Journal of Official Statistics
Vol. 3, No. 3, 1987, pp. 283–288

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

Methodological Problems in Designing Continuous Business Surveys:

Some Canadian Experiences

K.P. Srinath[1]

Abstract: Methodological problems associated with designing continuous business surveys are similar to the problems encountered in other continuing surveys. However, there are some problems which arise primarily because of the dynamic nature of the population of businesses and are more closely associated with continuing business surveys than with other surveys. Problems relating to the construction of sampling frames, sample selection and rotation, implementation of changes in classification information, and outliers are discussed. The purpose of this article is to highlight the problems rather than provide complete solutions.

Key words: Sampling frame; rotation; outliers; classification; establishment.

1. Introduction

Measurement of economic activity in different economic sectors by government agencies for the purpose of determining levels and trends is increasingly being done through sample surveys. More often, the surveys are repeated either monthly or quarterly to produce estimates of totals, averages, ratios, and estimates of change between two periods.

The problems associated with designing continuous business surveys are similar to the ones encountered in other surveys, one-time or continuing. However, there are some methodological problems which arise primarily because of the dynamic nature of the population of businesses. The purpose of this article is to highlight some of these problems. A brief discussion of some of these problems

[1] Senior Methodologist, Business Survey Methods Division, Statistics Canada, Ottawa, Canada. The views expressed by the author in this article do not necessarily reflect the policies of Statistics Canada.

Acknowledgements: The author thanks the referees and the editor for their constructive comments and suggestions.

can be found in Finkner and Nisselson (1978), Sanyal and Sinha (1977), and Konschnik et al. (1985) and no attempt is made to provide complete solutions to these problems. Although the problems mentioned in this paper seem straightforward, they require complex solutions. When these problems are not solved reasonably well, they can lead to substantial non-sampling errors, primarily frame errors.

Problems associated with the construction of sampling frames and changes in the frame over time are discussed in Section 2. In Section 3, the impact of a changing frame on sample selection and rotation is reviewed. In Section 4, the problem of implementing changes in classification information is discussed. In Section 5, attention is focused on the problem of determining outliers and estimating totals and change in the presence of outliers. Some of the observations in the following sections are the result of the author's experience in designing the Canadian establishment based Survey of Employment, Payrolls and Hours, and experience from the methodology team responsible for the survey design. A description of the survey along with detailed methodology is given by Schiopu-Kratina and Srinath (1986).

2. Sampling Frames

The first and the most visible problem for business surveys is the construction of sampling frames when sampling directly from universe-level lists. It is very difficult, if not impossible, to construct an up-to-date list of businesses including employers and non-employers with correct names, addresses, and standard industrial classification information. A frame has to be constructed from lists obtained from many sources of varying quality and age. Therefore the sampling frames used in business surveys are not free from defects like duplication, inaccuracy, incompleteness, and extraneous units. Even if a complete list of businesses, where businesses could be enterprises, companies, or establishments, was available, it would not be sufficient for many purposes. An establishment is defined as the smallest group of production units which produces a homogeneous set of goods and services. An establishment is also capable of reporting all elements of basic industrial data permitting the calculation of "census value added" as well as providing data on employment and payrolls. A company is a business organization consisting of one or more establishments. An enterprise may consist of one or more companies capable of controlling the allocation of resources and economic activities and providing a full set of financial accounts.

When a response is received from a business it is not always clear which part of the business the response relates to. For example, if an establishment has several locations and is selected in the sample, one must make sure that the response covers the intended location(s). It may also be necessary to have a separate response for each location. This means that any list obtained from a source must show or describe the entire structure of the business, whether it is an enterprise, company, or establishment. This is called "profiling." Again this structuring could be different for different surveys or even different characteristics. For example, if the survey is on employment, earnings, and hours, establishments having more than one location or payroll may have to be structured into what may be termed as employment reporting units capable of reporting employment, earnings, and hours. If the survey is on employee compensation, the earlier structure may not be entirely suitable because certain items relating to compensation may not be available for each employee. Agencies conducting business surveys have the difficult task of systematically profiling multi-unit businesses and keeping

the structure up-to-date. The profiling has to be done on a continuing basis as companies change in composition through mergers and acquisitions. Failure to take into account the entire structure of a business could lead to underestimation of the characteristic under study.

One characteristic of the population of businesses is that a small proportion of the businesses accounts for a very large proportion of the total (employment, sales, etc.). That is, the distribution of the population is generally highly skewed. Therefore the usual practice is to stratify businesses by some measure of size and take disproportionate samples from within strata (Raj (1972)). The largest businesses are included in the sample with certainty. The sampling fraction in the next largest stratum is less than one, but high and so on. The sampling fraction in the stratum of smallest businesses is usually the lowest and very small. The stratification and allocation are done to achieve the greatest design efficiency including a reduced response burden on smaller businesses. This strategy works better in a one-time survey than in a continuing survey because of the instability of the small business population. This is especially true in construction, trade, and service industries (Plewes (1982)) where the number of small businesses is large. In this stratum, the businesses enter and leave the population in large numbers each month. Their characteristics also tend to change over time more quickly than the large businesses. These features will have to be kept in mind while designing the survey. It is important to make sure that this stratum is adequately represented in the sample, so that changes in these businesses are not missed, especially during an upturn of the economy. In addition to the large businesses, it can also be useful to keep the multi-unit businesses in the sample with certainty. This helps keep their complex structure up-to-date, and avoids the problems inherent in sampling and collecting reports from parts of businesses.

Another major problem in continuing business surveys is the sampling of new businesses or what is called "births." Often, there is a substantial time-lag, sometimes several months, between the birth of new businesses and their inclusion in the sampling frame. In addition to creating the usual problem of underestimation, this time-lag leads one to consider whether new businesses should be classified to a separate stratum. In fact, when a business becomes eligible for selection, it may no longer be new but be well established and therefore not different from other non-sampled units in the population. The contribution of these businesses to estimates of change, which could be substantial in the initial months of their existence, are also missed.

The lag between the time a business ceases to operate and its removal from the frame may be even longer than the birth time-lag, possibly running into years for the non-sampled businesses. Therefore, one is always sampling from a frame containing a large number of "out of business" or extraneous units. This leads invariably to estimates of level and change not meeting the planned reliability requirement. It is difficult to guess the number of such units (businesses) in the population especially when the economy is on a downward trend. Identification of businesses which are no longer operating takes place primarily through the sample, that is, after they are included in the sample. Even here there is a time-lag, and the businesses are treated as non-respondents in the initial months of selection. The time-lag leads to a substantial underestimation of trends in the initial months of a new survey.

3. Sample Rotation

In order to obtain precise estimates of both level and change and to reduce response burden (especially on the small businesses), it

is desirable to have sample rotation. In a monthly survey, if it is desired to have businesses in the sample for approximately 12 months, then the usual practice is to replace 1/12 of the non-certainty sample each month. This partial replacement of the sample periodically is not as simple or as straightforward as it may seem. The rotation has to be adopted under certain constraints. For example, in a monthly survey it may be considered important to keep businesses out of the sample for at least a certain number of months after they rotate out because of response burden considerations. That is, they are not eligible for selection for a certain number of months. This requirement in a continuing survey may lead to keeping some businesses in some strata for more than 12 months. This is considered desirable since it is more of a response burden to come back into the survey within a certain number of months than to continue reporting each month. The dynamic nature of the business sampling frame due to births, deaths, and changes in the classification information makes rotation more complex than in surveys with stable or unchanging frames.

Sample rotation is usually carried out by first dividing the population of sampling units into a certain number of panels or rotation groups and then dropping and adding panels at regular intervals. It is possible for rotation groups originally to be approximately equal in size, but then become unequal over time. This might affect the estimates of change, especially if there is rotation group bias. Inaccuracies in the classification information for the rotating units tend to distort the estimates of change. If the sample size is smaller than the number of panels in the sample, then this may result in a number of empty panels and no rotation at regular intervals and again affect estimates of change.

The sample that was determined for the first month of the survey to give a prespecified reliability of the estimates cannot be maintained each month. The changing sample size may not meet the reliability requirements for the subsequent months. Nor can a constant sampling fraction always be ensured. Changes in population sizes and classification information may necessitate periodic recomputation of sample sizes or sampling rates, or both.

The problem of sampling births each month and eliminating deaths also presents difficulties. If births form a separate stratum and we wish to sample these with the same low rate used for other units, then births may be underrepresented in the sample. Since births are added to the frame continuously it would be difficult to treat births which are one month old, two months old, etc. separately for sampling purposes. Therefore the non-sampled births for a particular month have to be treated as members of the general population for the next month's sample.

In one-time surveys, extraneous units, such as units which are out of business (deaths), do not cause a major problem. When a sample is selected some of these units appear in the sample. But if the value of the variate of interest is taken as zero for these units, their presence will not vitiate the results. However, the sampling error of the estimates will increase. One procedure in a continuing survey is to remove deaths only if they are removed from the frame by an independent source. That is, the source identifies units as "dead" irrespective of whether the units are in the sample or not. The procedure is simple. But since such updating occurs infrequently, deaths tend to remain in the frame for a long time, and this is a clear disadvantage. It might add to the variance of the estimate. This procedure to remove deaths may not be applicable if the number of extraneous units in the population is changing due to an updating of the sampling frame coming from various sources. Therefore it may be important to ensure a proper representation for these units

to avoid over or underestimation of totals. This may involve estimating the number of extraneous units in the frame each month from the sample and applying some kind of a correction factor to the weights. Seal (1962) proposed adjusting for changes in the population by assuming a birth and death process.

4. Changes in Classification Information

As mentioned earlier, the sampling frame is changing continuously due to births, deaths, and changes in the classification information. Changes in the industry classification, location or measure of size could be real and reflect change in the activity, location, or the size of the business. These changes can occur in either the sampled or the non-sampled portion of the population. These changes could be detected immediately or only after several months. A change in classification could also come about because the unit was originally misclassified. Misclassification is detected more often in sampled units than in non-sampled units. Again, this misclassification might only be detected after the unit has been in the wrong stratum for several months. It is difficult to devise procedures for implementing the changes that would minimize the bias in the estimates of level and change. If it is important to estimate the monthly change, then one procedure is to store these changes each month and make these changes on an annual basis.

To reclassify correctly the units at the end of the year, unless followed by a complete redraw of the sample, could lead to biased estimates without the use of complicated estimation procedures. Redrawing the sample at the end of each year may be impractical since it would distort the rotation scheme. Reclassification may also distort the estimate of change from the month in which reclassification of the units was made to the previous month. There may have to be some compromise between not changing the classification information for a certain period of time and making these changes each month. Estimation procedures that take into account such changes could add to the complexity of the whole process of producing monthly estimates on a timely basis.

5. Outliers

The problem of determining and then dealing with outliers in a survey has been extensively investigated with no satisfactory solution. Even defining an outlier seems to be a part of the problem. This problem is very common in business surveys. For example, inaccurate measures of size and low sampling fractions in the stratum containing small businesses could lead to a serious distortion of the population total when the sample total is inflated using the weight. Thus it is important to deflate the weight for outliers at the estimation stage once they have been sampled and identified. Another possibility is to alter the observed values of the outliers in order to lessen their influence on the estimate of the totals. Survey managers seem to prefer techniques that alter the weights attached to the outliers rather than changing reported data. This is especially true in the case of repeated surveys in which the same unit is in the sample more than once. There are two problems associated with changing the weight. The first is determining whether an observation is an outlier in relation to its sampling weight. The second is how to "correct" the weights once they are identified as outliers. Hidiroglou and Srinath (1981) have proposed estimators of the population total that are robust to outliers in the sample. But the problem of classifying an observation as an outlier still remains. One strategy is to treat an observation above a prespecified value as an outlier, change its weight to one

and adjust the weights of the non-outliers. This may be reasonable for estimating levels. But the same unit may turn out to be a non-outlier the next month and get the usual weight, and thus seriously distort the estimate of change. The problems of determining boundaries for outliers and altering the weights when estimating month to month change is a topic for further research. Again, how does a cut-off based on one variable affect the other variables (Bershad (1960)). Possible solutions mentioned above for the problem of outliers assume simple random sampling. The treatment of outliers in samples drawn with unequal probability would also have to be considered. More work needs to be done to provide answers to these questions.

6. Concluding Remarks

Methodological problems which arise on account of the dynamic nature of business frames have been discussed. Some problems, like the maintenance of the sample under changes in classification information require complex solutions. Problems like outliers may need solutions specific to particular surveys. A number of potential solutions to these problems exist, although they have not been discussed in this paper.

7. References

Bershad, M. (1960): Some Observations on Outliers. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census.

Finkner, A.L. and Nisselson, H. (1978): Some Statistical Problems Associated with Continuing Cross-Sectional Surveys. In Survey Sampling and Measurement, N.K. Namboodiri (Ed.), Academic Press, New York.

Hidiroglou, M.A. and Srinath, K.P. (1981): Some Estimators of a Population Total From Simple Random Samples Containing Large Units. Journal of the American Statistical Association, 76, pp. 690–695.

Konschnik, C.A., Monsour, N.J., and Detlefsen, R.E. (1985): Constructing and Maintaining Frames and Samples for Business Surveys. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 113–122.

Plewes, T.J. (1982): Better Measures of Service Employment Goal of Bureau Survey Redesign. Monthly Labor Review, 105, No. 11, pp. 7–16.

Raj, D. (1972): The Design of Sample Surveys. McGraw-Hill, New York.

Sanyal, S.K. and Sinha, S.K. (1977): Methodological Problems in Large Scale Sample Surveys – Experiences From National Sample Survey. Sankhya, Series C, 39, pp. 47–70.

Schiopu-Kratina, I. and Srinath, K.P. (1986): The Methodology of the Survey of Employment, Payrolls and Hours. Working Paper No. BSMD-86-010E, Methodology Branch, Statistics Canada.

Seal, K.C. (1962): Use of Out-Dated Frames in Large Scale Sample Surveys. In Proceedings of the 48th Session of the Indian Science Congress, Cuttuck, January, 1982.

Received March 1986
Revised September 1987