WEIGHTING VS IMPUTATION: A SIMULATION STUDY

Sylvie Michaud, Statistics Canada

0. INTRODUCTION

The Census of Construction (COC) is an annual survey conducted by Statistics Canada. Among other things, it estimates expenses in the construction industry in Canada. Even though the survey is called a census, only large firms are completely enumerated (they are mailed a long questionnaire to get both financial and non-financial information). For small firms, administrative records are used both as a frame and to get information. Two stratified samples of small firms are selected from overlapping frames, to get basic information on the firms. Then two stratified sub-samples are selected independently from one of the samples, to obtain additional financial and non-financial information. More details about the design can be found in [3].

Two strategies could be used to estimate totals for variables collected in the sub-samples. One approach would be to weight the records. This would necessitate the calculation of different weights (at least one associated with each sub-sample, one associated with the sample). Another strategy could be to impute the appropriate missing data segments for records selected into the initial samples but not into the sub-sample(s). This approach creates a "rectangular" sample file that can then be weighted up to the population level, using one weight only. This strategy of "mass imputation" is the one currently used by the survey.

The purpose of the study is to compare the estimates obtained using the imputation strategy to the ones that could be obtained if a weighting strategy were used.

1. SIMULATION

A simulation study was done to compare the weighting and the imputation strategies. The simulation reproduced the steps of the survey (the sample design, the imputation and the estimation), but in a simplified manner. Only one of the sub-samples i.e.,the financial sub-sample, was studied for this simulation.

The study was restricted to unincorporated businesses, (a subset of the real population of unincorporated and incorporated businesses) due to practical reasons (the sample design for the incorporated businesses will be changed next fiscal year and will become similar to the one used for unincorporated businesses). It is hoped that the findings would be similar for the population of incorporated businesses.

The simulation has been done on a reduced real population, namely the records selected in the financial sub-sample for the fiscal year 1983. For that population, a value is present for every variable that is collected in either the sample or the financial sub-sample. The simulation population size is approximately 5,000 records. A stratified sample of 1,300 records has been drawn from that population (using approximately the same sampling fractions within the strata as are used in the survey). For the non-sampled records the variables of the financial sub-sample were blanked out. The sampled records were kept intact. The entire file (sampled and non-sampled records) was run through the imputation system. For the non-sampled records, the variables of the financial sub-sample were imputed. Estimates using the imputed values were computed and compared to weighted estimates. (More details about the weighted estimates are presented in section 3).

The process of selecting the sample, imputing the missing data and calculating estimates was repeated 30 times.

2. IMPUTATION

A brief overview of the imputation procedure of the COC is given here. More details can be found in [2] or [6].

The COC imputation strategy uses a "nearest neighbour" approach, within a "deck" of potential donors. More precisely, the file is sorted by stratum (geographical and classification variables). Within each stratum, the file is also sorted by income. For each candidate record[1], a deck of ten "potential" donors is found (the five donors before the candidate on the file, and the five after). A pre-defined distance function then determines which of the ten potential donors is the nearest neighbour (that is, has the smallest distance to the candidate). The imputed values are the values of the nearest neighbour, adjusted by the ratio of an auxiliary variable (which is present for both the donor and the candidate). This is actually a simplification of the real procedure. More details can be found in [6]. The variables are imputed in a certain order in order to insure that the edits constraints will be satisfied.

Following the imputation, estimates of characteristics were generated by summing over both the imputed and non-imputed data.

3. WEIGHTING

Three weighting procedures were considered for this study: a simple weighting estimate, where the weight is the inverse of the probability of selection, a ratio estimator and a regression estimator. For the ratio and the regression estimators, the auxiliary variable used is the same as that used in the imputation procedure. Formulas for the different estimates are given in Appendix 1. The variances of the weighted estimates are straightforward. They are also presented in Appendix 1.

The variance of the estimate obtained after the mass imputation is not as easy to derive. However, if one makes the assumption that imputing the nearest neighbour and adjusting by the ratio of an auxiliary variable is approximately equivalent to imputing and adjusting by the mean of the stratum, then the variance of the estimate obtained after imputation is equal to the variance of ratio estimate. The simulation will test that hypothesis.

4. RESULTS

There are seven variables collected in the financial sub-sample. Four of the variables have been studied:

ADD: Additions to fixed assets,
BEN: Employee benefits,
DEB: Bad debts,
RM: Repairs and maintenance.

The distributions of these variables are presented in Table 1. The variables are all skewed with a peak at the zero value, with the exception of repairs and maintenance, for which a zero value was considered invalid.

The variables used as auxiliary variables are different types of expenses. These variables are obtained from the information collected for the entire sample. More information on these variables can be found in [6].

For the mass imputation approach as well as for each of the weighting techniques, estimates of totals have been calculated. Table 2 presents the true population value (Y), and the different estimates of totals obtained by the various estimators ($\hat{Y}$) (average over the 30 replicates). From the 30 replicates, an estimate of the standard deviation (s) and of the bias ($\hat{Y}$-Y) of the estimates were also calculated. Tests were performed to see if there were significant differences between the estimates and if any of the estimates of bias were significantly different than zero. Before comparing the estimates however, the variances were tested to determine if they were equal. The hypothesis of equal variances was rejected for BEN, because of the variances of the ratio and regression estimates. Their high variances are mainly due to the distribution of the auxiliary variable used with BEN. That auxiliary variable was often zero with a few very small positive values. In certain strata, a "bad" sample can give a very small non-zero denominator which can result in an excessive increase in the estimate of BEN. Because the equality of variances was rejected for BEN, Welch and Brown-Forsythe statistics were computed. These test the equality of the estimates when variances are not assumed to be equal. The estimates were not found to be significantly different. For the three other variables, (ADD, DEB, RM) an ANOVA resulted in no significant differences between the estimates. The one exception was the ratio estimate of ADD, which was significantly underestimating the true value, compared to the other techniques. In terms of bias, only the ratio estimate has a significant bias, and only for the variable ADD. So, for the variables studied, weighting by the inverse of the probability of selection and mass imputation seem to be equivalent strategies to compensate for non-sampled records. Regression estimates may be used for some of the variables, but with caution (as demonstrated by the variable BEN). As for the ratio estimate, it is biased for ADD and overestimates the variance for BEN. The approximation of the imputation variance by the estimate of the variance of the ratio estimate appears to be a good one, but not for all variables.

Because there did not seem to be significant differences between estimates obtained by weighting by the inverse of the probability of selection and imputation, it was decided to evaluate more closely the imputation itself and try to see how the imputation "affected" the data. Different coefficients of correlation were calculated, before and after the imputation, to see if the imputation changed the correlational structure.

Coefficients of correlations were calculated for each replicate. Fisher's transformation [5] was applied to the coefficients. Table 3 presents the results. Even though many of the coefficients of correlation are significantly different (usually higher) after the imputation, only one (between RM and its auxiliary variable) showed a substantial increase. This could be explained by the fact that the variables are imputed in a particular order. Under the ordering algorithm, RM is often the last variable imputed. Because of the edit constraints and of the imputation procedure , RM is more confined into a model, and the correlation is increased because of that.

5. CONCLUSIONS

Weighting by the inverse of the probability of selection and doing mass imputation of non-sampled records appear to be equivalent strategies for the variables in this study. The estimates and the estimates of variances were never significantly different. Also, neither technique showed a significant bias. There could be various reasons for choosing one strategy over another. For example, in the actual COC survey design, two sub-samples are selected independently. Under weighting adjustments, cross-tabulations of

TABLE 1

DISTRIBUTION OF THE STUDIED VARIABLES

ADD	0	1-500	501-1000	1001-1500	1501-2000	2001-2500	2501-3000	3001-3500	3501-4000	4001-4500	4501-5000	5001+
Freq.	3132	185	172	107	81	72	58	58	65	47	43	719

BEN	0	1-100	101-200	201-300	301-400	401-500	501-600	601-700	701-800	801-900	901-1000	1001+
Freq.	4009	51	47	42	51	37	41	37	39	31	25	329

DEB	0	1-100	101-200	201-300	301-400	401-500	501-600	601-700	701-800	801-900	901-1000	1001+
Freq.	4301	45	44	40	22	33	18	12	14	11	18	181

RM	1-500	501-1000	1001-1500	1501-2000	2001-2500	2501-3000	3001-3500	3501-4000	4001-4500	4501-5000	5001+
Freq.	2065	829	454	272	184	136	106	79	88	49	477

TABLE 2

ESTIMATES OF TOTALS, FOR THE STUDIED VARIABLES
(Average over 30 Replicates)

Variables	Population (true value)	Weighted	Ratio	Regression	Imputation
ADD (x 1000)					
$\hat{Y}$	14,127	14,343	12,397**	14,028	14,056
s	-	1,111	852	1,104	1,149
$(\hat{Y}-Y)$	-	216	-1,730**	-99	-71
BEN (x 1000)					
$\hat{\bar{Y}}$	1,210	1,183	1,572	1,232	1,213
s	-	118	1,427*	379	145
$(\hat{\bar{Y}}-Y)$	-	-27	362	22	3
DEBTS (x 1000)					
$\hat{\bar{Y}}$	833	844	858	824	829
s	-	139	147	141	101
$(\hat{\bar{Y}}-Y)$	-	11	25	-9	-4
RM (x 1000)					
$\hat{\bar{Y}}$	10,571	10,564	10,666	10,466	10,433
s	-	452	411	504	487
$(\hat{\bar{Y}}-Y)$	-	-7	95	-105	-138

* significance level = 0.05
** significance level = 0.01

TABLE 3

COEFFICIENTS OF CORRELATION
BEFORE AND AFTER IMPUTATION

Coefficients of Correlation	Variables			
	ADD with WADD	BEN with SAW	DEB with EXP	RM with EXP1
ρ (Population)	.288	.378	.178	.481
$\hat{\bar{\rho}}$ (30 REP.) (After Imputation Mean)	.324	.394	.176	.633
$\hat{\bar{Z}}$ (30 Rep.) (Transformed Mean)	.336*	.419*	.178	.744*
Z (Transformed Pop. Value)	.299	.398	.180	.523

Coefficients of Correlation	Variables			
	ADD with RM	BEN with RM	DEB with RM	BEN with DEB
ρ (Population)	.423	.080	.053	.068
$\hat{\bar{\rho}}$ (30 REP.) (After Imputation Mean)	.378	.080	.077	.065
$\hat{\bar{Z}}$ (30 Rep.) (Transformed Mean)	.398*	.080	.077*	.068*
Z (Transformed Pop. Value)	.451	.080	.053	.065

* significantly different $\alpha = .05$

variables from the two sub-samples, would generally give inconsistent marginal totals. This problem could be solved by raking on the variables. However, the process may become cumbersome if raking must be carried out for every cross-tabulation. Mass imputation remedies the problem by creating a complete file, from which all tabulations would necessarily be consistent. In the COC case, the hierarchical manner in which the imputation is done increases the correlation between certain variables. The imputation strategy can allow more flexibility in the model imposed on each variable (as opposed to weighting by the inverse of the probability of selection, where every variable gets the same weight). On the other hand, variance estimates are more easily calculated under a weighting strategy. In addition, developing a weighting system often requires fewer resources than developing an imputation system.

For this study, the ratio or the regression estimators did not seem to improve the estimates (over weighting by the inverse of the probability of selection). For some variables they yielded inferior results (biased estimates, increased variances). Their use for estimating totals is not always to be recommended. Since the mass imputation and the weighting lead to similar estimates, the choice between them will in practice be dictated by resource constraints, the number of records to be processed, and the type of information required.

NOTE

1. A candidate is defined as a record that requires imputation either due to missing values or to edit failures and a donor, as a record that does not need imputation.

REFERENCES

[1] Cochran, W.G., (1977). Sampling Techniques. New-York, John Wiley & Sons

[2] Colledge, M.L., Johnston, J.H., Paré, R. and Sande, I.G., (1978). Large Scale Imputation of Survey Data, Proceedings of the Section on Survey Research Methods, American Statistical Association, 721-726.

[3] Giles, P., (1983). Construction Division: Census of Construction, Technical paper, Business Survey Methods Division, Statistics Canada.

[4] Michaud, S., (1986). Comparison of Weighting and Imputation for Non-Sampled Records, Technical paper, Business Survey Methods Division, Statistics Canada.

[5] Neter, J. and Wasserman, W., (1974). Applied Linear Statistical Models, Illinois, Irwin, pp. 404-407.

[6] Philips, J.L. and Emery, D., (1976). FIBCOC Technical Documentation, System Development Division, Statistics Canada.

APPENDIX 1

Notation

Y : variable of study (imputed)

X : auxiliary variable (present for both the donors and the candidates)

h : stratum

N : population size

N_h : population size in stratum h

n_h : sample size in stratum h

$s^2_{y_h}$: estimated variance of Y in stratum h

$s^2_{x_h}$: estimated variance of X in stratum h

s_{yx_h} : estimated covariance between Y and X in stratum h

The three weighted estimates can be expressed as:

weighted:

$$\hat{Y}_w = \sum_h \frac{N_h}{n_h} \sum_{j=1}^{n_h} y_{jh}$$

ratio:

$$\hat{Y}_{RAT} = \sum_h \left(\frac{\sum_{j=1}^{n_h} y_{jh}}{\sum_{j=1}^{n_h} x_{jh}} \right) \times \left(\sum_{j=1}^{N_h} x_{jh} \right)$$

$$= \sum_h \hat{R}_h \times \left(\sum_{j=1}^{N_h} x_{jh} \right)$$

regression:

$$\hat{Y}_{REG} = \sum_h N_h \left(\sum_{j=1}^{n_h} \frac{y_{jh}}{n_h} + b_h \times \left(\sum_{j=1}^{N_h} \frac{x_{jh}}{N_h} - \sum_{j=1}^{n_h} \frac{x_{jh}}{n_h} \right) \right)$$

where:

$$b_h = \sum_{j=1}^{n_h} \frac{(y_{jh} - \bar{y}_h)(x_{jh} - \bar{x}_h)}{\sum_{j=1}^{n_h} (x_{jh} - \bar{x}_h)^2}$$

The imputed estimate is simply:

$$\hat{Y}_I = \sum_{j=1}^{N} y_j \ .$$

The variance of the estimates can be expressed as follows:

$$v(\hat{Y}_w) = \sum_h N_h \frac{(N_h - n_h)}{n_h} s_{y_h}^2$$

$$v(\hat{Y}_{RAT}) = \sum_h \frac{N_h^2}{n_h^2} (N_h - n_h) \times (s_{y_h}^2 + \hat{R}_h^2 s_{x_h}^2 - 2\hat{R}_h s_{yx_h})$$

$$V(\hat{Y}_{REG}) = \sum_h \frac{N_h^2}{n_h^2} \frac{(N_h - n_h)}{(n_h - 2)} \times ((n_h - 1) s_{y_h}^2 - \frac{s_{yx_h}^2}{s_{x_h}^2}) \ .$$

For the imputation technique, if it is assumed that

$$y_{c_{jh}} = \frac{y_d}{x_d} * x_{c_{jh}} \simeq \frac{\bar{y}_h}{\bar{x}_h} * x_{c_{jh}}$$

where c: subscript to represent a candidate record
d: subscript to represent a donor record,

then,

$$\hat{Y}_I = \sum_{j=1}^{N} y_j = \sum_h \left(\sum_{j=1}^{n_h} y_{jh} + \sum_{j=n_h+1}^{N_h} y_{cjh} \right)$$

$$\simeq \sum_h N_h \left(\frac{\sum_{j=1}^{n_h} y_{jh}}{\sum_{j=1}^{n_h} x_{jh}} \right) \times \left(\sum_{j=1}^{N_h} \frac{x_{jh}}{N_h} \right) ,$$

and so,

$$v(\hat{Y}_I) \simeq v(\hat{Y}_{RAT}) \ .$$