USE OF THE SCORE FUNCTION TO OPTIMIZE DATA COLLECTION RESOURCES IN THE UNIFIED ENTERPRISE

Stuart Pursey[1]

ABSTRACT

The Unified Enterprise Survey (UES) is Statistics Canada's largest annual business survey. It is designed to measure provincial and national economic variables across a wide range of industries. Non-response and edit failure during data collection leads to significant follow up costs. The UES Score Function identifies collection units that are most important for follow up due to their impact on estimates by province and by industry. In this way the Score Function ensures that limited follow up resources are used effectively to reach minimum levels of quality. In this article we examine the Score Function by describing its method, explaining why it improves the process of data collection, analyzing its impact on the quality of the final estimates, and discussing its relationship to the UES sample design.

KEY WORDS: Data collection; Non-response follow up, Score Function

RÉSUMÉ

L'Enquête unifiée auprès des entreprises (EUE) est la plus grande enquête économique annuelle de Statistique Canada. Elle est conçue pour mesurer des variables économiques provinciales et nationales pour une grande variété 'industries. Pendant la collecte des données, les coûts reliés au suivi auprès des non répondants et à la vérification des données sont considérables. La fonction de caractérisation de l'EUE identifie les unités de collecte qui sont les plus importantes pour faire l'objet d'un suivi étant donné leur impact sur les estimations par province et par industrie. La fonction de caractérisation nous assure que les ressources consacrées au suivi sont utilisées efficacement de façon à atteindre un niveau minimum de qualité. Dans cet article, nous examinons la fonction de caractérisation en décrivant sa méthodologie, en expliquant pourquoi elle améliore le processus de collection des données, en analysant son impact sur la qualité des estimations finales et en discutant de son implication par rapport au plan de sondage de l'EUE.

MOTS CLÉS : Collecte des données; fonction de caractérisation; suivi auprès des non répondants.

1. INTRODUCTION

1.1 The Unified Enterprise Survey and data collection

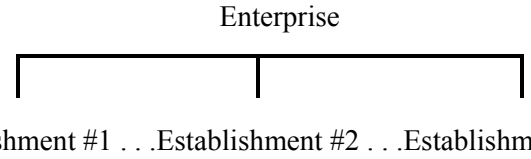
In business surveys at Statistics Canada there are significant costs to follow up non-response and edit failure during data collection. It is important to obtain data from all sampled units but some units are more important than others up due to their impact on the estimates. Statistics Canada has developed a Score Function for Reference Year 2002 of the Unified Enterprise Survey (UES) to identify the important units and to make priorities for follow up.

The UES is an annual business survey that unifies surveys from the Service industries, Manufacturing Wholesale and Retail, Transportation, Aquaculture, and Banking that were formerly designed and processed independently. There are about 20 surveys with a common frame, sample design, collection method, data processing, and estimation method. The questionnaires and data collection periods vary by industry. The process of collection is: design questionnaires for each industry, mail out questionnaires to respondents, and finally follow-up units that are non-response and edit failures.

Statistics Canada models the operating structure of a business using four hierarchies: enterprise, company, establishment, and location. The vast majority of businesses are simple in structure and thus all four levels of the business are one and the

[1] Stuart Pursey (stuart.pursey@statcan.ca), Statistics Canada, 11th Floor, R. H. Coats Building, Tunney's Pasture, Ottawa, Canada, K1A 0T6

same. Nonetheless the fewer businesses with complex operating structures dominate the economy. Only the enterprise and establishment levels are important to the UES.



The UES sampling unit (SU) is the group of establishments belonging to the same NAICS industry (North American Industry Classification System), same province, and same enterprise. About 65,000 business establishments are selected in the UES sample. They are linked to about 36,000 business enterprises. The selected sampling units generate about 30,000 collection units. Generally a collection unit is a questionnaire for a business that makes collection with that business efficient and appropriate based its structure and data availability. Thus many selected establishments may be associated with a single questionnaire and occasionally several questionnaires are associated with a single business.

The first UES was conducted for reference year 1997 as a pilot of seven surveys. Collection follow-up activities were conducted without a formal structure. This has two difficulties. First, the collection process, especially as it begins to reach its closure, is held hostage to the relentless demands from the UES survey mangers to make priorities for follow up. Second, an important measure of the quality of a survey is the amount of coverage it has managed to obtain through responding units as opposed to non-responding units (that will have been imputed). That is, if the response rate turned out to be 70% it would be preferable that the responses cover 95% of the industry's economy than it cover 35%. This is the basis behind the UES Score Function: give priority to collection units that have high impact by revenue.

Section 2 describes the methodology of the UES Score Function and Section 3 discusses the impact of the Score Function on data quality and on the cost of collection. This section 4 also provides some conclusions and ideas for an improved UES Score Function.

2. METHOD OF THE UES SCORE FUNCTION

2.1 Method of the UES Score Function: the initial scores

The UES sampling unit (SU) is the group of establishments belonging to the same NAICS industry (North American Industry Classification System), same province, and same enterprise. The UES score function attaches a numerical measure of importance to each UES sampling unit. R_i is revenue from the Business Register (the UES sampling frame) and the W_i is the sample weight from the UES sample design, of the i^{th} establishment in the sample.

The initial score is:

$$S_i = (W_i * R_i) / (\text{Total Weighted Sum of } R_i \text{ of all Units in Cell})$$

The higher the score the more important the unit and the more likely it will receive follow-up for non-response or collection edit failure. Since the score is a measure of the unit's weighted contribution to coverage, it is similar to the weighted response rate or weighted coverage rate. This is a useful measure because it reflects a collection goal of reaching a high coverage of the economy.

For each UES survey there is a target coverage rate. The sampling units within each cell – a cell contains the sampling units of a 5-digit NAICS and province combination – are listed in descending order, from highest score to lowest score. Then beginning at the top of the list, the units that accumulate their scores (i.e. contribution to coverage) up to the target coverage receive Priority=1 for Follow-Up and the remaining units receive Priority=0 for Do-Not-Follow-Up. This is the initial score.

Table 1 shows an example of the 20 sampled establishments of a cell in the UES sample with the goal of reaching 65% coverage. After consolidation of establishments to sampling units, there are 15 sampling units in this cell. The weighted sampling unit revenue for each sampling unit is calculated. For simple establishments this is just W_iR_i. Then the sampling units are sorted in descending order, then the accumulating percentage of revenue (or coverage of the economy)is calculated, and finally the Priority 1s and 0s are assigned so as to reach the minimum target coverage rate of 65%.

Table 1 Example of determining scores

Sorting of a Cell's Sampling Units by Scores and Determination of Priorities for a 65% Target Coverage				
Enterprise ID (sampling unit)	Number of establishments associated with the sampling unit	Weighted Sampling Unit Revenue	Accumulating percentage coverage	Priority
SU21	**3**	**14666**	**27%**	**1**
SU52	**1**	**11730**	**49%**	**1**
SU29	**4**	**4632**	**57%**	**1**
SU33	**1**	**3334**	**63%**	**1**
SU54	**1**	**2343**	**68%**	**1**
SU25	1	2327	72%	0
SU42	1	2251	76%	0
SU23	1	2106	80%	0
SU30	1	2036	84%	0
SU35	1	1989	87%	0
SU38	1	1825	91%	0
SU16	1	1562	94%	0
SU07	1	1187	96%	0
SU28	1	1182	98%	0
SU46	1	1120	100%	0

With the priorities set they are linked back to the establishments and then linked forward to the collection units associated with the establishments. The process is complex as there are many-to-many links between establishments and enterprises and collection units. The basic rule is that if at least one establishment of a collection unit is Priority 1 then the whole collection unit is deemed Priority 1. Further, sometimes an enterprise receives more than one collection unit (i.e. survey questionnaire). That is, often several questionnaires are sent to the same contact person in the business. In follow-up we should not say to a business or respondent "Ignore this questionnaire because it is "Priority 0" but please respond to the others because they are Priority 1." Therefore for consistency during collection, if any Questionnaire of an enterprise is Priority 1 (even if all the rest of the enterprise's Questionnaires are Priority 0) then all Questionnaires of that enterprise become Priority 1.

During collection the Priority 1s are followed up and thus incur a cost and a response burden. The Priority 0s are not followed up and thus incur no cost and no response burden due to follow up. In the example, five sampling units are initially Priority 1. If we get responses for all 5 then we reach 65% coverage of the total revenue of this cell and we can stop follow up, even if the others have not yet responded. But full response is not expected. Thus there is a second scoring processing – the dynamic process.

2.2 Method of the UES Score Function: the dynamic scores

During the collection process, there is feed back on the status of each questionnaire. There are four possibilities: Response (completed or received), Out of scope (industry misclassification, deaths, duplicates, etc.), Non-response (will not receive data) or "Still in Limbo" (not yet received but there is still hope).

Based on the current state of collection, once or twice a week, the scores are re-calculated. This is the dynamic part of the Score Function. The responding questionnaires contribute to meeting the threshold of the cell and thus the coverage target is reduced by their contribution. The out-of-scope questionnaires are removed entirely from the process – the scores are

re-calculated as though these questionnaires had never been part of the sample – this reduces the cell's revenue total. The non-responding questionnaires contribute nothing to the threshold. The "still-in-limbo" questionnaires remain eligible for dynamic scoring.

Before doing the re-calculation we ask: Have the responding questionnaires been sufficient to reach the threshold of a cell? If yes, the cell is closed for follow up and all sampling units become Priority 0. This allows follow up resources to be allocated to cells that have not yet reached their threshold. If not the scores are re-calculated based on the "Still in Limbo" questionnaires. If there were at least some responding questionnaires or some out-of-scope questionnaires from the previous week, there will be a reduced coverage target and the revised cell total. A "Still-in-limbo" Priority 1 questionnaire may fall back to Priority 0 if enough of the previous week's Priority 0 questionnaires have responded (despite no follow up). But more likely a Priority 0 will be promoted to a Priority 1 to substitute for Priority 1 questionnaires that are confirmed as non-response. The dynamic process continues week by week. As cells close, collection resources become concentrated toward cells that have not yet made their coverage targets. Some cells must close unfortunately, there being no more "still-in-limbo" questionnaires available.

3. EVALUATION OF THE SCORE FUNCTION AND IDEAS FOR IMPROVEMENTS

This section provides an outline of the approaches used to understand the impact of the score function on data quality and on collection costs. The analysis is limited since the processing of the UES RY2002 is not yet complete at the time of preparing this note. Conclusions and ideas for future variants of a score function are noted.

There are two views of data quality. The strong move toward the high impact units leads to a response bias. But is the bias minimal or significant and will estimation adjustments be necessary? (Simulations on the previous year's UES indicated that bias would be small.) Further, the score function as currently designed does not provide scores by stratum (take-all, large take-some, and small take-some) within the each cell. There is concern that the estimates for small, medium, and large size businesses will be weak. Thus there is some consideration of providing scores at the stratum level. On the other hand, the move toward the high impact units leads to a greater coverage of the economy from respondents. We avoid imputation for a higher proportion of the economy (measured by revenue), and from this perspective we improve data quality.

The target threshold of coverage for each survey needs to be considered more carefully. Currently they are set without objective analysis – generally the more skewed to distribution of revenue the higher the threshold (so as to get some follow up for the smaller units). More effective thresholds should be determined based on attempting to answer: how much quality is essential (regardless of cost) and how much can we afford to pay?

Although the analysis is preliminary it does appear that overall costs have been reduced slightly. It is clear that costs have moved toward the Priority 1 collection units – thus operationally the score function has worked as advertised! This of course is at the expense of the smaller collection units. It does appear that the response rate of the smaller units is less that than the previous year's UES, although perhaps not significantly.

There are some approaches that may be worth considering. Why "discard" some of the sample (from the view of not doing follow up)? Should the sample design be revised so that it generates a smaller sample size – so that all collection units can be followed up? Currently the score is based on the revenue of a business. This seems appropriate for the follow up of non response but for received units with edit failures that require follow up, perhaps a score based on the severity or frequency of edit failures is better.

3.1 Conclusion

From the view of collection operations it is clear that the score function is providing a useful day to day operational plan for collection staff to follow. Thus follow up activities proceed more smoothly and are no longer subject to a variety of ad hoc and sometimes sudden requests from the survey managers. There is an agreed upon stopping point (before an unobtainable 100% response rate is reached).