REDESIGN OF THE MONTHLY SURVEY OF MANUFACTURING

Ritu Kaushal, Mark Majkowski and Steven Thomas, Statistics Canada
Ritu Kaushal, Business Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6

Keywords: Establishment Survey, Survey Redesign, Multivariate Outlier Detection, Series Linkage, Parallel run, Estimation and Benchmarking.

Abstract

After three years, the redesign of the Monthly Survey of Manufacturing (MSM) has come to a conclusion with a parallel run launching the new sample into production. The redesign has included a new frame, an improved sample design, multivariate outlier detection, regression-based imputation and an integrated estimation strategy.

The frame for the new survey is the Business Register (BR), which is a list frame used by most business surveys at Statistics Canada. The redesigned survey is based on the North American Industrial Classification System (NAICS), which will soon replace the 1980 Canadian Standard Industrial Classification System (SIC80). The new sampling strategy uses improved stratification and introduces rotation. For estimation and benchmarking, an integrated approach has been used. The redesigned MSM makes extensive use of Statistics Canada's suite of generalised systems to facilitate the sampling and estimation process. The redesigned survey went into production in December 1998, initially running in parallel with the old survey for a period of eight months.

The focus of the paper will be on frame, sampling, imputation and estimation issues faced during the redesign and parallel run process.

1. INTRODUCTION

The Monthly Survey of Manufacturing is a sample survey, of about 11,500 manufacturing establishments, that collects monthly data on shipments, inventories and unfilled orders. MSM estimates are considered an important indicator of economic activity in Canada. The estimates of production derived from the survey form a substantial portion of the monthly estimates of Gross Domestic Product. Industry Canada, the Department of Finance, the provinces and manufacturing companies are the main users of the data produced.

The redesign of the MSM took place over a three-year period from the summer of 1996 to the summer of 1999. An eight-month parallel run of the redesigned MSM and the old MSM was held for the reference months of December 1998 to July 1999. Following this parallel run, the redesigned MSM replaced the old MSM as the production system.

The next section gives the reasons for the redesign. The section following it describes the new survey methodology with frame, sampling, processing and estimation issues detailed in the subsections. The last two sections present some preliminary results and conclude with further methodology issues.

2. WHY REDESIGN?

The MSM has been completely redesigned for several reasons, the main reason being that the old design had been around for about thirty years. This stale sample had units dropped from the sample when they died or were a chronic non-respondent. These units were often replaced by larger units which caused the pre-benchmarked MSM estimates to be biased (see Majkowski & Metzger 1997). In addition, units in the old survey design could have different weights for different characteristics because some units were not asked to respond for characteristics that were not available on a monthly basis. The weights for these unavailable characteristics were modified and applied only to the units that had the characteristic available each month. This procedure has caused inconsistencies to occur between the estimates for different characteristics.

Another reason for the redesign was the fact that the old processing systems needed updating. The processing systems needed to be year 2000 compliant. As well, the old processing systems, which were located on the mainframe, required a number of manual processes and they produced large numbers of printed reports to be analysed by the subject matter specialists. The new systems are paperless PC based systems.

A final reason for the redesign of the MSM was to take advantage of new initiatives at Statistics Canada affecting business surveys. One new initiative was the use of a centralised frame known as the Business Register (BR). The redesigned MSM uses the BR as a frame. Another initiative was the implementation of a new classification system known as the North American Industry Classification System (NAICS). The new sample design for the redesigned MSM is NAICS-based. A final new initiative that indirectly impacted the MSM was the

Project to Improve Provincial Economic Statistics (PIPES). Because of PIPES, the redesigned MSM had a sample selected so that the MSM provincial estimates would be improved.

3. NEW SURVEY METHODOLOGY

3.1 Frame

The old MSM was based on the frame that was created and updated yearly by the Annual Survey of Manufacturing (ASM). The ASM collects total annual shipment values and detailed commodity information on about 40,000 businesses in Canada. The ASM industrial classification was SIC80, which was based on the most current commodity mix available from the last ASM. Previous year's reported ASM shipments were used as a size criterion for the stratification of the MSM population. One of the problems with the ASM as a frame was that it was updated only annually. The ASM estimates of shipments and inventories are used to benchmark the MSM. In fact, the MSM sample itself was not updated, only the benchmark correction factors were.

Using the Business Register (BR) as the frame for the new survey is one the biggest changes between the new and the old survey design. The BR is a list frame of Canadian businesses updated by administrative data from Revenue Canada. Apart from administrative data updates and survey feedback, the BR is also updated by regular profiling of large businesses. The BR has both NAICS and SIC80 classifications and revenue as a size measure. While revenue strongly correlates with shipment, a previous year's shipment from the ASM is a more accurate measure of size and is used in the new survey where available.

The change to the BR added about 70,000 new units while adding less than 10% to the level of the estimate. However, the BR brings with it problems of using administrative data for surveys. Some of the apparent increase in coverage is likely to be duplicates, misclassified units and dead units. While this presents one of the biggest challenges to producing estimates using the BR, linkage (section 3.4) to the old series will minimize the impact of coverage problems.

3.2 Stratification, Allocation and Sampling

The new stratified simple random sample design is similar to the previous design except for new NAICS industrial classification and a different allocation scheme. The stratification, carried out using the Lavallée-Hidiroglou (Lavallée and Hidiroglou 1988) algorithm, is based on four-digit or five-digit NAICS by province by size. In order to be able to produce SIC80 estimates, units that are in SIC80 manufacturing but not in NAICS manufacturing are included in strata of their own. The size measure used is derived from the frame or the most recent ASM.

To minimise response burden on small businesses and reduce cost, different strategies are adopted for different sizes of business. Businesses were ordered by shipments and the very small businesses, constituting the second percentile, are excluded from sampling. For small to medium size businesses, a low sampling fraction (max. weight of 30) is used where possible and rotation will be instituted in the year 2000. The largest businesses within province and NAICS industry are sampled with a probability of one.

The allocation strategy was designed with an increased emphasis on industry by province estimates whereas the old sample was allocated to get accurate Canada level estimates for industries. Power allocation with a value of 0.5 was used to distribute the new sample of 11,500 units. As a result, the new sample contains on average smaller units and is distributed more finely throughout Canada. However, there is a significant overlap of about 50% between the old and new sample. This overlap consists mostly of the larger, self-representing units.

3.3 Outlier Detection and Imputation

In updating the processing systems, the outlier detection system was changed from a univariate to a multivariate approach. The multivariate outlier detection (Franklin and Brodeur 1997) is based on the calculation of the Mahalanobis distance, a multivariate distance measure, which measures an observation's distance from some measure of location.

Multivariate outlier detection has met with limited success because of the high level of item non-response. The item non-response occurs from establishments not being able to report for characteristics other than total shipments on a monthly basis. While imputation systems were also constrained by the same problem, hierarchical imputation classes based on geography, industry and size stratum were used and insufficient units resulted in the utilisation of a higher imputation class. Regression on shipments or size measure was used when previous month's data was not available.

3.4 Estimation and Linkage

With no monthly auxiliary data available, the Horvitz-Thompson (HT) estimator is used in the new design. It

provides an unbiased estimate of level and of month-to-month change and can be used in conjunction with the Denton approach for time series benchmarking (Majkowski and Kaushal 1999). Other estimators were investigated during the course of the redesign (Majkowski *et al.* 1997). In addition to the HT estimator, a linkage factor will be used to maintain continuity between the old and the new series.

A constraint for the redesign is that new NAICS estimates are required for the longer term while continuing to produce SIC80 estimates series for the next year. Hence, domain estimation is used to produce SIC80 estimates. As a result of domain estimation, the coefficients of variation (CVs) are higher compared to the NAICS estimates.

The series are linked at the lowest level of estimates published and then rolled up to produce aggregate level estimates. For shipments, the lowest level at which series are published, referred to as a cell, is detailed industry by province. To avoid a break in the ongoing series, the levels from the old series are maintained, but the trend and movement is from the new sample. In order to achieve this, the linkage factor, L_c, is calculated by taking a mean of the ratio of the estimates from the old and new sample over the months of the parallel run as given in the following equation:

$$L_c = \frac{1}{M} \sum_{m=1}^{M} \frac{\hat{Y}_{mC}^{OLD}}{\hat{Y}_{mC}^{NEW}}$$

where, $\hat{Y}_{mC}$ is the estimate from the old or new sample for month m (from 1 to M) and cell C (SIC80 cell). The linkage factor is applied at the establishment level to units in the sample during the parallel run. Birth units are given a linkage factor of one.

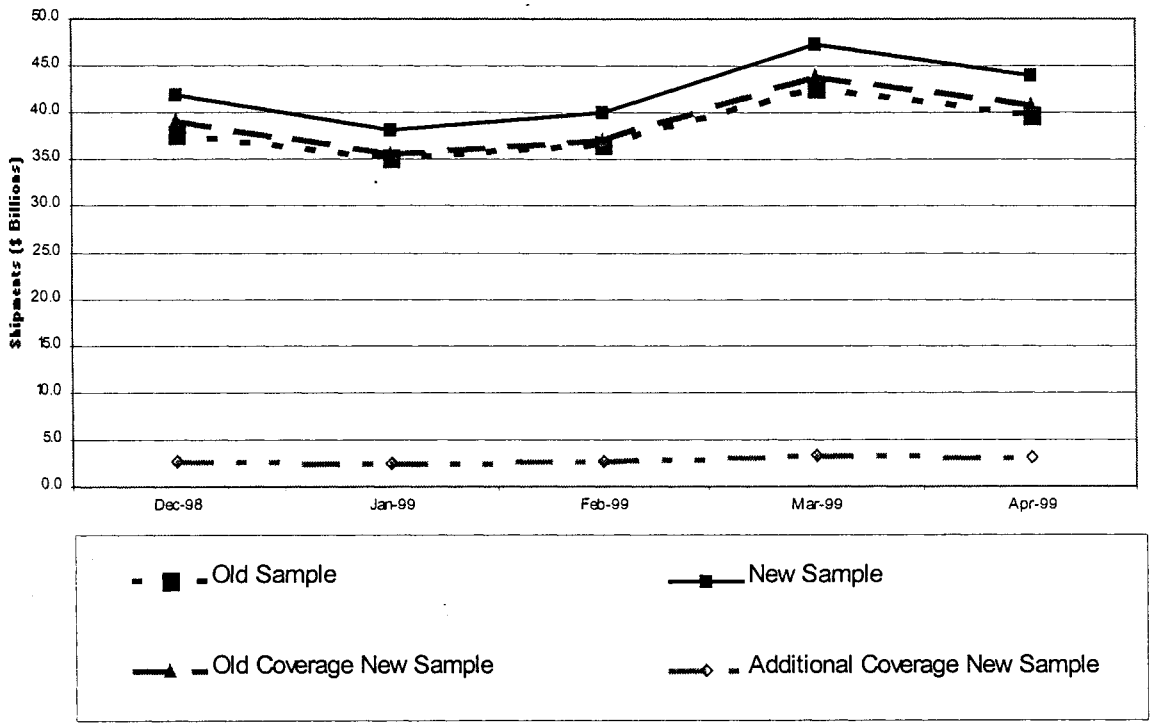
4. PRELIMINARY PARALLEL RUN RESULTS

At the time of the presentation, data from only five out of the eight months of the parallel run were available. These data were used to check the relative level of estimates, trend, movement and the variances associated with the old and new sample. The linkage factors were analysed to identify possible data problems. Graph 1 gives the estimates for Canadian shipments in billions of dollars during the months of the parallel run. The old and the new sample follow a similar movement but are consistently apart by just under three billion dollars. To investigate this difference, estimates with different domains were compared. The old sample estimates are representative of the 1996 ASM frame. A domain estimate, using the new sample, based on the population common to the new frame and the 1996 ASM gave the series referred to as "Old Coverage New Sample" in Graph 1. This series is very close to but just higher than the benchmarked old sample estimates. The difference is small but could be in part due to respondent or imputed data not investigated by subject matter experts. The last series shown in graph 1 is the contribution of the additional coverage from the new frame. The additional coverage represents some large births in manufacturing since 1996, out-of-scope units and duplication on the new frame. The analysis of the additional coverage is ongoing and the magnitude of the estimate for additional coverage is expected to decrease as the frame and data are analysed and cleaned.

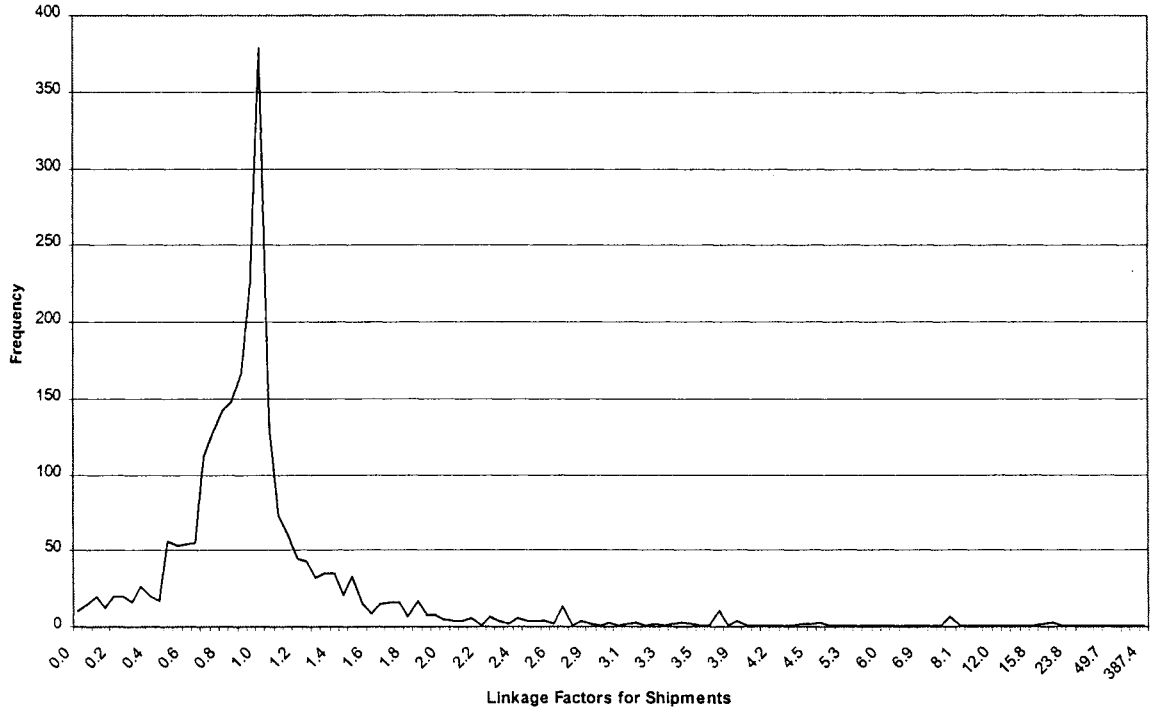
The distribution of linkage factors for all characteristics is plotted in Graph 2. As anticipated, the peak of the distribution is close to one because most series are not affected by the change in coverage. Linkage factors with extreme values, close to zero or much greater than 1, could be a result of significant changes in sampled units and data problems. With the change in population coverage and sample distribution, certain series will have units entering or leaving cells resulting in series commencing or terminating. Missing critical units, reporting problems or inaccurate imputation can also be highlighted by extreme values of factors. All extreme values are being investigated.

The final graph compares the distribution of CVs for NAICS and SIC80 cells. In Graph 3, one of the distributions is for the SIC80 using the old sample and the other two distributions, SIC80 and NAICS, are based on the new sample. All these distributions are based on industry by province cell level data for all months for the parallel run. To be able to examine the tails, zero CVs have been excluded from Graph 3. A comparison of the CVs from the three sets of estimates is difficult to make because there has been a significant change in coverage, the NAICS and SIC80 estimates are not comparable and the old sample estimates are benchmarked while the new sample estimates are not. Given these constraints, we can still draw some general conclusions. On comparing the SIC80 old design and NAICS new design, we find that distributions are similar but with the new design a larger percentage of the CVs are publishable (for example CV <10%). With the use of power allocation, more detailed estimates are of an acceptable quality without damaging the quality of higher level estimates. Including zeros, the average CVs are 5.3% and 3.5% respectively for the old and new samples. When the new estimates are benchmarked, the CVs are expected to decrease. As anticipated, we see that SIC80 cell estimates from the new sample have higher CVs than both other sets of estimates because of domain estimation.

Graph 1: Shipment Estimates from Parallel Run

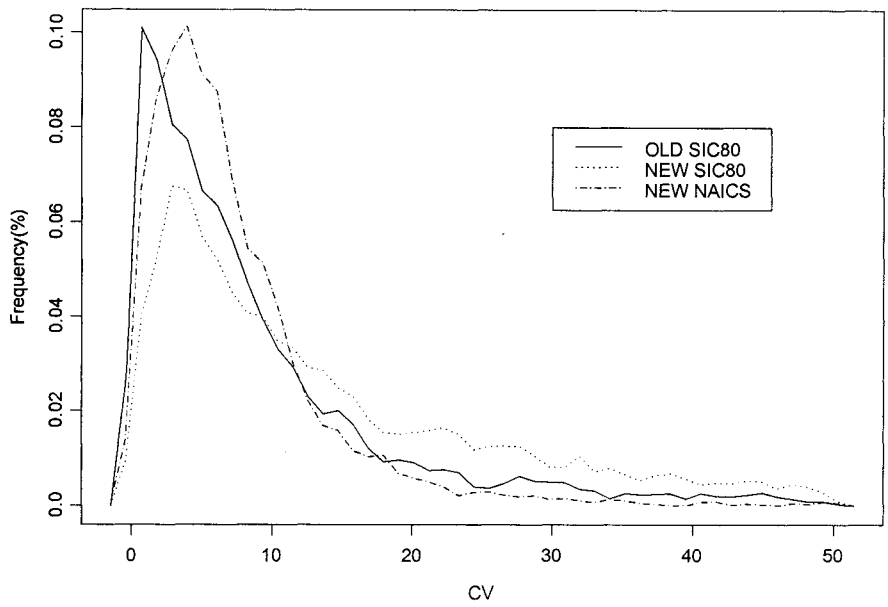


Graph 2: Distribution of Linkage Factors



Graph 3:

Parallel Run CVs



5. CONCLUSION

Upcoming methodology issues are integrated benchmarking for all characteristics and possible uses of new monthly auxiliary data. Benchmarking is currently being done one characteristic at a time. Therefore, a unit can have different weights for different characteristics. This creates inconsistencies between variables and hence, it is necessary to develop a method to get common weights. Auxiliary data are not used in estimation because the available size measures are updated annually. In the next couple of years, new auxiliary data (Goods and Services Tax collected) will become available. Various uses of this information will be investigated including ratio estimation and replacement of survey data for small businesses with administrative data.

As shown, the redesign and the parallel run have been successful. At the end of the parallel run, the SIC80 estimates based on the new survey methodology will be published in October 1999. The publication of the new NAICS series will commence in June 2000.

Acknowledgements: The authors would like to thank Mike Hidiriglou for technical advice during the redesign process; and, Harold Mantel and Pierre Lavallée for their helpful comments on this paper.

References:

Cuthill, Ian (1996). The Statistics Canada Business Register Statistics. *Statistics Canada internal document: Original August 1990, revised 1996.*

Franklin, S., and Brodeur, M. (1997). A Practical Application of a Robust Multivariate Outlier Detection Method. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 186-191.

Lavallée, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Populations, *Survey Methodology*, 14, 33-43.

Lee,H., Majkowski M., and Duddek, C. (1997). A Study of Sampling and Estimation Strategies for the Redesign of the Monthly Survey of Manufacturing, *Proceedings of the Section on Survey Research Methods, Statistical Society of Canada Conference.*

Majkowski M., and Kaushal R., (1999). Estimation and Benchmarking for Monthly Surveys, *Proceedings of the Survey Methods Section, Statistical Society of Canada Conference.*

Majkowski M., and Metzger R., (1997), Study of the Bias in Estimates, *Statistics Canada Internal Report.*