CALIBRATION METHODS FOR FOREIGN AND DOMESTIC INVESTMENT IN CANADA

Leon Jang and Marie-Claude Duval, Statistics Canada
Leon Jang, R.H. Coats Building, 11th Floor, Ottawa, ON, Canada, K1A 0T6

Key Words: Calibration, Country of control

1. INTRODUCTION

Auxiliary information gathered prior to survey sampling is often used by survey statisticians as an efficient method to improve the quality of their estimates while ensuring consistent results. Though normally acquired prior to sampling, this auxiliary information can be collected during sampling or transferred from an administrative source by matching to the sampling frame. Assuming that the auxiliary variable contains some intrinsic information about a variable of interest, the use of auxiliary information can be seen as particularly advantageous. Provided that the availability of an auxiliary variable is complete for the entire population or simply that population totals are available, the use of the general regression estimator allows for auxiliary information about that population to be applied. Through a set of calibration weights that respect a series of constraints, this calibration estimator has been shown to be very appealing in producing precise estimates.

Statistics Canada's Capital Expenditures Survey (CES) measures investment in fixed capital by businesses, governments and institutions within Canada. The survey sample is stratified (and survey estimates are published) at various levels by industrial classification and by geographic division. Statistics Canada was asked by end data users to provide their estimates on capital spending for several specified countries of control in order to evaluate the capital in Canada allocated for spending by foreign countries. Although the initial design stratification did not take foreign control into consideration, the country of control designation was available for all units in the target population. Therefore, it was possible to produce country level estimates on capital investment.

While information about the country of control could be directly applied to produce estimates at the domain level, this method could possibly prove to be unsatisfactory if the selected sample did not accurately represent the population in terms of country of control. However, when combined with auxiliary income information (revenue) which had also been obtained for all population units, the country of control designation allowed for calibration methods to be considered in order to provide the data users with estimates of a higher quality. Different levels of calibration that make use of the auxiliary income are possible, including calibration methods based upon cell totals or marginal totals. The ultimate choice of which calibration level to use can often be dependent on having enough units to contribute to producing a worthwhile estimate at the specified level.

In Section 2, a brief description of Statistics Canada's CES is given, while the concept of country of control is introduced in Section 3. The proposed calibration methods to improve the capital expenditure estimates by country of control are reviewed in Section 4 and the selected calibration procedure and comparison measures are outlined with respect to the chosen surveys. In Section 5, comparisons between calibrated and noncalibrated estimates at the industrial classification level are presented. Section 6 presents the subsequent enhancements to the CES as a result of the calibration work that had been done and a comparison of the subsequent estimates with the estimates produced with calibration. Finally, some concluding remarks are provided in Section 7.

2. THE CAPITAL EXPENDITURES SURVEY (CES)

Capital expenditure figures are a useful indicator of market conditions for the economy as a whole as well as for particular industries. In addition, information about the productive capacities of the economy such as the size and content of investment can be derived from expenditure information since they account for a large and somewhat variable proportion of gross domestic expenditures. Any changes in capital expenditures over time can give an indication about the management view of future market demand in relation to the current productive capacity.

The CES collects data on intended capital spending for the coming year and on actual expenditures for the previous year for the economy in Canada, covering all types of Canadian industries. Capital expenditures are the gross expenditures on fixed assets used in the operations of an organization or for lease or rent to others. These include the cost of procuring, constructing and installing plants as well as machinery and equipment, whether for replacement of worn or obsolete assets or for additions to existing assets. Also included are capitalized costs such as feasibility studies, architectural, legal, installation and engineering fees, and the value of any capital assets.

Among the types of expenditure estimated are capital expenditures for new construction (henceforth abbreviated as CC) and capital expenditures for new machinery and new equipment (henceforth abbreviated as CM). CC refers to new expenditures incurred for construction, modifications, additions and major renovations, conversions and alterations to permanent structures such as plants, office buildings, roads, bridges, electric power lines, etc. CM refers to new expenditures incurred for machinery and equipment which use some type of energy to complete work or a work environment for people. This represents the cost of items such as automobiles, professional and scientific equipment, furniture, motors, compressors, generators, etc. The sum of CC and CM is often referred to as total capital.

All Canadian businesses and governments from all provinces and territories in Canada are targeted in this survey. The target population of over 2,000,000 establishments is stratified by industrial classification (using the North American Industry Classification System), geographic classification (the 10 Canadian provinces and three Canadian territories) and three size factors based upon income. A sample of approximately 29,000 units is then selected and used to produce estimates for three distinct surveys covering three reference years: the Actual survey for the previous year, the Preliminary Actual survey for the current year and the Intentions survey for the coming year.

Questionnaires are mailed out to all sampled units. Once all the data has been received from respondents, some consistency editing is done for each establishment. This includes consistency edits for complete responses and validation of investment data. Missing data for the two variables of interest (CC and CM) is imputed using one of three methods: imputation by historical value from a previous survey, imputation by current ratio (based on the investment-to-income ratio for the stratum) and imputation by current mean (based on the average of values among units in the same stratum). It should be noted that any units deemed to be outliers during data editing were considered valid but were not permitted to be used to impute other units.

Estimates are produced for the economy in Canada at large as well as by sector (various aggregations of industrial classification levels) and geographic classification for CC, CM and total capital. The ratio estimator is used for estimation with income as the auxiliary variable. Using this method ensures that the final sampling weight for each unit multiplied by its income corresponds to the known total income for the population at each estimation level. Variances are estimated using Taylor's linearization formula in the case of the ratio estimator.

Three surveys are published at a time to allow for a comparison among the Actual, Preliminary Actual and Intentions surveys, covering three distinct reference years. Additionally, in order to comply with Statistics Canada's mandate of non-disclosure of information provided by respondents, confidentiality rules are used to suppress any information in the publication that might lead to disclosure of data supplied by a respondent.

Further information on the CES and its concepts and methodology can by found in the annual survey publication (Statistics Canada, no. 61-205-XIB, 2004).

3. COUNTRY OF CONTROL

The control of a company is based on the potential to make strategic decisions for a business. While some companies are rather autonomous from their parent corporation, it is the potential to control on which the country of control classification is based. In general, this control is viewed as having the ability to select a majority of the board of directors of a company. Usually, ownership of voting equity is used to determine the ability to elect a board of directors. This concept of control is used in determining foreign control and assigning a country of control to a company.

In most cases, the country of control classification is the country of residence of the ultimate parent corporation. This is based on tracing ownership links to the ultimate parent corporation. As a result, it is sometimes possible for a parent of a corporation to be controlled by another corporation residing in a different foreign country. In this case, the country of control of this ultimate parent is assigned to the corporation. Each subsidiary within the global enterprise is assigned the same country of control as its parent.

If one corporation owns more than 50% of the voting shares of a company, there is a majority voting ownership and the country of control for that corporation is assigned to the company. In cases where minority control exists, the owner of a block of equity has at least 33% of the voting rights and which exceeds the sum of the next two largest blocks. The situation of minority control is also sufficient enough to assign a country of control to a company. However, the overall percentage of cases where control has been assigned due to minority control is relatively low.

When a company's voting rights are owned equally by Canadian and foreign controlled corporations, it is assigned the foreign country of control. Therefore, Canadian corporations must own over half of the voting rights for a company to be designated with a Canadian country of control. If two foreign controlled corporations jointly own an equal amount of the voting rights, the country of control is assigned according to an order of precedence based on their aggregate level of foreign direct investment in Canada.

A total of nine countries were specified by Statistics Canada's data users for which estimates were required. These countries were Canada, the United States, Germany, France, the United Kingdom, Italy, Japan, the Netherlands and Sweden. These nine countries represent 99.8% of units within the CES target population and over 98% of total income in the same population. Any units with a country of control not included among these nine countries were aggregated into a tenth category representing all other countries.

Country of control level estimates at different sectors and geographic classifications for the Actual 2000 CES can be found in the published report (Statistics Canada, no. 61-232-XIB, 2003).

4. METHODS OF CALIBRATION

For the purposes of this paper, only the 2000 and 2001 Actual surveys for the CES were considered for each sampling year.

After combining information from the survey and from administrative sources, there are a couple of known factors to be noted. First, the CC and CM investment totals have been published at the sector level. These are the totals to which all sector level estimates should correspond. Second, the income totals can be calculated at the country of control level. With this knowledge, the goal is to find the CC and CM investment totals at the country of control level.

We denote the population by $U = \{1, \ldots, k, \ldots, N\}$ which is partitioned into H independent strata. The sample chosen from the population is $s = \bigcup_{h=1}^{H} s_h$ where s_h represents the sample in each stratum h. For each investment variable, we wish to estimate the population total $Y = \sum_{k \in U} y_k$ where y_k is the value of y for any unit k. We estimate Y by $\hat{Y} = \sum_{k \in s} a_k y_k$ where a_k represents the design weight defined as $1/\pi_k$ and takes into account the corresponding stratum h.

4.1 Domain Estimation

Aside from using calibration techniques, the country of control information can be applied to the existing estimates by means of domain estimation. Even though the country of control was not used in the sample design, estimates in these domains can be produced.

Let d denote the new domain of interest (country of control). Then, a new variable y_{dk} can be used in place of y_k as follows:

$$y_{dk} = \begin{cases} y_k & \text{if } k \in d \\ 0 & \text{if } k \notin d. \end{cases}$$

All other facets of domain estimation for the country of control remain the same as with the estimation of the total Y using $\hat{Y}$. The estimated total of y in domain d is given by $\hat{Y}_d = \sum_{k \in s} a_k y_{dk}$.

However, it is possible for the sample to be unrepresentative of the country of control distribution within the entire population. This will result in domain estimates that do not accurately reflect the population. Therefore, calibration methods that use the income totals at the country of control level are considered.

4.2 Calibration Estimation

The auxiliary information can be introduced to the existing estimates by using the calibration estimator. With the motivation of efficiently using auxiliary information, calibration estimators attempt to put a restriction on the final weight $a_k g_k$, where g_k represents the control or g-weight, in order to satisfy certain constraints based on the auxiliary variable. This is done by producing a new set of weights that are as close as possible to the design weights.

If the population total of the auxiliary income variable is X, the g-weights for the sampled units satisfy the estimated population total for the auxiliary variable which is the same as the known total.

$$\hat{\boldsymbol{X}} = \sum_{k \in s} a_k g_k \boldsymbol{x}_k = \boldsymbol{X}$$

When calculating g-weights, it is possible that there are units with negative or zero g-weights. From an analytical perspective, this is difficult to interpret and often impossible to accept. One method of avoiding this phenomenon is to control the magnitude of the g-weights so that the adjustment is as close as possible to 1. In turn, this results in the final weight $w_k = a_k g_k$ being similar to the sample design weight a_k. This is done by introducing constraints known as calibration equations. In terms of the final weights w_k, they can be expressed in the following form.

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \boldsymbol{X}$$

If the calibration equations are consistent, there are generally an infinite number of sets of w that will satisfy the equations. In order to find a single set of final weights, distance measures can be introduced as an objective method to define an optimal set that ensures the final weights are as close as possible to the design weights.

Statistics Canada has developed a product called the Generalized Estimation System (GES) that is capable of calculating calibration estimates by producing a set of calibrated final weights. The distance measure D used in GES is the generalized least squares function where

$$D = \sum_{k \in s} \frac{c_k (w_k - a_k)^2}{2a_k}$$

with c_k interpreted as a specified unit positive factor. Detailed explanation of this factor and the distance measure can be found in the documentation for GES (Statistics Canada, internal document, 1997).

If we consider the income information at the country of control level, we have a different perspective of the auxiliary variable that can be used to improve the weights. This can be interpreted by using several calibration groups U_i which can be different from their respective stratification groups.

All units in the sample have a known income x_k and since we also have the income value for every unit in the population, the auxiliary totals X_i for each U_i can be calculated as well. These factors are used as the main constraint in the calibration equations. With our set of design weights a_k and specified positive factors c_k, the set of calibrated final weights w_k is obtained by minimizing

$$D = \sum_{k \in U_i} \frac{c_k (w_k - a_k)^2}{2a_k}$$

with respect to w_k subject to $\sum_{k \in s_{U_i}} w_k \boldsymbol{x}_k = \boldsymbol{X}_i$.

Applying the calibrated final weights, the estimated total of y in domain d now becomes $\hat{Y}_d = \sum_{k \in s} w_k y_{dk}$ and reflects the auxiliary income and country of control information.

4.3 Calibration Groups

There are several different ways in which calibration groups can be selected in order to represent the country of control. Since we have available the auxiliary income variable for all population units, it is possible to construct calibration groups in any fashion and to have corresponding income totals for those groups. However, the ten specific categories for the country of control should form the basis for the calibration groups since the expenditure estimates are desired by the end data users for these categories.

One possible grouping method is calibrating on known cell counts N_{U_i} , which is distinguished as "complete poststratification" by Deville and Särndal (1992). Since we want to use the auxiliary income variable rather than the counts themselves, we extend the calibration to be performed on known cell income totals X_{U_i} . With the expenditure estimate totals already available by sector, U_i is defined to be at the intersection between the sector and the country of control. We identify this method of calibration as "cell level calibration".

Table 4.1 gives an example of cell level calibration using a reduced level with four sectors and five country of control categories. In this example, there are 20 calibration groups and the income total is known for each of them.

Table 4.1 – Cell Level Calibration

	Country of Control				
Sector	CAN	USA	GER	JPN	Other
A	X_{U_1}	X_{U_2}	X_{U_3}	X_{U_4}	X_{U_5}
B	X_{U_6}	X_{U_7}	X_{U_8}	X_{U_9}	$X_{U_{10}}$
C	$X_{U_{11}}$	$X_{U_{12}}$	$X_{U_{13}}$	$X_{U_{14}}$	$X_{U_{15}}$
D	$X_{U_{16}}$	$X_{U_{17}}$	$X_{U_{18}}$	$X_{U_{19}}$	$X_{U_{20}}$

A drawback to this type of calibration is when there are some cells with only a few units or with small income values. Sometimes, it becomes difficult to solve the calibration equations or can result in negative calibration weights which are generally undesirable. Among the 29,000 sampled units, fewer than 4,000 have a non-Canadian country of control and, of these, over 2,000 have a United States designation. With these units spread thin among the countries of control and 20 sectors, there were numerous small cells. Therefore, it was decided to take a less ambitious approach to forming calibration groups.

Deville and Särndal (1992) discuss an alternative to this problem by looking at calibration on known marginal counts, which they term "incomplete poststratification". In our case, we extend this calibration on known marginal income totals instead of counts. U_i now depends only on the ten country of control categories so there are fewer calibration groups and more sampled units, and in turn more auxiliary income in the sample, for each calibration group. There is a better chance for an optimal solution to be found for the calibration equations and, as a result, negative calibration weights are usually avoided. We accept this method of calibration and identify it as "marginal level calibration".

Table 4.2 presents an example of marginal level calibration using the same reduced level seen earlier with four sectors and five country of control categories. This example produces five calibration groups (instead of 20 with cell level calibration) and once again, the income total is known for each group.

Table 4.2 – Marginal Level Calibration

	Country of Control				
Sector	CAN	USA	GER	JPN	Other
A					
B					
C					
D					
Total	X_{U_1}	X_{U_2}	X_{U_3}	X_{U_4}	X_{U_5}

Having only a few sampled units for some countries of control in the 2000 and 2001 surveys, further refinement of the selected calibration method was necessary. Initial attempts at marginal level calibration using all country of control categories were compared with the auxiliary income totals for these same calibration groups. An examination of the results showed some country/sector cells with one or two sampled units were quite large or small, resulting in an overestimate or underestimate in the country of control expenditure total. To correct for this, the ten categories for country of control were reduced to seven categories:

- Canada
- United States
- Germany
- Japan
- Great Britain and Sweden (overestimation using income)
- France, Italy and The Netherlands (underestimation using income)
- Other countries

By using only seven countries of control, there are fewer calibration constraints to respect with the same amount of units in the sample. As such, more units were available in the country categories and better estimates could be produced for these countries.

4.4 Variance Estimation

The calculation of a variance for the calibration estimator is complex. However, an estimate of this variance is calculated using the GES and the procedure behind this estimate can be found in the documentation (Statistics Canada, internal document, 1997). As well, Arcaro (1998) presents the variance estimation using GES for a two-phase sample which can be applied to our estimation when looking at the first phase example.

A design based model for variance estimation is used that requires the calculation of residuals under the model. This residual for each sampled unit e_{dk} is then calculated for all units, including those not in the domain, and is expressed as

$$e_{dk} = y_{dk} - \mathbf{x}'_k \left(\sum_{k \in s} \frac{a_k \mathbf{x}_k \mathbf{x}'_k}{c_k} \right)^{-1} \sum_{k \in s} \frac{a_k \mathbf{x}_k y_{dk}}{c_k} .$$

The variance of the calibration estimator can be expressed as follows.

$$\hat{V}(\hat{Y}_d) = \sum_{k \in s} \sum_{l \in s} \frac{\Delta_{kl}}{\pi_{kl}} (a_k g_k e_{dk})(a_l g_l e_{dl})$$

$$\text{where } \Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k \pi_l & \text{if } k \neq l \\ \pi_k (1 - \pi_k) & \text{if } k = l \end{cases}$$

In the formula above, π_{kl} denotes the joint selection probability of units k and l and $\pi_k = 1/a_k$ and $\pi_l = 1/a_l$ denote the individual selection probabilities for units k and l respectively.

As a measure of the variance relative to the estimate, the coefficient of variation (CV) is used to determine the precision of the expenditure estimates. For each estimate, the corresponding CV is defined as

$$CV(\hat{Y}_d) = \frac{\sqrt{\hat{V}(\hat{Y}_d)}}{\hat{Y}_d} .$$

When the capital expenditure estimate is zero, both the variance and the CV are undefined. The CV will be used as a key statistic in determining the efficiency gains with the calibration method.

5. RESULTS

For both of the capital expenditure variables in each of the 2000 and 2001 CES, marginal level calibration was performed with the previously outlined seven country of control categories as the calibration groups. Using the GES, estimates and CVs were obtained for each country of control by sector cell. For the purposes of determining efficiency gains in the method of applying the country of control, the CV from marginal level calibration CV_{MLC} for each cell was compared to the CV for the same cell from domain estimation CV_{DE}.

The comparison of CVs can be categorized into three groups: where marginal level calibration performs better than domain estimation ($CV_{MLC} < CV_{DE}$), where marginal level calibration performs worse than domain estimation ($CV_{MLC} > CV_{DE}$) and where marginal level calibration performs as well as domain estimation ($CV_{MLC} = CV_{DE}$). When the difference between CV_{MLC} and CV_{DE} is less than or equal to 0.1%, marginal level calibration is deemed to be as efficient as domain estimation. Table 5.1 gives some results in term of counts and relative percentages for the comparison between CV_{MLC} and CV_{DE} for the two expenditure variables and for both surveys.

Table 5.1 – CV Comparison
Country of Control by Sector Level

	2000		2001	
	CC	CM	CC	CM
Better $CV_{MLC} < CV_{DE}$	15 (14.3%)	35 (31.3%)	7 (6.1%)	25 (20.7%)
Same $CV_{MLC} = CV_{DE}$	35 (33.3%)	38 (33.9%)	53 (46.1%)	53 (43.8%)
Worse $CV_{MLC} > CV_{DE}$	55 (52.4%)	39 (34.8%)	55 (47.8%)	43 (35.5%)

Table 5.1 shows that a large proportion of cells at the country of control by sector level perform neither better nor worse when comparing marginal level calibration to domain estimation. Between 33% and 34% of cells in 2000 and 44% and 46% of cells in 2001 have a CV that is similar when comparing both estimation methods. However, there are more cells where domain estimation produces a lower CV in comparison to marginal level calibration. It appears that this difference is quite evident for the CC expenditure variable, where 52% of cells in 2000 and 48% of cells in 2001 have a worse CV from marginal level calibration than from domain estimation.

Although there are more cells where the CV is higher with marginal level calibration, one should consider if the magnitude of change for cells that improved following calibration (CV_{DE} - CV_{MLC}) is better than for cells that had a better CV with domain estimation (CV_{MLC} - CV_{DE}). Table 5.2 gives the average CV changes among cells where CV_{MLC} was better and where it was worse for the two expenditure variables for both surveys.

Table 5.2 – Average CV Change
Country of Control by Sector Level

	2000		2001	
	CC	CM	CC	CM
Better $CV_{MLC} < CV_{DE}$	0.5%	3.1%	0.2%	6.9%
Worse $CV_{MLC} > CV_{DE}$	3.4%	3.6%	2.5%	2.5%

In Table 5.2, we see that among the country of control sector level cells where CV_{MLC} was worse than CV_{DE}, the marginal level calibration CV for the CC expenditure variable was larger on average by 3.4% in 2000 and 2.5% in 2001. For cells where CV_{MLC} was better than CV_{DE}, the marginal level calibration CV was nearly identical to the domain estimation CV for both years. However, for the CM expenditure variable, the average difference when CV_{DE} performed better than CV_{MLC} was about 3%. For cells where CV_{MLC} did better than CV_{DE}, the average difference was 3.1% in 2000 and 6.9% in 2001. This average difference along with the number of cells showing an improved CV demonstrates that marginal level calibration appears to be more effective for CM than for CC.

With marginal level calibration, there appear to be a few limitations that prevent this method from further improving the expenditure estimates. Since the representation of the country categories that were used as calibration groups was not considered during sampling, it happens that some cells do not have any units sampled. So, in this case expenditure values that should belong to these cells are distributed among the represented cells in the sample. As well, the large number of cells where CV_{MLC} was worse than CV_{DE} tends to suggest a poor relationship between the auxiliary income variable and the two expenditure variables.

6. ENHANCEMENTS TO THE SURVEY

This work to improve the country of control level estimates was initially done in response to demands from end data users. While the selected calibration method made effective use of auxiliary income information, we found that the method was not overly efficient in producing satisfactory estimates. In order to improve the quality of any future estimates, the country of control designation needed to be adequately represented by the selected sample.

Beginning in 2002, the CES implemented country of control into the stratification along with the existing industrial classification, geographic classification and size factors. In total, there were 12 countries of control: the nine previously specified countries, Switzerland, Hong Kong and a category for all other countries. An additional 4,000 units were added to the sample in order to account for country of control in the stratification, mainly among the non-Canadian and non-American strata. However, other improvements were made to the CES sample design in order to maintain the overall sample size at 29,000 units.

Table 6.1 gives a comparison between the cell level CV from 2002 estimation with the country of control represented by stratification (CV_{2002}) and from 2001 estimation with the country of control represented by calibration (CV_{2001}) for the two expenditure variables. The counts in Table 6.1 include only the country by sector cells that appear in both years.

Table 6.1 – CV Comparison
Country of Control by Sector Level

	CC	CM
Better $CV_{2002} < CV_{2001}$	55 (48.7%)	49 (40.5%)
Same $CV_{2002} = CV_{2001}$	26 (23.0%)	33 (27.3%)
Worse $CV_{2002} > CV_{2001}$	32 (28.3%)	39 (32.2%)

The results in Table 6.1 indicate that 49% of cells for the CC variable and 41% of cells for the CM variable have CV_{2002} that was less than CV_{2001}. Conversely, there are 28% of cells for CC and 32% of cells for CM where CV_{2002} was greater than CV_{2001}. This suggests a gain in efficiency when accounting for the country of control during stratification in 2002. Furthermore, there are no longer any unrepresented countries of control at the sector level so an estimate can be produced for every cell.

7. CONCLUSION

Using calibration methods to produce expenditure estimates at the country of control level is reasonable. Applying cell level calibration would have been the favoured method but was not considered due to the number of cells with zero or small auxiliary variable values. Marginal level calibration proved to be more feasible but did not produce the expected efficiency gains. With improvements to the selected calibration groups and to the auxiliary variable, this calibration method would likely produce estimates of improved quality.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Ann Lim and John Foley for their useful comments and insight. They would also like to thank Wesley Yung for his encouragement and continued support.

REFERENCES

Arcaro, C. (1998). GES Estimation Specifications for Two-Phase Sampling with Auxiliary Information. Business Survey Methods Division, internal document, Statistics Canada.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Statistics Canada (1997). Generalized Estimation System, Methodology Information (v4.0). Business Survey Methods Division, internal document, Statistics Canada.

Statistics Canada (2003). Foreign and domestic investment in Canada, Report March 2003. Statistics Canada catalogue no. 61-232-XIB.

Statistics Canada (2004). Private and public investment in Canada, Intentions 2004. Statistics Canada catalogue no. 61-205-XIB.