

## A STUDY OF SAMPLING AND ESTIMATION STRATEGIES FOR THE REDESIGN OF THE MONTHLY SURVEY OF MANUFACTURING

H.Lee, M.Majkowski and C.Duddek<sup>1</sup>

### ABSTRACT

The Monthly Survey of Manufacturing (MSM) is being redesigned. A study was conducted to identify the best sampling and estimation strategy to use in the redesigned MSM. The sampling methods studied were: (1) the cumulative square root of  $f$  rule, which is the current method, (2) the Lavallée-Hidiroglou method, which is good for skewed populations, and (3) some model-based stratifications. Based on the study, the Lavallée-Hidiroglou method performed the best. Various ratio estimators were also studied.

KEY WORDS: Lavallée-Hidiroglou method; cumulative square root of  $f$  rule; model-based stratification; separate ratio; combined ratio

### RÉSUMÉ

L'Enquête mensuelle des industries manufacturières (EMIM) fait actuellement l'objet d'un remaniement. Une étude a été effectuée dans le but d'identifier la meilleure stratégie d'échantillonnage et d'estimation à utiliser lors du remaniement de EMIM. Les méthodes d'échantillonnage étudiées étaient: (1) la racine cumulée de  $f$ , qui est la méthode utilisée actuellement, (2) la méthode de Lavallée-Hidiroglou, qui est une bonne méthode pour les populations asymétriques, et (3) stratification basée sur le modèle. Selon cette étude, la méthode donnant le meilleur résultat est celle de Lavallée-Hidiroglou. Différents estimateurs par le ratio ont aussi été étudiés.

MOTS CLÉS: Méthode de Lavallée-Hidiroglou; méthode de la racine carré cumulée de  $f$ ; stratification basée sur le modèle; ratio séparé; ratio combiné.

### 1. INTRODUCTION

The Monthly Survey of Manufacturing (MSM) is a sample survey of about 11,300 manufacturing establishments collecting monthly data on shipments, inventories, orders and progress accounts. The estimate of manufacturing production derived from the survey constitutes a substantial portion of the monthly estimate of Gross Domestic Product.

The sample is stratified by province (PROV) and by 4-digit Standard Industrial Classification (SIC4). The resulting strata are called the basic-strata. Each basic-stratum is further stratified using a size measure based on the shipment value from the Annual Survey of Manufactures (ASM). Benchmarking is performed once a year using the most recent ASM data.

The current design is about 30 years old and needs to be updated and modified. The MSM Redesign (REMSM) project is underway.

One of the most important components of the REMSM is the sample design and estimation methodology. We conducted this study to identify the most suitable sampling and estimation methodology among a number of possibilities. The identified methodology would also have to be one that would be efficient and rather easy to implement. A lesser objective of the study was to determine the feasibility of producing reliable estimates (especially for the inventory variables) at the SIC4 by PROV level.

A total of four sampling and estimation strategies were studied. In Section 2, descriptions of the considered methods are given as well as a listing of the basic assumptions for the study. The results from implementing these strategies are presented in Section 3. Finally, Section 4 contains the recommendation for the sampling and

---

<sup>1</sup> M.Majkowski, H.Lee and C.Duddek, Statistics Canada, 11<sup>th</sup> floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

estimation strategy to use for the REMSM as well as some concluding remarks.

## 2. STRATEGIES STUDIED

### 2.1 Assumptions

The assumptions we made were:

- 1) To implement the strategies for six important variables from the current MSM: (i) Goods of own manufacture (GOM), (ii) Raw materials (RM), (iii) Goods in process (GIP), (iv) Finished products (FP), (v) Total inventory held (TIH), and (vi) Unfilled orders (UO).
- 2) To have the smallest domain of estimation as the SIC4 by PROV level. This level is the basic-stratum in the sample design. A common feature of the four strategies is the ability to stratify the basic-strata by a size measure.
- 3) To use the ASM shipment value as the size measure for stratification and as an auxiliary variable for ratio estimation.
- 4) To have a sample size equal to the current level (11,300 establishments).

### 2.2 Strategy 1

The first strategy considered was stratified simple random sampling (SSRS) with the cumulative square root of  $f$  rule for size-stratification and Neyman allocation (Dalenius and Hodges, 1959; Cochran, 1977, pp. 127-129). This strategy (S1) is similar to the current MSM sample design, except that stratified simple random sampling without replacement will be used instead of stratified systematic sampling.

For the time being, we assume that the size stratification is done using the GOM variable (denoted by  $y$ ). The cumulative square root of  $f$  rule, as given in Cochran (pages 127-129), is derived by assuming that Neyman allocation is used. Neyman allocation (Cochran 1977, p.98) determines the size-stratum sample size ( $n_h$ ) by

$$n_h = n \frac{N_h S_h}{\sum_h N_h S_h}, \quad (2.1)$$

where  $n$  is the sample size of the basic-stratum,  $N_h$  is the size of the  $h$ th size-stratum and  $S_h$  is the standard deviation of the  $y$ -variable for the  $h$ th size-stratum.

The stratum boundaries are calculated in such a way so that the cumulative frequency within each stratum will be made equal for all strata.

With Neyman allocation and the stratum boundaries calculated as above and assuming the unit sampling cost is the same for all size-strata, the minimum sample size that is required to meet a specified coefficient of variation (CV), say  $c$ , for each basic-stratum to estimate the population total,  $Y$ , of the  $y$ -variable is determined by

$$n = \frac{\left( \sum_{h=1}^H N_h S_h \right)^2}{(cY)^2 + \sum_{h=1}^H N_h S_h^2}. \quad (2.2)$$

Using this sample size, the size-stratum sample sizes ( $n_h$ ) are then determined by the allocation formula (2.1). Note that frequently  $n_h \geq N_h$ , particularly for the size-stratum with the largest units. This "over-allocation" is dealt with by making the size-stratum a take-all stratum.

Since the population  $y$ -values are not available, the formulae given above cannot be used directly. Normally, an auxiliary variable ( $x$ ), which is highly correlated with the  $y$ -variable, is used instead. Therefore, in the formulae given above,  $y$  should be replaced by  $x$ . It is then important to understand that the optimization is with respect to the Horvitz-Thompson (H-T) estimator for the  $x$ -variable and thus, an achieved CV can be quite different from the specified CV.

Since the regression model passing through the origin fitted well for the GOM variable and the auxiliary variable from ASM, it seemed obvious that a ratio estimator should be used. Three different ratio estimators were studied: a separate ratio estimator and two combined ratio estimators, one combined over take-some strata only and the other over all size-strata including the take-all. We also included the H-T estimator as the base estimator. The criterion to compare these estimators was the variance.

For strategies 1 and 2 (see section 2.3), we studied the following four estimators.

$$\begin{aligned} \hat{Y}_{HT} &= \sum_{h=1}^{H-1} \hat{Y}_h + Y_H & (i) \\ \hat{Y}_{SR} &= \sum_{h=1}^{H-1} \frac{\hat{Y}_h}{\hat{X}_h} X_h + Y_H & (ii) \\ \hat{Y}_{CR1} &= \frac{\sum_{h=1}^{H-1} \hat{Y}_h}{\sum_{h=1}^{H-1} \hat{X}_h} (X - X_H) + Y_H & (iii) \\ \hat{Y}_{CR2} &= \frac{\sum_{h=1}^{H-1} \hat{Y}_h}{\sum_{h=1}^{H-1} \hat{X}_h} X & (iv) \end{aligned} \quad (2.3)$$

where  $\hat{X}_h$  and  $\hat{Y}_h$  are the simple expansion estimators for the stratum totals  $X_h$  and  $Y_h$ .  $X$  is the population total of the  $x$ -variable.

The estimators in (2.3) are respectively referred to as: (i) Horvitz-Thompson, (ii) separate ratio, (iii) combined ratio 1, and (iv) combined ratio 2.

### 2.3 Strategy 2

Strategy 2 (S2) is based on stratified simple random sampling (SSRS) with the Lavallée-Hidiroglou method and power allocation (Lavallée and Hidiroglou, 1988).

The Lavallée-Hidiroglou method uses an algorithm that is a combination of two earlier algorithms. The first is an algorithm developed by Sethi (1963) that optimally stratifies a continuous distribution into a number of take-some strata. The second is an algorithm by Hidiroglou (1986) which divides a skewed finite population into a take-all stratum and into one take-some stratum so as to minimize the overall sample size for a specified CV. Hidiroglou's algorithm assumes the H-T estimator with simple random sampling in the take-some stratum.

The Lavallée-Hidiroglou method has been developed to stratify highly skewed populations. An iterative algorithm is used and its objective is to determine optimal stratification boundaries, with respect to the H-T estimator, which split the population into a take-all stratum and a number of take-some strata. Simple random sampling for the take-some strata is assumed. Given the CV of the H-T estimator and the allocation scheme of the sample to the take-some strata, the stratification boundaries are computed so as to minimize the overall sample size.

The solution to the algorithm will minimize the overall sample size  $n$  for a given coefficient of variation  $c$  and a specific allocation scheme labeled as  $a_h$ . The allocation scheme chosen for S2 is power allocation. The formulae for this strategy are the following:

$$n = N_H + \frac{\sum_{h=1}^{H-1} N_h^2 S_h^2 / a_h}{(cY)^2 + \sum_{h=1}^{H-1} N_h S_h^2} \quad (2.4)$$

where  $a_h$  defined by a power allocation ( $P$ ) is

$$a_h = \frac{Y_h^P}{\sum_{h=1}^{H-1} Y_h^P}, \quad 0 < P < \infty. \quad (2.5)$$

Here, again,  $y$ -values are not available at the population level and must be replaced by the auxiliary variable ( $x$ ). The stratum boundaries will be derived in terms of the  $x$ -variable. The optimality of the boundaries will diminish for variables that are not well correlated with the

auxiliary variable. The estimators given in (2.3) were also examined under this strategy.

### 2.4 Strategy 3

Strategy 3 (S3) is based on a model-assisted approach (Kott, 1995; Sigman and Monsour, 1995, pp. 133-152) to stratification and allocation. Stratified simple random sampling without replacement is the sampling procedure for this strategy. S3 assumes from the beginning that stratification uses an auxiliary variable and that estimation is through a ratio estimator.

Suppose that there is a model that describes the relation between the  $x$ - and  $y$ -variables in a basic-stratum as follows:

$$\xi : y_i = \beta x_i + \varepsilon_i \quad (2.6)$$

where  $V(\varepsilon_i, \varepsilon_j) = x_i^{2q} \sigma^2$  if  $i = j$  and  $=0$ , otherwise. The allocation rule that minimizes the anticipated variance (the model expectation of the design variance) of the combined ratio estimator, ignoring the cost factor, is given by

$$n_h \propto N_h [M_h(x^{2q})]^{1/2}, \quad (2.7)$$

where  $M_h(x^{2q})$  is the mean of  $x^{2q}$  in size-stratum  $h$ . The combined ratio estimator is appropriate in this situation since one model is assumed for all size-strata.

Wright (1983) proposed another procedure based on the sample allocation given by

$$n_h \propto \sum_{i=1}^{n_h} x_{hi}^q. \quad (2.8)$$

This allocation is somewhat less efficient than Kott's but it is easier to implement.

This strategy was not implemented in the study primarily due to lack of time. Moreover, the results obtained from Strategy 4, which are based on the same model-based stratification, were not expected to differ much under S3. It is expected that the results under S3 would place it somewhere between Strategies 2 and 4 in terms of efficiency.

### 2.5 Strategy 4

Strategy 4 (S4) stratifies the population by the use of unequal probability sampling with the Pandher transfer algorithm (Pandher, 1996). This transfer algorithm implements the optimal probability proportional to size (PPS) sampling under the model  $\xi$  (Equation 2.6). Considering the highly skewed nature of the survey population, the algorithm first divides the population into take-all (let its size be  $n_A$ ) and take-some strata optimally for a given sample size ( $n$ ). Next it calculates a necessary

sample size to attain a specified CV with the stratum boundary determined previously and then with this sample size the optimal take-all and take-some stratification boundary is recalculated. This iterative procedure continues until the smallest sample size is obtained for the given CV. Being a model-assisted approach, it uses anticipated variance to define the desired CV ( $c$ ) as follows:

$$(cY)^2 = \sum_{i \in U_B} \left( \frac{1}{\pi_i} - 1 \right) \sigma_i^2 = \sigma^2 \sum_{i \in U_B} \left( \frac{1}{\pi_i} - 1 \right) x_i^{2q} \quad (2.9)$$

where  $U_B$  is the take-some stratum and

$\pi_i = n_B x_i^r / \sum_{U_B} x_k^r$  for some constant  $r$ . Thus,

$$n_B = \frac{\sigma^2 \sum_{U_B} x_k^r \sum_{U_B} x_k^{2q-r}}{(cY)^2 + \sigma^2 \sum_{U_B} x_k^{2q}} \quad (2.10)$$

Note that the required sample size ( $n_B$ ) for the given CV ( $c$ ) depends on the stratum boundary. On the other hand, the optimal stratum boundary depends on the sample size. Thus, the algorithm proceeds iteratively starting with an arbitrarily large sample size. For this sample size, the optimal stratum boundary is computed and then for this boundary, the required sample size is computed using (2.10). This continues until the sample size starts increasing. The stratum boundary that gives the minimum sample size is the optimal boundary. The optimal sample size is given by  $n = n_A + n_B$  at the time of stopping. Note that the take-all stratum simply consists of those units whose associated  $\pi_i = nx_i^r / \sum_{U} x_k^r$  is greater than or equal to 1.

In this study, this strategy was considered in conjunction with sequential Poisson sampling which implements PPS sampling rather easily (Ohlsson, 1995). Any basic-strata with population size less than or equal to 4 were made take-all as was done for Strategies 1 and 2. We tried three different PPS sampling schemes: probability proportional to  $x^r$  with (1)  $r = \hat{q}$ , (2)  $r = 0.75$  and (3)  $r = 0.5$ . These sub-strategies are denoted as S41, S42 and S43, respectively.

### 3. STUDY RESULTS

The results of the study were summarized at different levels of aggregation. However, in this paper we will focus on those aggregations that are regarded as the most important in terms of publication and user need. Particularly, we will place more importance on the results at the SIC4

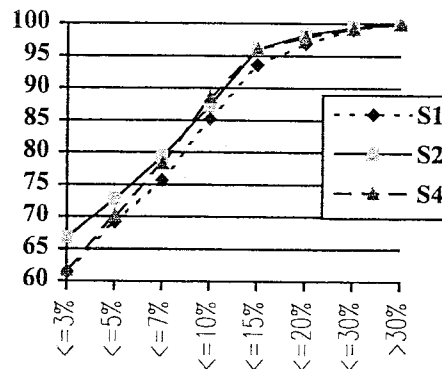
by province level since this is the level for which the basic-stratum is defined and the CV requirement is set. Moreover, it is the building block for all other levels of aggregation. The highest level of aggregation, the national level, is covered because of its economic importance. The summary data at these two levels are presented in figures and with discussion.

The main criterion to compare different strategies was the achieved coefficient of variations of the three different ratio estimators and the Horvitz-Thompson estimator as calculated using (2.3). The three ratio estimators were very similar in their efficiencies but the separate ratio was slightly better than the others. Therefore, unless otherwise indicated, the results presented are based on CV's of this estimator. Furthermore, since the results for S41, S42 and S43 were quite similar, S43 will be referred to as S4 in the remainder of this paper.

#### 3.1 Comparison of the Strategies at the SIC4 x PROV Level

There were a total of 1,588 non-empty SIC4 x PROV (basic-stratum) cells for which estimation of totals and associated CV's was required for each strategy. The achieved CV's were calculated using population values. We used the distribution of CV's to compare the strategies. The higher the percentage of cells meeting a certain CV level, the better the strategy is. Figure 1 shows the relative frequencies of cells that satisfy pre-chosen levels of CV for the GOM variable. From this figure, we can see that S2 is the best in the sense that more cells satisfy those pre-chosen CV levels for GOM. The next best was S4. It

Figure 1. Relative Frequency (%) of SIC4 x PROV Cells That Satisfy a Given CV Level (GOM)

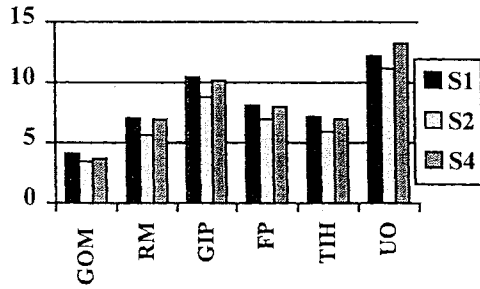


should be noted that similar figures appeared (with S2 as the best) at this level of aggregation for the other five variables studied.

The relative comparison among the three strategies can also be confirmed from average cell CV's. Figure 2 shows these averages calculated for the six studied variables. The only aberrant case is UO where S4 was the worst.

However, the figure shows that S2 had smallest average cell CV's for all the variables.

Figure 2. Average of SIC4 x PROV Level CV's (%)

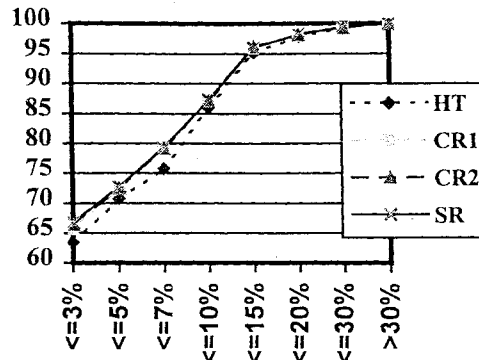


Since S2 performed the best at the SIC4 by Province level where the target CV was set, we decided to take a closer look at its performance by looking at how different estimators performed under S2. Figure 3 shows the relative frequency distribution of SIC4 x PROV cells that satisfy given CV levels for the 4 estimators we studied (see Equation 2.3).

The three ratio estimators are very close to each other in the graph even though the separate ratio is slightly better than the combined ones. However, the combined estimator is preferred for the operational reason in that it is less likely to be undefined due to non-response. The problem can be serious when the size-stratum sample size is small.

It is somewhat surprising that the H-T estimator performed quite well in the sense that it is not so much behind of the ratio estimators. At the beginning of our study, we expected that a bigger difference would exist between the Horvitz-Thompson estimator and the ratio estimators. Our explanation is that the same auxiliary data was used for both stratification and ratio estimation. This is also one of the reasons why S4, which was supposed to be "optimal", was not the best. The over-reliance of S4 on the size measure made it vulnerable when the size measure was not good. This problem associated with S4 would have been mitigated with the use of S3. However, the performance of S3 would not be better than that of S2 because of the reason given above.

Figure 3. Relative Frequency (%) of SIC4xPROV Cells That Satisfy a Given CV Level by S2 (GOM)

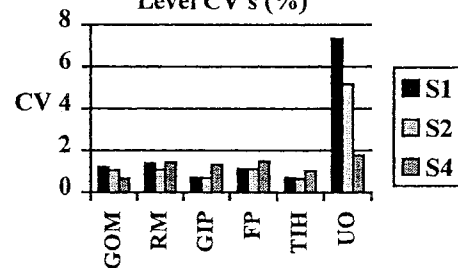


### 3.2 Comparison of the Strategies at the Canada Level

Figure 4 shows the national level CV's for the 6 variables by different strategies. According to what is displayed in the figure, S4 is the best for GOM and UO while it is worse than S1 and S2 for inventory variables (RM, GIP, FP and TIH). When S1 and S2 are compared, S2 is always better than S1.

It is intriguing to observe from the above Figure 4 that S4 is best for GOM and UO at the Canada level aggregation. This is somewhat contrary to what we saw at the SIC4 x PROV level comparison. In that comparison, the average CV under S4 was worse than S2 for GOM and for UO. Obviously, there were some less important cells in terms of total values, that had much larger CV's under S4 and this resulted in high average CV's for S4. However, CV's for more important large cells were generally smaller under S4, resulting in a smaller Canada level CV.

Figure 4. Comparison of Canada Level CV's (%)



## 4. CONCLUSION

The results obtained through this study can be summarized in the following way. At the SIC4 x PROV level, S2 performed the best in terms of average CV's and the relative frequency distribution of CV's. For S2, the percentage of SIC4 x PROV cells that had CV's less than 10% was over 87% for the GOM variable and between 75% and 80% for the other variables. At the Canada level,

S1 and S2 performed similarly, with S2 slightly better. S4 performed best for GOM and UO, but worse than S1 and S2 for the inventory variables. Different ways of forming the ratio estimate did not make much difference but the separate ratio estimator was slightly better than the combined ones. The ratio estimator was more efficient than the Horvitz-Thompson estimator in terms of average CV at the SIC4 x PROV level.

Based on these results, we recommend the Lavallée-Hidiroglou method with power allocation ( $p=0.5$ ) is recommended. Its performance was best at the SIC4 x PROV (basic-stratum) level and generally good at other levels. Another reason for its recommendation is its wide acceptance at Statistics Canada.

For the choice of estimator, we have seen that the separate ratio was slightly better in terms of relative frequency distribution and in terms of average CV at the basic-stratum level. However, we recommend the combined ratio estimator 2 that is formed over the size-strata excluding the take-all stratum. We think that the operational convenience of this estimator will more than offset the minimal loss in efficiency. This estimator has a better chance to be defined than the separate ratio when the total non-response rate is high in a small take-some size-stratum.

## REFERENCES

- Cochran, William G. (1977), *Sampling Techniques*, 3rd edition, New York: Wiley.
- Dalenius, T., and Hodges, J. L. (1959), "Minimum Variance Stratification," *Journal of the American Statistical Association*, 54, 88-101.
- Hidiroglou, M. A. (1986), "The Construction of a Self-Representing Stratum of Large Units in Survey Design," *The American Statistician*, 40, 27-31.
- Lavallée, P., and Hidiroglou, M. A. (1988), "On the Stratification of Skewed Populations," *Survey Methodology*, 14, 33-43.
- Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers," in *Business Survey Methods*, eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, New York: John Wiley, pp. 153-170.
- Pandher, Grupdesh S. (1996), "Optimal Sample Design under GREG in Skewed Populations with Application," *Survey Methodology*, 22, 199-204.
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- Sethi, V. K. (1963), "A Note on Optimum Stratification of Populations for Estimating the Population Means," *Australian Journal of Statistics*, 5, 20-33.
- Sigman, Richard S., and Monsour, Nah J. (1995), "Selecting Samples from List Frames of Businesses," in *Business Survey Methods*, eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, New York: John Wiley, pp. 133-152.
- Wright, R. L. (1983), "Finite Population Sampling with Multivariate Auxiliary Information," *Journal of the American Statistical Association*, 78, 879-884.