthodes consistant à imputer la valeur d'un plus proche voisin ajustée par une variable auxiliaire et les méthodes consistant à imputer la moyenne des répondants ajustée par pondération sont recommandées. Seules les méthodes consistant à faire l'imputation par la moyenne des répondants ajustée par pondération sont recommandées lorsque le taux de non-réponse est élevé, et que des estimations de moyennes et de totaux sont désirées.

Pour ce qui est de la structure de corrélation entre les variables on doit conclure que les méthodes utilisant la technique du plus proche voisin sont particulièrement recommandées. Pour ces méthodes, la corrélation entre deux variables ayant été imputées est en général maintenue.

D'après les résultats obtenus dans cette étude il semble que le choix de la variable auxiliaire utilisée dans les méthodes d'ajustement par le ratio ne n'affecte pas les résultats obtenus. En effet les estimateurs utilisant une variable auxiliaire très corrélée avec la variable nécessitant imputation ne donnent pas des résultats différents de ceux qui utilisent une variable auxiliaire moins corrélée. Il semble que plus le taux de non-réponse augmente, moins les estimateurs sont efficaces. Par contre, d'autres tests pourraient être exécutés en utilisant différentes structures de corrélation et différents niveaux de non-réponse pour vérifier si ces conclusions sont toujours valables.

Le tableau I donne une idée rapide des différentes qualités de chacun des estimateurs, pour l'ensemble de données utilisé. Il est important de noter que les conclusions de cette étude ne s'appliquent que dans le contexte décrit ici. La robustesse des résultats, quant au choix des distributions des variables, à la structure de corrélation entre les variables, au taux de non réponse, à la taille de la population et à la taille de l'échantillon n'a pas été déterminée. Par conséquent, la généralisation des résultats obtenus dans cette étude ne peut se faire qu'avec une très grande prudence.