Survey Methodology, December 1986
Vol. 12, No. 2, pp. 197–206
Statistics Canada

Comparison of Weighting and Imputation Methods for Estimating Unsampled Data

SYLVIE MICHAUD[1]

ABSTRACT

The Canadian Census of Construction (COC) uses a complex plan for sampling small businesses (those having a gross income of less than $750,000). Stratified samples are drawn from overlapping frames. Two subsamples are selected independently from one of the samples, and more detailed information is collected on the businesses in the subsamples. There are two possible methods of estimating totals for the variables collected in the subsamples. The first approach is to determine weights based on sampling rates. A number of different weights must be used. The second approach is to impute values to the businesses included in the sample but not in the subsamples. This approach creates a complete "rectangular" sample file, and a single weight may then be used to produce estimates for the population. This "large-scale imputation" technique is presently applied for the Census of Construction. The purpose of the study is to compare the figures obtained using various estimation techniques with the estimates produced by means of large-scale imputation.

KEY WORDS: Weighting; Large-scale imputation; Unsampled.

1. INTRODUCTION

The Census of Construction (COC) is an annual survey which attempts to estimate expenses in the construction field. Although it is called a "census", in fact only businesses having a gross income exceeding $750,000 are surveyed. Various financial and non-financial data are collected by means of a long questionnaire mailed to these firms. For businesses with a gross income between $10,000 and $750,000, expenses are estimated from a sample of administrative data. First, two samples are selected independently from overlapping sample frames. Two subsamples are then drawn from one of the samples in order to obtain additional information.

Variables collected in the subsamples may be estimated in two different ways. The method currently used for the Census of Construction is to impute values for the businesses included in a sample, but not in a subsample. This creates a complete "rectangular" file, from which estimates for the overall population may be produced using only one weight. An alternative would be to calculate weights based on the probabilities of selection; these would have to be calculated separately for different subsets of data. The purpose of this study is to compare the estimates obtained by weighting with the estimates obtained by imputation.

The study was carried out on a population of unincorporated businesses only because, for fiscal year 1983, the sample selection strategies for unincorporated and incorporated businesses were different. The strategy used for corporations will be modified for fiscal 1984 to be equivalent to the strategy for unincorporated businesses. The strategy for unincorporated businesses was therefore examined. One hopes that the conclusions of this study will remain the same for incorporated businesses.

[1] S. Michaud, Business Survey Methods Division, Statistics Canada, 11th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

2. DESCRIPTION OF THE SAMPLING PLAN

As mentioned above, two independent samples are drawn from overlapping sample frames. The first is the prespecified sample selected for the Census of Construction; it is stratified by gross business income (GBI), province and 3-digit 1970 Standard Industrial Classification (SIC) code. The sample frame used is not completely up-to-date. It contains some "deaths", i.e. businesses which are no longer within the scope of the COC for various reasons (a firm which no longer exists, is no longer engaged in a construction activity, or whose gross income is below $10,000). Furthermore, the sample frame does not contain "births" or businesses which have changed activities and are now part of the construction industry. The second sample is a "cross-sectional" sample, selected independently by Revenue Canada from a complete database containing businesses in all SIC groups (not only construction). It is used to estimate "births". This sample is stratified by Gross Business Income ranges. Figure 1 below illustrates the situation.

Two independent subsamples are selected from the units of the prespecified sample: a financial subsample and a subsample of "other characteristics" (OC). The OC subsample is drawn directly from the prespecified sample, while the financial subsample is selected using data transcribed from the sample (and so "deaths" are not subsampled). Further details concerning the sampling plan may be found in Giles (1983).

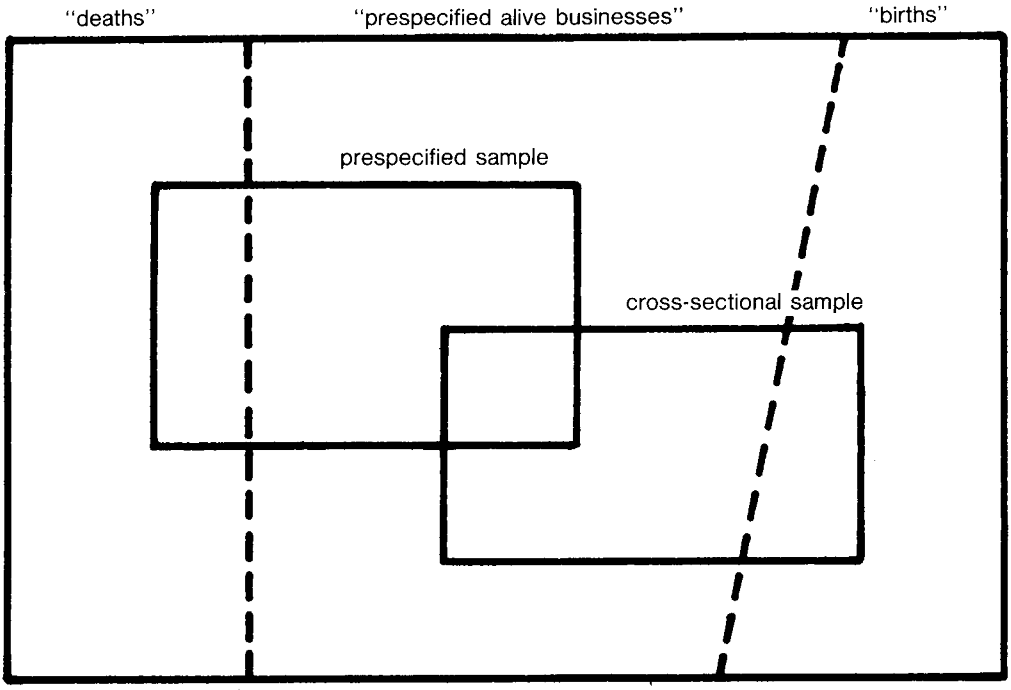

Figure 1. Representation of RC Sampling Plan

3. IMPUTATION TECHNIQUE

The COC uses a large-scale imputation technique to estimate the variables selected in a given subsample (i.e. values are imputed for each variable, for all records not selected in the subsamples). The imputation is carried out independently for each subsample. (The imputation is done in phases, and the imputation phases of the various subsamples are mutually independent and apply different techniques.) In each phase, the nearest neighbour is chosen from a subset of potential donor records, and is used to impute the variables which were not sampled.

The imputation is carried out differently for each subsample.

In the case of the financial subsample, the imputed value is the donor's value, adjusted by the ratio of an auxiliary variable which is available for both the donor and the candidate (the candidate being the record which is missing data to be imputed). (Note: The actual procedure is more complicated: the variables are imputed hierarchically and linear constraints are placed on the imputed values (the second variable is dependent on the value imputed to the first variable, etc.). Additional information on this procedure may be found in Philips and Emery (1976). A more detailed overview is also provided in Colledge *et al.* (1978)).

Suppose we use the following notation:

Y: the variable of interest (known for the donors, to be imputed for the candidate)
X: an auxiliary variable available for both the donor and the candidate
c: denotes the candidate
d: denotes the donor
I: denotes an imputed value.

For the financial subsample variables, the imputed value Y_c^I is defined to be:

$$Y_c^I = Y_d \frac{X_c}{X_d}$$

For the OC subsample variables, the imputed value is simply the value on the donor record:

$$Y_c^I = Y_d$$

The imputation procedure produces a complete rectangular file (the records of all the businesses that were selected in one of the samples contain values for all the variables of the samples/subsamples). Sampling weights may then be used to generate estimates for the overall population.

The weight assigned to a given record is the inverse of the probability of it being selected into at least one of the samples. If we use the following notation:

$P(\text{presp}_h)$: the probability of a record being selected in stratum h of the prespecified sample
$P(\text{cross}_k)$: the probability of a record being selected in stratum k of the cross-sectional sample
hk : cross-classification of records
h : denotes the stratum of the prespecified sample
k : denotes the stratum of the cross-sectional sample,

then the weight associated with each unit may be expressed as:

$$W_{hk}^{-1} = 1 - [1 - P(\text{presp}_h)]\,[1 - P(\text{cross}_k)]$$

Births and deaths cannot be cross-classified. Deaths have a zero weight $W_h = 0$ and the weight of a birth, W_k, is the inverse of the probability of being selected in stratum k of the cross-sectional sample. More details may be found in Bankier (1982).

Therefore, when the imputation technique is used, the estimator of the total is

$$\hat{Y} = \sum_{h,k} W_{hk} \sum_{j=1}^{n_{hk}} y^{*}_{jhk}$$

where $y^{*}_{jhk} = y_{jhk}$ if $j \in$ subsample

$= y^{I}_{jhk}$ if $j \notin$ subsample.

4. WEIGHTING TECHNIQUE

If a weighting technique were used to estimate subsample variables, there would be a number of possible estimators. The estimators are in the same form for both subsamples, but different weights are used.

The first estimator ($\hat{Y}_1$) would be based on the sampling plan used, adjusted for undercoverage of the population. In each of the SIC, PROV and GBI strata (Standard Industrial Classification, province, gross business income), a prespecified sample is selected. Once they have been transcribed (units sampled and still alive), the units are classified to two strata: "outside survey field" and "within survey field".The subsamples are chosen from the "within survey field" stratum. (We may assume that all the units in the "outside survey field" stratum have been subsampled and have a mean equal to zero.)The estimator contains a correction factor that compensates for undercoverage of the sample frame (calculated using information from the cross-sectional sample).

The second possible estimator ($\hat{Y}_2$) is a simplified version of the first estimator, $\hat{Y}_1$. Instead of assuming a double sampling to determine "within survey field" and "outside survey field" units, we could assume that a prespecified stratified sample is selected from "within survey field" units. A subsample is selected from the prespecified sample. The estimator must once again be adjusted to take undercoverage into account. If the differences between the first and second estimator turn out to be insignificant, the second would be a better choice because it is simpler.

The third possible estimator ($\hat{Y}_3$) is an estimator based on data from the cross-sectional sample only. We could assume that the units selected in both the subsample and the cross-sectional sample are selected from the cross-sectional sample. The reasoning behind such an estimator is that the cross-sectional sample is drawn from a complete sample frame. However, since the subsamples are selected from the prespecified sample, and not from the cross-sectional sample, the size of the subsamples in the cross-sectional sample will be small.

Finally, a fourth estimator ($\hat{Y}_4$) could be obtained by supposing that the subsample is selected from the complete sample (prespecified sample + cross-sectional sample), and that the complete sample comes from multiple frames. This fourth estimator is the one that most closely resembles the estimator obtained after large-scale imputation. Indeed, both of these estimators assume that births and new businesses "react" like the rest of the population. The imputation procedure does not make any special adjustment for such businesses, and the weighted estimator is not stratified in such a way as to distinguish these units. In addition, both estimators take into account the fact that the sample comes from a number of frames. The same sampling weight is therefore used in both cases to produce data up to the population level.

As mentioned above, the variables collected in the financial subsample are adjusted by the ratio of an auxiliary variable during the imputation.

We could therefore propose another type of estimator for the variables collected in the financial subsample: a ratio estimator. The auxiliary variable used would be the same one used for the imputation. As is the case for the simple weighting, different estimators could be calculated.

The various estimators and their variances are described in mathematical terms in the Appendix.

5. RESULTS

In the study, four of the seven variables in the financial subsample were considered.

As for the subsample of other characteristics, eight variables are collected for all businesses, while other variables are available for certain SIC groups only. The study was therefore limited to these eight variables.

The variables in the financial subsample presented in this report are ''ADD'' (additions to fixed assets) and ''RM'' (repair and maintenance). For the OC subsample, results are given for the variable ''PCON'' (percentage of construction in a specific field). However, the PCON variable is not published directly, but is multiplied by total expenses to obtain expenses in a specific field: PEXP. This second variable was the one studied.

As mentioned earlier, the variables in the OC subsample are not adjusted by a ratio during the imputation procedure. The ratio estimators will therefore not apply to these variables.

Tables 1, 2 and 3 provide values for the different estimators and estimates of their respective variances, based on 1983 tax data for unincorporated businesses.

In the first place, we see that there are no significant differences between the first two estimators. (According to the predetermined definitions, the second estimator is a simplified version of the first one.)The simplified version will therefore be retained.

Table 1

Estimated Values of PEXP (%EXP*EXPCONS) and Standard Deviation of PEXP

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_1$
Estimate (× 10^{11})	3.44	3.43	3.96	3.66	3.70
Standard deviation (× 10^9)	3.5	3.5	8.4	3.2	

Table 2

Estimated Values of ADD and Standard Deviation of ADD

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_{Q2}$	$\hat{Y}_{Q3}$	$\hat{Y}_{Q4}$	$\hat{Y}_1$
Estimate (× 10^8)	2.08	2.10	2.14	1.84	7.82	5.06	5.2	1.4
Standard deviation (× 10^7)	1.9	1.9	2.0	1.0	0.8	2.2	0.8	

Table 3

Estimated Values of RM and Standard Deviation of RM

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_{Q2}$	$\hat{Y}_{Q3}$	$\hat{Y}_{Q4}$	$\hat{Y}_1$
Estimate (× 10 ^{8})	1.5	1.5	1.43	1.55	0.9	1.63	1.67	1.75
Standard deviation (× 10^6)	6.9	6.9	8.9	5.3	3.1	11.0	4.3	

In general, for the variables in the financial subsample, the imputation technique appears to yield results similar to those produced by the weighting method ($\hat{Y}_4$). The estimator obtained by considering only units drawn from the cross-sectional sample ($\hat{Y}_3$) seems more variable than the other estimators. This variability could be explained by the smaller number of units used to calculate this estimator. It should be pointed out that these comparisons are based only on an observed sample, and so the conclusions are somewhat limited. However, owing to the nature of the data (often percentages and subdivisions of activity in the construction field), which is relatively stable in the strata (3-digit 1970 SIC, province and GBI), it was considered unnecessary to analyse these variables in greater depth.

For the variables in the financial subsample, it was found that the estimators adjusted by the ratio do not always seem applicable (for example, the ADD variable). The estimates which they produce are extremely biased. One possible explanation is that the ADD variable and the auxiliary variable used have a high frequency of zero values. A "bad" sample in certain strata can thus inflate the estimates inordinately.

Some problems were also encountered with the imputation system (data imputed when they should not have been, data not imputed), which in certain instances may have affected the estimates obtained by the imputation method. Since the results were based on an observed sample only, and because it was difficult to estimate the impact of the system-related problems, it was decided that a simulation would be done.

6. SIMULATION

The simulation was carried out using a data subset, namely those businesses that had been selected in the financial subsample (all of the variables studied are present for this data subset). Then an attempt was made to apply a simplified version of the technique used by the Census of Construction. A stratified sample was selected, using sampling rates similar to those of the survey. The variables of the financial subsample, for the data not selected in the sample, were considered as missing, and then imputed by the system. The sample selection process and the imputation were repeated thirty times.

Estimates were produced, allowing us to compare the results obtained by summing the non-imputed and imputed data with the estimates produced using sampling weights equal to the inverse of the sampling rate. Since the value for the population is known, the bias and the variance of the estimates were calculated. The results for the ADD and RM variables are shown in Tables 4 and 5.

For the ADD variable, the value produced by ratio estimation differs significantly from the estimates obtained by imputation or by weighting. The bias of the estimate is also significantly not null. For the RM variable, all the estimators are equivalent (equal variances, bias not significant at a 5% level, estimates not significantly different).

Table 4

ADD Estimates Obtained by Simulation

	Population	Weighting	Ratio	Imputation
Estimate ($\times$ 10^7)	1.41	1.43	1.24	1.41
Standard deviation ($\times$ 10^5)		1.11	.85	1.15
Bias ($\times$ 10^5)		.22	−1.73	−0.07

Table 5

RM Estimates Obtained by Simulation

	Population	Weighting	Ratio	Imputation
Estimate ($\times$ 10^7)	1.06	1.06	1.07	1.04
Standard deviation ($\times$ 10^5)		4.52	4.11	4.87
Bias ($\times$ 10^5)		−0.07	−0.95	−1.38

7. CONCLUSIONS

According to the study results, there do not appear to be significant differences between the large-scale imputation technique and the weighting technique, for the variables in the other characteristics subsample. This was foreseeable, inasmuch as the variables studied seem to be relatively stable within each stratum.

The conclusions for the variables in the financial subsample are based on the results of the simulation. These seem to indicate that the estimates obtained by weighting by the inverse of the probability of selection are comparable to the estimates obtained from large-scale imputation.

The ratio estimator does not appear appropriate for the ADD variable (or for the other variables analysed, but not discussed in this report). Continuation of the study will try to determine whether a regression estimator would be more appropriate, and to evaluate the impact of the imputation on the variable correlation structure.

ACKNOWLEDGEMENT

The author would like to thank P. Giles for his comments and suggestions given during this study, and also the referees for their helpful comments.

APPENDIX

The following notation may be used for the proposed estimators:

h : stratum of the prespecified sample

k : stratum of the cross-sectional sample

N_h : size of the "prespecified" population in stratum h

$\hat{N}_{1h}$: size of the "prespecified" population with "alive businesses (within the scope of the survey) in stratum h (estimated)

$\hat{N}_{2h}$: size of the "prespecified" population with businesses "outside the scope of the survey" in stratum h (estimated)

$\hat{N}_k$: size of the population in stratum k, estimated using information from the cross-sectional sample

$\hat{N}'_k$: size of the population in stratum k, estimated using information from foth samples (multiple frames)

n_h : number of units sampled in stratum h of the prespecified sample

$\hat{n}_{1h}$: number of units sampled and transcribed in stratum h of the prespecified sample

$\hat{n}'_k$: number of units sampled and transcribed in stratum k

$\hat{m}_{1h}$: number of units subsampled from among "alive" businesses in stratum h

y : variable of one of the subsamples

x : auxiliary variable available for all units of the samples

s^2_{yh} : estimate of the variance of y for the units of the subsample in stratum h

s^2_{xh} : estimatee of the variance of x for the units of the subsample in stratum h

s_{yxh} : estimate of the covariance of x and y in stratum h.

i)

$$\hat{Y}_1 = \left(\frac{\hat{N}_{1\ \text{pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\ \text{pre-spec}}}\right) \sum_h \frac{N_h}{n_h} \frac{\hat{n}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} Y_{hj}$$

$$V(\hat{Y}_1) \simeq \left(\frac{\hat{N}_{1\ \text{pre-spec}} + \hat{N}_{\text{births}}}{\hat{N}_{1\ \text{pre-spec}}}\right)^2 \sum_h N_h n_h \left(\frac{N_h - 1}{n_h - 1}\right)$$

$$\times \left[W_{1h} S^2_h \left(\frac{1}{\gamma_h} - \frac{1}{N_h}\right) + \frac{G_h}{n_h} S^2_h \left(\frac{W_{1h}}{N_h} - \frac{1}{\gamma_h}\right) + \frac{G_h}{n_h} W_{1h}(1 - W_{1h})^2 \bar{y}^2_h\right]$$

where

$$G_h = \left(\frac{N_h - n_h}{N_h - 1}\right),\ \gamma_h = n_h \frac{\hat{m}_{1h}}{\hat{n}_{1h}},\ \text{and}\ W_{1h} = \frac{\hat{n}_{1h}}{n_h}.$$

ii)

$$\hat{Y}_2 = \left(\frac{\hat{N}_{1\ \text{pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\ \text{pre-spec.}}}\right) \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} y_{hj}$$

$$V(\hat{Y}_2) \simeq \left(\frac{\hat{N}_{1\ \text{pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\ \text{pre-spec.}}}\right)^2 \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} \left(\hat{N}_{1h} - \hat{m}_{1h}\right) s^2_{yh}$$

iii)

$$\hat{Y}_3 = \sum_k \frac{\hat{N}_k}{\hat{m}_{1k}} \sum_j Y_{kj}$$

$$V(\hat{Y}_3) = \sum_k \hat{N}_k \left(\frac{\hat{N}_k - \hat{m}_{1k}}{\hat{m}_{1k}}\right) S^2_{yk}$$

iv) $$\hat{Y}_4 = \sum_k \frac{\hat{N}_k'}{\hat{m}_{1k}} \sum_{j=1}^{\hat{m}_{1k}} y_{kj}$$

$$V(\hat{Y}_4) = \sum_k \hat{N}_\kappa' \left(\frac{\hat{N}_k' - \hat{m}_{1k}}{\hat{m}_{1k}}\right) s_{yk}^2.$$

Ratio estimators may be calculated and, like simple estimators, they may take on different forms, depending on the hypotheses postulated. For example, the ratio estimator corresponding to estimator 4 would be:

$$\hat{Y}_{Q4} = \sum_k \hat{N}_k' \, \bar{Y}_{\text{sub}_k} \frac{\bar{X}_{\text{samp}_k}}{\bar{X}_{\text{sub}_k}}$$

where $\bar{X}_{\text{samp}_k}$ is the mean of variable X for the units selected in the complete sample, which are in stratum k

$\bar{X}_{\text{sub}_k}$ is the mean of X for the units selected in the subsample, which are in stratum k

$\bar{Y}_{\text{sub}_k}$ is the mean of variable Y in stratum k of the subsample.

$$V(\hat{Y}_{Q4}) = \sum_k (\hat{N}_k')^2 \left(\frac{1}{\hat{m}_{1k}} - \frac{1}{\hat{n}_{1k}'}\right) \left[s_{y_k}^2 + \hat{R}_k^2 s^2{}_{x_k} - 2\hat{R}_k s_{yx_k} + \left(\frac{1}{\hat{n}_{1k}'} - \frac{1}{\hat{N}_k'}\right) s_{y_k}^2\right]$$

where $\hat{R}_k = \dfrac{\bar{Y}_{\text{sub}_k}}{\bar{X}_{\text{sub}_k}}$.

REFERENCES

BANKIER, M., (1982). Variance formula for an estimator based on any number of independant stratified samples of which some are Poisson samples. Technical document, Business Survey Methods Division, Statistics Canada.

COCHRAN, W.G. (1977). *Sampling Techniques.* New York: John Wiley & Sons.

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.

GILES, P. (1983). Construction division: Census of Construction. Technical document, Business Survey Methods Division, Statistics Canada.

PHILIPS, J.L., and EMERY, D. (1976), FIBCOC documentation. Technical document, Systems Development Division, Statistics Canada.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika,* 60, 125-133.