WP. 28

ENGLISH ONLY

UNITED NATION STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iii): Data editing processes within survey processing

THE USE OF A SCORE FUNCTION IN A DATA COLLECTION CONTEXT

Supporting Paper

Submitted by Statistics Canada [1]

I. Introduction

1. Statistics Canada's initiative in reducing the number of follow-ups of survey respondents due to non-response and edit response errors started in the early 1990's. Various approaches were considered, based mostly on score functions that determine priorities for collection units. Details on the use of a score function are given by Latouche and Berthelot (1990). Since then, applications have been developed and fine-tuned for a few surveys that had significant respondent follow-up costs. This paper focuses on two recent applications of score functions: The Unified Enterprise Survey (UES) and the Annual Survey of Manufactures (ASM). Both surveys utilize the fact that in business surveys, typically only a few units represent a large percentage of the population characteristics. It is believed that concentrating on the units with larger scores for priority follow-up, would not compromise the overall data quality, as long as the process is monitored. Here, the impacts on the sample design, the collection process and the quality of the end product are considered in setting the priorities of individual units.

II. The Unified Enterprise Survey

2. The Unified Enterprise Survey (UES) is an annual business survey that was created in 1997 in order to combine many surveys into one (Beelen, Royce and Hardy, 1997) while improving the quality of their results (Tourigny, Pursey and Whitridge, 2001). It covers Service industries, as well as Manufacturing,

[1] Prepared by Claude Poirier (poircla@statcan.ca), Robert Philips, and Stuart Pursey

Wholesale, Retail, Transportation, Aquaculture, and Banking industries. A total of 20 surveys are integrated into the UES with a common sampling frame, sample design, data collection, data processing and estimation.

3. All questionnaires include the same key financial variables, but with various sets of secondary variables. The questionnaires vary by industry. Due to seasonal patterns observed in the business activities, the collection periods are also customized by industry. The overall sample size represents about 65,000 establishments spread over 45,000 sampling units that are grouped into approximately 30,000 questionnaires.

4. The data collection process consists of (a) pre-contacting new enterprises to confirm their activity codes, (b) mailing out questionnaires to all selected enterprises, and (c) following-up non-responses and edit failures according to budget and time constraints. Since the sampled units do not have equal importance in terms of revenue, a good response rate does not guarantee good revenue coverage. Therefore, priority is given to units that have higher impact on the estimated revenue.

What is the score function?

5. The score function is a measure of importance. For 2002, UES introduced a score function approach to reduce its collection costs (Pursey, 2003). It is based on the sample weighted revenue of the sampled units. While the revenue is obtained from historical information or administrative sources, the weights are derived from the sample design. Within a specific industry, the strategy is to set a target threshold (in percentage terms) and to make sure the cumulated weighted revenue obtained from the respondents is above this threshold. In order to get a homogeneous coverage, this strategy is applied for each industry sub-class[2] and each province.

6. Before starting the collection process, the percentage contribution – or score – of each unit is calculated within its cell and the units are sorted in descending order given their score. Starting with the highest scores, a Priority 1 is sequentially assigned to the units until their cumulated scores reach the target threshold. All remaining units get a Priority 0. In the case where a questionnaire covers more than one unit, all its units get a Priority 1 if at least one received a Priority 1 in the first step. In the data collection process, the units with Priority 1 will be followed-up in the case of non-response or edit failure. The units with Priority 0 will receive a questionnaire but no follow-up will be attempted should problems occur with their data or they do not respond.

[2] An industry sub-class consists of a 5-digit code of the North American Industrial Classification System. Statistics Canada (1998) gives the details of this coding.

The dynamic aspect of the score function

7. As collection goes on, a questionnaire results in either (a) a response, (b) a non-response, or (c) the unit is declared out-of-scope. The out-of-scope units and the non-response units bring a dynamic aspect to the score function. Out-of-scope units reduce their cell's total revenue. If, after several attempts at follow-up, a Priority 1 unit does not respond, sufficient highest scoring Priority 0 units are selected such that they equal the non-responding units score. Therefore, the Priorities 1 and 0 units are recalculated twice a week given the responses and the changing cell's total revenue. In this process, responses contribute to the cell threshold, while non-responses contribute nothing and out-of-scope units reduce the cell revenue. The achieved threshold may therefore go up or down depending on the cell. The follow-up process is completed when the sum of the scores of the responding questionnaires reaches the cell threshold. As the collection for certain cells is completed, collection resources are reassigned to the cells that have not yet attained their thresholds.

Evaluation of the process

8. It is still too early to evaluate the benefit of the score function as the data collection is still going on. Due to the size of some specific cells, it is expected that a few will not reach their target thresholds as a result of time constraints and difficulties to change some non-responses to responses. The evaluation will identify which target thresholds may be unreachable. Another aspect of the evaluation is related to the bias. Since typically small units are not followed-up, their potential to become larger may not be observed in the case of non-response, and thus they may introduce a negative bias in the estimates. Estimation adjustments may be considered, for instance from logistic regressions. Is the score function worthwhile in terms of cost saving and are costs better optimized among the industry by province cells? These questions will be addressed as well.

III. The Annual Survey of Manufactures

9. The Annual Survey of Manufactures (ASM) is a survey that collects and provides both financial and commodity information (Whitridge and Nadeau, 2000). While its financial portion is a primary source for the System of national accounts, its commodity portion is very important to understand the dynamics of the industry. ASM is based on a sample but the end product is not a series of estimates but rather a full matrix of micro-data for all variables (financial and commodities) and all units in the population. It is challenging to have the imputation process get this complete set of micro-data through the modelling of historical values.

The ASM score function

10. The ASM collection process has used a score function for much longer than UES. In fact, the ASM methods were used as the basis for the development of the UES 2002 initiative. The interest in the commodity information motivated the development team to introduce a refined method making good use of the complete historical micro-data file.

11. Similarly to the UES scores described in section II, the goal of the ASM scores is to provide a numerical summary of a unit (Philips, 2003). Given the several variables of interest, the score is a composite measure which increases with the relative importance of the units.

12. For a given unit i, the ASM score is based on the value of its m commodities, say $C_{i1},..,C_{im}$, as follows:

$$S_i = \sum_{j=1}^{m} (C_{i+}/C_{++}) (C_{+j}/C_{++}) C_{ij}$$

where C_{i+} , C_{+j} , C_{++} are the sums of C_{ij} over j, i and ij respectively. From this equation, we observe that the score depends on the relative size (C_{i+}/C_{++}) of the unit and that each commodity is "weighted" by its relative importance (C_{+j}/C_{++}).

13. As collection proceeds, the scores are kept dynamic by using the best commodity value available. For clean records that have been received, the commodity values, C_{ij}, are the ones reported by the respondent. For other records, a modelled total shipment of commodity values $\hat{C}_{i+}$ is derived from historical information, with the help of estimated trends, which allows for the calculation of its estimated score. An additional feature of the ASM scores is the use of the number of follow-up attempts within the calculation of the non respondent's score. The principle is to reduce the score as follow-up failures occur and identify comparable substitutable units for priority follow-up.

14. With the dynamic approach, the follow-up process is completed when the sum of the clean commodity values surpasses the cell target coverage.

Evaluation of the process

15. Since the ASM scores have been used for several years now, evaluation studies have been completed. It has been shown that the effective implementation of the dynamic function was useful in prioritizing units for the follow-up process. During the first year of implementation, the ASM collection period was reduced and $250,000 was saved. While the response rates were reduced slightly when compared to previous years, the overall coverage rates remained stable.

IV. Relation with other processes

16. Both the UES and the ASM follow-up processes, with their score functions, were developed to reduce missing data problems, either caused by non-respondents or edit failures. When planning the survey, managers must consider its impact on other processes, monitor it across several survey occasions, and fine-tune the whole survey stream according to the performance of the score function.

17. On that matter, a first aspect is the link with the sample design. The stratification must be detailed enough to be efficient. Strata, however, especially strata with smaller units which are not prioritized by the score function, must get enough sampled units to face a lower response rate. Up to now, the UES design couldn't consider the new score function but the plan is to use the 2002 results to feedback the 2004 design. Outcoming response rates will be used to either increase stratum sample sizes or derive adequate expected CVs. On the other hand, the 2003 design targets the replacement of survey data with administrative data, with a modelling mechanism for variables not available from administrative sources. About 50% of sampled units will get such replaced values with modelling. Since the other 50% for which data will be collected will be used to derive the models, they become much more important then they used to be. For that reason, this is used as an input for the setting of the follow-up thresholds.

18. As for the estimation process, research studies are being planned to take into account the non-random response mechanism. To that end, post-stratification or calibration groups may be defined at a better level to reduce the risk of bias. Since the overall response rate is reduced by the use of a score function, the estimation process is also considering the calculation of the variance due to imputation. Here again, the plan is to monitor the impact on the estimation process and to adjust the score function strategy should problems occur.

19. The metadata, like variable definitions, their relations or their mapping with administrative sources, is the main source of information to allow the survey steps to help each other. In the case of UES, a comprehensive infrastructure was developed in 1997 and has evolved since then to store the survey data and metadata. Although the follow-up information does not automatically feed other processes, it is made available in the form of metadata to let the survey designers take it into account when adjusting various processes.

V. Conclusion

20. The use of a score function helps Statistics Canada to prioritize follow-up actions for larger units. This reduces the number of contacts, and then

reduces both the survey cost and the response burden. In the context of UES and ASM, follow-up information affects other processes through metadata. The sampling, collection and estimation processes can then react to it. On the other hand, the score function also reacts to other processes. The UES tax replacement initiative is an example showing how information travels from the sampling process to the collection, the follow-up, the imputation, and the estimation.

VI. References

Beelen, G., Royce, D. and Hardy, F. (1997). "Project to Improve Provincial Economic Statistics" presented at the Symposium 1997 of Statistics Canada, November 1997.

Latouche, M. and Berthelot, J.-M. (1990). "Use of a Score Function For Error Correction in Business Surveys at Statistics Canada" presented at the International Conference on Measurement Errors in Surveys, November 1990.

Philips, R. (2003). "The Theory and Application of a Score Function for Determining the Priority of Follow-up in the Annual Survey of Manufactures". Proceedings of the Survey Methods Section, Statistical Society of Canada, to appear.

Pursey, S. (2003). "Use of the Score Function to Optimize Data Collection Resources in the Unified Enterprise Survey". Proceedings of the Survey Methods Section, Statistical Society of Canada, to appear.

Statistics Canada (1998). North American Industrial Classification System, Canada 1997. Statistics Canada publication no. 15-501

Tourigny, J., Pursey, S., and Whitridge, P. (2001). "The Unified Enterprise Survey – its Approach to Quality" presented at the Symposium 2001 of Statistics Canada, October 2001.

Whitridge, P. and Nadeau, C. (2000). "The Redesign of the Annual Survey of Manufactures". Statistics Canada internal document, November 2000.