# LARGE SCALE IMPUTATION OF SURVEY DATA

M.J. Colledge, J.H. Johnson, R. Paré and I.G. Sande, Statistics Canada

## 1. Introduction

The demand for statistical information to aid government and management decision-making has been increasing for many years. In the past, Statistics Canada was able to cope with this situation by expanding the scope and number of their surveys. Recently, such expansion has become inhibited as a result of two factors. Firstly, there is an increasing sensitivity to complaints from respondents about the burden of completing questionnaires. Secondly, current fiscal policies prevent growth in manpower. There is no indication that either of these factors is likely to be short-lived. Thus, in order to cater to an increased demand for information without raising costs or response burden, Statistics Canada is committed to making the best possible use of existing data, including data collected by other agencies for administrative purposes. One particular manifestation of this policy was the decision to use financial data from Revenue Canada to supplement two annual surveys of businesses for the 1975 reference year. This paper deals with the systems which evolved as a result.

The Census of Construction (COC) is concerned with about 80,000 businesses in Canada whose primary activity is construction. Despite its name, the COC is, in fact, a survey. For the 1975 reference year, only businesses with gross business income (GBI) of at least $5,000 were considered in scope. These were divided into two groups: "small" businesses having a GBI of less than $500,000 and "large" businesses. The latter group were the subject of a census operation; all large businesses were mailed a questionnaire asking for a comprehensive set of data. Small business information was derived from two sources: from Revenue Canada and from a mailout as follows. A sample was selected and basic financial data were obtained from Revenue Canada tax files; for a subsample of these, secondary (more detailed) financial data were also obtained. A second subsample was selected and mailed a survey questionnaire requesting only non-financial data. Thus, in comparison with a full census, the COC response burden was reduced by sampling and by reducing the number and type of questions asked. This was made possible by the availability of relevant administrative data, namely tax files.

Arrangements for the Motor Carrier Freight Survey (MCF) were along the same general lines. The significant differences were that the universe of about 25,000 was divided into "small" and "large" by a GBI threshold of $100,000, no subsample of secondary financial data was obtained and the survey questionnaire requested a full range of information (not just non-financial).

The decision to utilize administrative tax data for the COC and MCF came quite abruptly and in advance of experience, existing software, data or a feasibility study. The short time scale combined with a restricted budget dictated certain constraints on the design. Firstly, program development and testing had to be substantially achievable before any real data were available. Secondly, the programs had to be robust and easily modifiable in order to allow adjustment for unexpected characteristics of the data. Thirdly, the programs had to interface with existing systems associated with the surveys, in particular, the tabulation systems which had been developed for census operations in previous years. Thus, the following design decisions were made:

(i) data from tax and survey sources would be combined at the micro level, i.e., level of individual businesses;

(ii) a complete set of data (all financial and non-financial items) would be imputed at micro level for all businesses using a "hot deck" technique with constraints to ensure that imputation was consistent with prescribed edit rules;

(iii) the data would be inflated to universe level by replication to allow tabulation by existing systems which had not been developed to handle weights; and

(iv) programs would be modular and readily adaptable to new or modified imputation and edit rules.

The following sections of this paper elaborate upon the design features and describe the systems implementation which processed 1975 data for the COC and MCF. An evaluation of the procedures is given in section 5.

## 2. Overview

The central feature of the system is the imputation procedure, discussed in detail in sections 3 and 4. The purpose of this section is to outline the environment within which the procedure operates by describing the complete system. The scale of processing is illustrated by reference to figures for the small business portion of the COC universe.

MERGE brings together data records from tax and survey sources. The input data files have been individually cleaned and edited. The output is a set of records, one per business, each of which contains a basic tax data segment and may (or may not) contain secondary tax data or survey data segments. The existing segments may have sporadic missing entries in various fields; also, some entries may be inconsistent with one another.

CHECKIN prepares data for imputation by screening out unusable or unwanted data. The module reformats records, strips off irrelevant fields, identifies out-of-scope or duplicate records, checks entries against a set of prescribed edit rules, blanks out inconsistent entries and identifies all missing fields. Any record which is out of scope or a duplicate or contains insufficient useful data is flagged ("dropped"); the remainder are subject to processing by the next module, IMPUTE.

Columns 1 and 2 of figure 1 illustrate the results of processing COC data. Some 9106 of the 50,538

Figure 1: Summary of Results of Processing Census of Construction, 1975 Reference Year

| | | | | 1 At Input to System | 2 After Checkin | 3 At Output from System | 4 Blown up to Universe |
|---|---|---|---|---|---|---|---|
| Out of Scope | | | | 9,106 | 9,106 | 9,106 | - |
| In Scope | Data not Good | | | 0 | 462 | 656 | - |
| | Good | | | | | | |
| Segments Present | Tax | | Survey | | | | |
| | Basic | Secondary | | | | | |
| | XXX | | | 34,181 | 33,937 | 0 | - |
| | XXX | XXX | | 2,316 | 2,186 | 0 | - |
| | XXX | | XXX | 4,027 | 3,963 | 0 | - |
| | XXX | XXX | XXX | 908 | 884 | 40,776 | - |
| Total Good | | | | 41,432 | 40,970 | 40,776 | 78,563 |
| Total | | | | 50,538 | | | |

merged records were declared out of scope. Of the remainder, 462 were dropped leaving 40,970 "good" records.

IMPUTE imputes all missing fields on every record. For the COC data, 884 records contained all segments, 3,963 records required imputation of just the secondary financial segment, 2,186 records required imputation of just the survey segment and 33,937 records required both (see figure 1, column 3). In addition, some entries in existing segments were missing.

CHECKOUT checks records against the same prescribed set of edit rules as were applied to the data at input, and identifies and "drops" records containing inconsistent or missing entries. This is necessary because inconsistent values may be imputed due to shortcomings in specification or programming, or the imputation may fail to define a consistent value for a field in some circumstances.

From columns 2 and 3 of figure 1 it can be deduced that 194 COC records were inconsistent or incomplete and had to be dropped.

INFLATE raises the sample of good records to the population level and thereby generates an output file which can be tabulated by the census tabulation system. Inflation is achieved by replicating each record according to its weight after "correction". All records entering the system carry a weight which is the inverse of the probability with which the record entered the basic tax sample. Three types of corrections are applied prior to replication:

(i) Duplication correction. Some businesses are represented by more than one record, as in the case of partnerships.

(ii) Out-of-scope correction. There are instances where the tax data information suggests the business is in scope, whereas the survey data indicates it is not. The survey data is assumed to be more reliable. In order to allow for possible inclusion of out of scope records containing tax data only, a correction factor is applied based on data from businesses for which tax and survey information is obtained.

(iii) Dropped record correction. Records for some in-scope businesses are dropped because of inadequate or inconsistent data. Thus the imputation procedure need not be successful in all cases as a correction can be made.

Figure 1 indicates that after weight correction and inflation a file of 78,563 small construction businesses was obtained.

3. Imputation Methodology

For purposes of imputation, the record for each business can be considered as consisting of four types of segments:

(i) Key fields. These consist of fields used for classification or matching and are collected or derived from the tax return. The actual fields used were the standard industrial classification (SIC), province, salaries and wages indicator (SWI, set to 1 or 0 according as there is any indication that salaries or wages were paid or not), gross business income (GBI), net business income (NBI). If any of these fields were missing, the record was not used in the imputation.

(ii) Basic financial data collected from the tax return, e.g., depreciation, purchases, closing inventory. An attempt is made to collect this data for all businesses sampled, but information available with the return may be insufficient or unclear. Thus the segment may be complete (all fields present) or incomplete (one or more fields missing).

(iii) Secondary financial data, collected from tax returns for a subsample of records. These detailed financial data, e.g., balance sheet, detailed expense breakdowns, were collected only for the Census of Construction; but, potentially, one or more such subsamples might exist. This segment may be either complete, incomplete (some fields present) or missing (no fields present, as in the case of records not in the subsample).

(iv) Survey data, collected for a subsample of records. This segment may be complete, incomplete or missing.

The imputation problem is to complete the incomplete segments and to supply the missing segments.

A possible imputation procedure would be to model the missing fields in terms of those that are present. If the number of fields were very large (as it is here) and the constraints (or edit rules) on the fields were at all complex, structuring and fitting the model would be very difficult and would have to be done after the data had been collected and edited. As a result, a great deal of time would be spent experimenting with the data just when one could least afford it - when the publication deadlines were approaching and a great deal of processing had yet to be done.

Thus, modelling the data did not seem a very attractive option and a type of hot-deck technique was devised. In this procedure, a record requiring imputation (candidate) is matched with a complete record (donor). The donor supplies the missing fields, possibly with some adjustment so that the edit rules are satisfied. This procedure produces realistic looking data and can be expected to preserve the underlying distributions, whereas modelling tends to produce smoothed data and distorts distributions. Another advantage is that the imputation can be set up and ready to run before the data collection is finished.

The hot-deck requires a reasonable supply of complete records, but in fact there are few records with all segments complete. If one attempted to impute for all missing fields in a single pass, the same donors would be used excessively, no use would be made of records with partial information, and the matches would be poor. In addition, a matching procedure appropriate for one segment may not be appropriate for another. Therefore, the imputation is broken up into several phases, each corresponding to a segment or sub-segment:

Phase 1. Candidates are records with Segment A incomplete (but not missing). Donors are records with Segment A complete. At the end of Phase 1, all records have Segment A complete or missing.

Phase 2. Candidates are records with Segment A missing. Donors are records with Segment A complete (including records which were Phase 1 candidates). At the end of Phase 2, all records have Segment A complete.

Phase 3. Candidates are records with Segment B incomplete (but not missing). Donors are records with Segment B complete. At the end of Phase 3, all records have Segment B complete or missing. Those with Segment B complete are eligible as donors in Phase 4.

In order to match candidates with donors, the file of all records is stratified by Province (or Region), SIC and SWI. The collection of potential donors (i.e., the hot-deck) as well as the collection of candidates are identified for the particular phase. Within the stratum, the records are ordered by GBI. A sequence of records from a stratum might be represented like this:

GBI: $25K $26K $27K $28K $29K
...CCDCCDCCDCCDCDCCDCCDCC$_o$DCCDCCDCCDCCDCCDC...

The C's are candidates and the D's are donors (other records not involved in this phase are not represented). In order to impute $C_o$, only the nearest 5 potential donors on "either side" of $C_o$ are considered, a total of 10 possible donors which are all about the same size (in terms of GBI) as the candidate. The number 5 is quite arbitrary. From the "nearest" 10 donors, that one is chosen which minimizes a distance function DIST ($C_o$,D). DIST can be quite a complex function, but the basic structure used was

$$DIST\ (C,D) = |\ \log EXP_C - \log EXP_D\ |$$

where EXP = GBI - NBI = total expenses, and the subscripts C and D denote values from the candidate and donor records, respectively. EXP was used because many of the fields to be imputed are detailed expense breakdowns or correlated with expenses.

Note that GBI and NBI are key fields, so that DIST is always determined. DIST may also depend on other key fields, or fields which have already been imputed in an earlier phase, or even metadata. In particular, the distance function may be modified to spread donor usage by making it also an increasing function of the number of times the potential donor D has already been used as an actual donor in the phase.

After a suitable donor has been identified, the candidate's missing fields are supplied from the corresponding fields in the donor record. Some adjustment or transformation may be necessary to ensure that the constraints (edits) are satisfied. For example, three fields, X, Y and Z, may have to satisfy $X + Y \leq Z$ with X, Y and Z all non-negative. The donor's values for these fields are $X_D$, $Y_D$ and $Z_D$ while the candidate has X and Y missing and the value $Z_C$ in the Z field. If the values $X_D$ and $Y_D$ are simply written into the corresponding candidate fields, we may find that $X_D + Y_D > Z_C$, which violates the edit. Therefore it is better to prorate $X_D$ and $Y_D$ to ensure that the edit holds:

$$X_C = (X_D / Z_D)\ Z_C;\ Y_C = (Y_D / Z_D)\ Z_C.$$

In other words, the proportions $X_D / Z_D$ and $Y_D/Z_D$ are transferred to the candidate. A common example is

$$FUEL_C = (FUEL_D / EXP_D)\ EXP_C$$

where FUEL is the amount spent on fuel and EXP is the total expenses. This imputation estimates that the candidate spent the same proportion of his total expenses on fuel as did the donor.

The transformation needed to impute a field may be more complex if the field is involved in several edits. In some cases a decision table may be required, where the form of imputation depends on which set of conditions holds. In desperate situations, a table of default values may be used.

If a field is not involved in any edits, it may be prorated using a correlated variable in the case of a numeric field. Categorical data, averages or proportions are usually copied from donor to candidate.

The imputation specifications are written separately for each field - no generalized transformation is used. They are written in such a way as to produce consistent data and this involves not only accommodating constraints, but also ensuring that constraints are not violated due to roundoff error.

## 4. Implementation

The systems design was based on the following principles:

(a) The breakdown into phases each of which is functionally the same, except in detail, suggested a general system which would be tailored separately for each phase.

(b) To simplify data-set control, the output produced from a phase would have the same record description as the input and all records would be carried forward. Each phase would identify its donors and candidates, perform imputation, and copy all other data as is.

(c) Instrumentation of the system would mostly be done offline by analysis of a log file describing imputation "events", and by investigation of the output of each phase.

(d) Fields would either have a value or be missing. If missing, any value which it might have had would be ignored for imputation purposes.

(e) Fields would be identified as missing only at beginning of processing. Once imputed to a value, the field stays imputed. Thus, inconsistencies must be removed at the beginning and never introduced by imputation.

(f) The control language should be quite flexible to allow unusual imputation rules, but should still be quite readable since it would be the final specification of side effects in unusual situations.

(g) One donor only would be used for each candidate in each phase.

The effect of these considerations on the design was to simplify the systems development and operation of the system while retaining flexibility in the details of imputation. This would facilitate final tuning without holding up production more than necessary.

Consideration (a) resulted in a general phase structure where basically four modules are involved along with three utility sorts:

(i) CNVT is responsible for identifying that subset of the file that is to be involved in imputation. For each donor or candidate it writes out an "Imputation Control Segment" (ICS) which contains match fields for donor assignment as well as space for indicating the donor actually assigned.

(ii) NEBR performs the assignment of donor to candidate on the basis of match fields. The ICS file has been stratified by sorting on a KEY. A local search is performed in a large circular buffer (about 2,000 segments) and the best match according to some measure is selected.

(iii) MERG combines a copy of the appropriate donor record to each ICS record.

(iv) IMPT then performs consistent imputation using the donors assigned.

Consistent imputation (for linear edits) was aided by a routine that kept track of the current upper and lower bounds for each field, determined by the edits and the fields already assigned. For each field to be imputed, assignment would be done if the value were in range, and the ranges of the remaining unassigned fields would then be adjusted appropriately. The routine caused the actual assignment to be made and a log entry to be written.

Where it could be applied, this approach simplified the work enormously. Unfortunately, it could not be made universally applicable without in effect solving an integer programme at each field assignment. Nonetheless, the edit rules which occurred were predominantly positivity restrictions and simple sums. Some conditional edits could be handled by selectively activating edits. Others were handled by taking great care with the imputation rules. However, the potential for an inconsistent imputation still remained.

Flexibility (consideration (f)) was ensured by allowing the control language to be a number of inclusions into the general programmes which could then be compiled to produce executable modules. The environment of each inclusion is carefully documented and service routines are provided for certain common functions.

## 5. Evaluation

The imputation procedure described in section 3 will produce estimates of the population totals (or means), but some assessment of the quality of these estimates, in terms of bias and variation, is required. One would like to know how the quality of the estimate varies with (a) the sampling bias, (b) the population size, (c) the sampling rate, (d) the correlation or relationship between the imputed variable and the auxiliary variable used for prorating, (e) the size of the window used to determine the number of eligible donors, (f) the complexity of the edits, (g) the distance function, and (h) the control of donor usage. One would also like to compare the "imputation" estimate with some natural competitors, such as the usual sampling (expansion) estimate and the ratio estimate.

A small simulation study has been done to examine the effects of sampling bias (in a nominally simple random sample) and sampling rate for a population of fixed size.

A population of 1000 units was created, each consisting of five variables corresponding to GBI, NBI and the "expense items": "salaries", "depreciation" and "purchases". GBI and NBI were the auxiliary variables. All quantities except NBI are non-negative and, in addition, we have the edit rule:

$$\text{salaries} + \text{depreciation} + \text{purchases} \leq \text{GBI} - \text{NBI} = \text{EXP}.$$

We omit the gory details, but the distribution of the non-negative variables is skewed towards zero.

Sampling was either unbiased or biased. Biased samples were created by ordering the population on GBI and (a) selecting 25% of the sample from below the median GBI and 75% of the sample from above the median GBI (bias up), or (b) reversing the percentages in (a) (bias down).

The sampling fractions were 10%, 20% and 50%.

For each sampling bias and sampling rate, twenty-five independent samples were selected from the same population. For each sample, a new file was created for the population in which GBI and NBI were retained for all records, and salaries, depreciation and purchases were included for the sampled records only. Salaries, depreciation and purchases were then imputed for the non-sampled records, using the sampled records as the hot-deck and prorating on EXP. For each replicate, the imputation, sampling and ratio estimates of the population means were calculated. These could then be compared with the known population values.

Table I gives the mean over 25 replicates divided by the population mean for each type of estimate, bias condition, sampling rate and variable. The population correlation between the imputed variable and the prorating variable is given in parenthesis in the first column. For the unbiased case, all types of estimates do quite well. For the biased cases, the imputation estimate clearly does better than the ratio estimate. The sampling estimate does very badly as one would expect.

Table II gives the coefficient of variation of the estimates in the form of the standard deviation calculated for the 25 replicates divided by the population mean. For the unbiased case, the coefficients of variation are about the same for the imputation and ratio estimates, while that of the sampling estimate is much larger. This is also true for the upward biased case. In the downward biased case, the position is less clear and the estimates appear to be roughly equivalent; but if one considers the root mean square error divided by the population mean, the bias dominates and the imputation estimate is clearly superior.

The implication of Table II is that in order to estimate the variance of an imputation estimate one may formally use the estimate of the variance of the corresponding ratio estimate as a reasonable approximation.

It will be noticed in Table I that the correlations between the imputed and prorating variables are quite high, higher than one might expect in "real" data. We would expect the difference between the imputation and the ratio estimate to become less pronounced as the correlation decreased; but no systematic work has been done to investigate this.

When the correlations are high, the size of the window appears to have no effect on the quality of the imputation estimate.

We have some evidence to suggest that when the correlations are low and the sampling rates are very low, all estimates are bad.

## 6. Conclusion

Planning for the 1975 imputation system started in April 1976 and the final output data were delivered in August 1977. Most of the delays were due to problems with data collection and survey processing. Publications based partly on the imputed data have been released.

For 1976 data the imputation system and methodology were refined and at least one survey, the Census of Construction, should run on virtually the same system with 1977 data.

Large-scale imputation appears to be a useful new weapon in the arsenal; but further evaluation should precede more widespread use. At the moment, assessment of its feasibility in any situation is based more on hunches than facts. Unfortunately, thorough and systematic evaluation promises to be a lengthy process and the best we can hope for are piecemeal results.

Table I: Percentage Bias of Estimates

| Variable (ρ) | Sampling Fraction % | UNBIASED | | | BIASED UP | | | BIASED DOWN | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Imputation | Sampling | Ratio | Imputation | Sampling | Ratio | Imputation | Sampling | Ratio |
| Salaries (.95) | 10 | -.4 | .4 | -.2 | -.5 | 32.9 | 2.6 | .5 | -31.0 | -3.9 |
| | 20 | -.2 | 0.0 | -.1 | -.3 | 33.0 | 2.9 | - .1 | -32.3 | -5.3 |
| | 50 | 0.0 | 0.0 | -.4 | -.4 | 33.8 | 3.0 | - .4 | -33.0 | -5.6 |
| Depreciation (.89) | 10 | .4 | .3 | 0.0 | .4 | 25.4 | -2.9 | - .7 | -25.4 | 4.1 |
| | 20 | .1 | 0.0 | .1 | .4 | 25.1 | -3.2 | .2 | -24.6 | 5.6 |
| | 50 | 0.0 | -.7 | .3 | .5 | 25.8 | -3.1 | .4 | -24.9 | 5.9 |
| Purchases (.82) | 10 | -.7 | .8 | .2 | 0.0 | 34.2 | 3.6 | .4 | -31.9 | -5.4 |
| | 20 | -.1 | 0.0 | -.1 | -.7 | 34.1 | 3.8 | -2.1 | -34.1 | -7.9 |
| | 50 | -.1 | -.4 | -.8 | -.5 | 34.3 | 3.4 | .6 | -34.2 | -7.3 |

Table II: Coefficients of Variation of Estimates (Percent)

| Variable | Sampling Fraction % | UNBIASED | | | BIASED UP | | | BIASED DOWN | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Imputation | Sampling | Ratio | Imputation | Sampling | Ratio | Imputation | Sampling | Ratio |
| Salaries | 10 | 3.9 | 10.6 | 3.9 | 2.4 | 9.9 | 3.1 | 4.3 | 6.6 | 4.1 |
| | 20 | 2.1 | 8.1 | 2.2 | 2.3 | 5.2 | 1.7 | 4.0 | 4.7 | 3.8 |
| | 50 | 1.0 | 2.8 | .6 | 1.1 | 3.0 | .7 | 1.8 | 1.4 | 1.2 |
| Depreciation | 10 | 3.9 | 7.8 | 3.8 | 2.4 | 4.2 | 3.1 | 4.3 | 5.0 | 4.1 |
| | 20 | 2.1 | 5.9 | 2.1 | 2.4 | 2.9 | 1.8 | 4.0 | 2.6 | 4.0 |
| | 50 | 1.0 | 2.5 | .7 | 1.2 | 1.5 | .9 | 1.9 | 1.5 | 1.2 |
| Purchases | 10 | 6.6 | 12.7 | 6.5 | 5.5 | 13.2 | 6.5 | 10.1 | 9.2 | 8.4 |
| | 20 | 3.9 | 9.1 | 4.2 | 3.9 | 7.3 | 3.7 | 7.3 | 6.6 | 6.5 |
| | 50 | 2.1 | 3.6 | 1.6 | 1.4 | 3.8 | 1.3 | 4.1 | 2.4 | 2.9 |