BUSINESS SURVEY DATA COLLECTION RESEARCH AT STATISTICS CANADA

Jeannine Claveau, Laurie Reedman and Xinye Yang[1]

ABSTRACT

In an effort to cost-effectively collect high quality data, Statistics Canada regularly reviews its collection practices. Over the past few years, Statistics Canada has conducted several analytical studies using paradata. The work done until now was mainly for social surveys. Statistics Canada has decided to extend the research to business surveys as well. The objective of the research is to identify opportunities for operational efficiency that could improve the quality of data collected. In this paper, we discuss some paradata analyses done at Statistics Canada for business surveys and provide some preliminary results.

KEY WORDS: Blaise, Business Surveys, Collection Practices, Paradata.

RÉSUMÉ

Afin de recueillir des données de grande qualité de manière rentable, Statistique Canada procède régulièrement à la révision de ses pratiques de collecte. Ces dernières années, Statistique Canada a mené plusieurs études analytiques en se servant de paradonnées. Le travail fait jusqu'à présent a porté principalement sur les enquêtes-sociales. Statistique Canada a décidé d'étendre la recherche aux enquêtes-entreprises. L'objectif de cette recherche est d'identifier les opportunités opérationnelles d'amélioration de la qualité des données recueillies. Cet article s'intéresse aux recherches sur les paradonnées faites à Statistique Canada pour les enquêtes-entreprises et présente quelques résultats préliminaires.

MOTS CLÉS : Blaise; enquête-entreprises; paradonnées; pratiques de collecte.

1. INTRODUCTION

In an effort to cost-effectively collect high quality data, Statistics Canada regularly reviews its collection practices. The use of paradata (i.e. data related to the survey process) is seen as a key component of this analysis. The work done until now was mainly for social and agriculture surveys. Statistics Canada has decided to extend the research to business surveys as well. The Business Survey Data Collection Research project was conceived. The first step in this research project was to assess the availability and quality of business survey paradata, and to determine if it holds the potential for identifying operational efficiencies in the data collection process itself. The second step was to determine if and how the paradata can be used to improve the quality of the collected data.

This paper discusses some paradata analyses done at Statistics Canada for business surveys and provides some preliminary results. Section 2 provides the reader with a brief description of the collection process. Section 3 gives examples of paradata analysis. Section 4 discusses future work planned. Conclusions are followed in Section 5.

2. COLLECTION PROCESS

Business survey collection consists of many steps and uses more than one collection mode. At Statistics Canada, many business surveys continue to use mail questionnaires for initial data gathering. Some respondents to business surveys prefer to give their data over the telephone or via an electronic data collection instrument. Electronic reporting allows

[1]Jeannine Claveau, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, jeannine.claveau@statcan.gc.ca;
Laurie Reedman, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, laurie.reedman@statcan.gc.ca;
Xinye Yang, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6, xinye.yang@statcan.gc.ca;

companies to extract information directly from their data systems or to complete a questionnaire online and transmit it electronically to Statistics Canada.

Before the mail-out of questionnaires, a pre-contact is often made for new enterprises selected in the sample to confirm their activity codes and contact information. The information for the new enterprises is provided by the Business Register and may not always be correct, therefore, a pre-contact call is required to make any corrections to this information when necessary. The pre-contact is also important because it determines which enterprises should not be part of the sample due to changes of ownership, business closures and amalgamations/mergers of companies or when their main business activity does not fall within scope of the survey they are selected to participate in.

When the questionnaires are mailed back digital images of the paper questionnaires are created and then the data are captured. Data capture and preliminary editing are performed simultaneously to ensure the validity of the data. Telephone follow-up is conducted to resolve edit problems with mailed back questionnaires and to collect data from respondents who have not returned the questionnaires after a pre-specified period.

3. PARADATA ANALYSIS

The Blaise system records the history of all calls made to enterprises during pre-contact, calls made for follow-up activities to collect data from respondents who have not returned the questionnaire, or calls made to resolve edit problems with mail-back questionnaires. Every time an operator accesses a collection unit, this system records a wealth of information. The data file generated by this system is called the Blaise Transaction History (BTH). Examples of paradata included in the BTH are collection unit identification, the date and the amount of time a case was open, the status of completion of an operational phase and relevant information about each call (e.g. the number of call attempts made, the result of the call, the appointment reason, etc.).

In order to assess the potential uses of paradata, some data analysis of paradata in BTH files of business surveys was performed. By analyzing the BTH files, we wish to learn more about various issues surrounding the data collection process. Examples of analysis are presented below for the 2008 Unified Enterprise Survey (UES). The UES is an annual economic survey that collects financial and characteristic data from Canadian businesses. It combines 60 surveys from different industries (Services, Distributive Trades, Manufacturing, Agriculture and Transportation). Most analyses presented here are related to the Annual Survey of Service Industries and the Annual Survey of Manufactures and Logging (ASML).

3.1 Pre-contact

For some business surveys, before the mail-out, a pre-contact is made for birth units to confirm their activity codes. However, survey managers question the cost-effectiveness of pre-contact. By using the BTH pre-contact file, we can carry out paradata analysis in an attempt to answer this and other such questions.

Pre-contact is conducted over the telephone. Confirmation can be obtained from a receptionist, an administrative assistant or any person within the organization who knows the details of the business. A maximum of three attempts is made for each pre-contact unit and then the case is finalized. If all the contact information (legal name, operating name, contact name, mailing address, phone number, fax number, e-mail address) and the main business information are confirmed then the case is finalized as “confirmed”. If all information except the main business activity is confirmed then the case is finalized as “not confirmed”. A case could also be finalized with an outcome status of “out-of-scope” or “never resolved”.

As shown in Figure 1, for the 2008 Annual Surveys of Service Industries, 27% of collection units in the sample did not require pre-contact. Almost half of the sample had their contact information and main business activity confirmed. Usually if the contact information is confirmed, the main business activity information is also confirmed. Few units (3%) had the contact information confirmed but not the business activity information (coded as North American Industry Classification System (NAICS) industry). Pre-contact also identified that 5% of the units in the sample were out-of-scope. No questionnaire was mailed out to the out-of-scope units.

As shown in Figure 2, the percentage of response is 5% higher for collection units where both contact information and business activity were confirmed during the pre-contact compared to those where no pre-contact was performed. Those

cases with only NAICS not confirmed and the unresolved cases during the pre-contact were more often non-response and out-of-scope units.

We cannot definitely say that pre-contacting units will increase the response rate. While it seems to increase a little bit for those units where all information was confirmed, we need to remember that questionnaires were not sent to the out-of-scope units (approximately one thousand) identified during pre-contact. However, many units are pre-contacted (73%) and it could be interesting to evaluate which ones really need to be pre-contacted.

Figure 1
Distribution of Pre-Contact Result

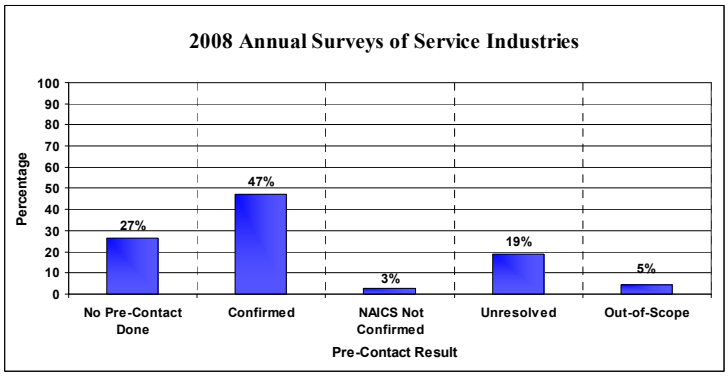
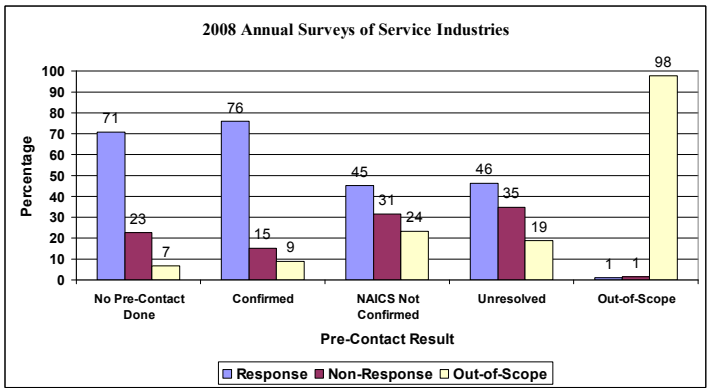


Figure 2
Distribution of Response Status by Pre-Contact Result



3.2 Distribution of response rates

To understand how the unweighted and weighted response rates change during the collection period, the distribution of cumulative unweighted and weighted response rates by month are presented. It is also interesting to see the distribution of return rate by month. We use this graph to determine if we can stop the collection process earlier and perform follow-up calls for non-response earlier in the process. These indicators are illustrated in Figure 3 for the 2008 Annual Survey of Manufactures and Logging.

The unweighted response rate equals the number of responding units over the number of collection units minus the number of out-of-scope units (discovered until that month). During collection, before discovering that a unit is out-of-scope, the unit is considered unresolved. If the unit at the end of collection is still unresolved, the unit will become a non-responding unit because of the survey deadline. We will never know whether that unit is an out-of-scope one. If the collection had stopped one month earlier, out-of-scope units that would have been discovered in the last month of collection would have been consider unresolved instead. For that reason, when we compute the response rate by month we just consider out-of-scope units discovered until that month.

The return rate indicates the percentage of questionnaires completed and returned. It is determined when the questionnaire is returned or if the unit is declared a respondent via another collection mode. The unweighted response rate and the return rate up to month *m* are given by:

$$\text{Cumulative Unweighted Rate}_{\text{m}} = \frac{\sum_{i=1}^{m} r_i}{n - \sum_{i=1}^{m} o_i}$$

Where r_i is the number of responding (or returned) units in month i; o_i is the number of out-of-scope units in month i; n is the total number of units; and m the month of the cumulative rate.

Response rate indicates the percentage of questionnaires completed, returned and verified. Response is determined once the questionnaire is returned and all edits (and subsequent follow-up edits) are performed or if the unit is declared a respondent via another collection mode. Business surveys have highly skewed populations, meaning a relatively small number of units can account for a large portion of the economic activity. Therefore, response rates ought to be calculated both on a weighted and unweighted basis. Essentially, the weighted response rate is the sum of the weighted revenue of

the respondents over the sum of the weighted revenue[1] of all units minus the sum of the weighted revenue of out-of-scope units. More explicitly, the weighted response rate up to month *m* is given by:

$$\text{Cumulative Weighted Response Rate}_{\text{m}} = \frac{\sum_{i=1}^{m}\sum_{k\in R_m} w_{ik}x_{ik}}{\sum_{k\in S} w_k x_k - \sum_{i=1}^{m}\sum_{k\in O_m} w_{ik}x_{ik}}$$

Where w_{ik} ,w_k are sampling weights and x_{ik} , x_k are revenues; R_m is the set of responding units at month *m*; O_m is the set out-of-scope units at month *m*; and *S* is the set of all of units in the sample.

As seen in Figure 3, the weighted response rates are higher than the unweighted response rates except in the the last months. This means that we got more responses from large contributing businesses (with larger weighted revenue) than from smaller ones in the last few months of collection. The return rates are, until August, higher than the weighted and unweighted response rates. That happens because after a questionnaire is received from a collection unit, the editing step has to be completed before the case is finalised. A collection unit is identified as returned before the final status is outputed. For ASML, it seems that the collection could stop no earlier than August since the return rate continues to increase until that month.

The distributions of cumulative unweighted return rate, unweighted and weighted response rates by number of call attempts made for collection units are presented in Figure 4. To create a graph of unweighted and weighted response rates by number of attempts, we need to compute the response rates for each subset of sample units: one subset with units that need 0 attempts to return their questionnaires, 1 attempt, 2 attempts….The unweighted response rate and the return rate up to the number of attempts, *j*, are given by:

$$\text{Cumulative Unweighted Rate}_{\text{j}} = \frac{\sum_{i=1}^{j} r_i}{n-o}$$

Where r_i is the number of responding (or returned) units with *i=0,1,2,...j* attempts; *o* is the number of out-of-scope units in the sample and *n* is the total number of units. Note that since we compute response rate by the number of call attempts made during the entire collection period, we subtract all out-of-scope units discovered during collection. The weighted response rate up to the number of attempts, *j*, is given by:

$$\text{Cumulative Weighted Response Rate}_{\text{j}} = \frac{\sum_{i=1}^{j}\sum_{k\in R_j} w_{ik}x_k}{\sum_{k\in S} w_k x_k - \sum_{k\in O} w_k x_k}$$

Where w_{ik} , w_k are sampling weights and x_{ik}, $x_k$are revenues; R_j is the set of responding units with *j* attempts; *O* is the set out-of-scope units; and *S* is the set of all of units in the sample.

As we can see in Figure 4, it seems that making six or more attempts on a case does not increase the response or return rates significantly. This is because few cases have six or more attempts, not because the businesses refuse to respond or are absent. Currently, there is a limit on the number of call attempts during collection. After five contact attempts, where either contact is made or an answering machine message is left, the case is resolved as final non-response. However, if another contact person is reached at any of the five contact attempts, if the respondent contacts Statistics Canada or if the unit is considered influential to the survey, the number of attempts could be greater.

---

[1] The weighted revenue used is the revenue available on the Business Register of Statistics Canada

Figure 3
Distributions of Response Rates by Month

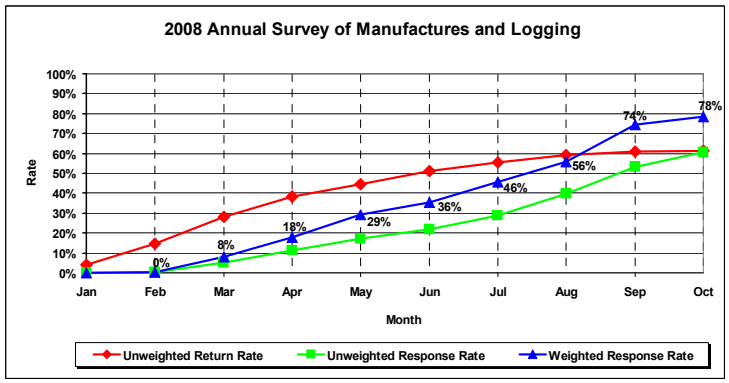
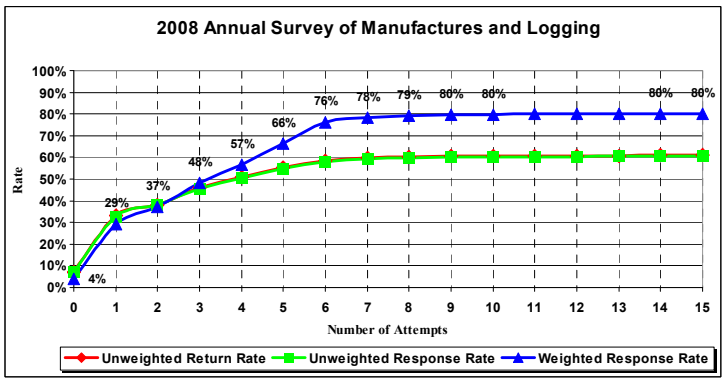


Figure 4
Distributions of Response Rates by Attempts



3.3 Distribution of Time spent

We have also examined the distribution of time spent by collection step and groups of surveys. Figure 5 presents the total number of hours spent for each collection step (pre-contact, non-response follow-up and failed edit follow-up) for the 2008 UES. We remark that ASML spent more time in failed edit follow-up than in non-response follow-up but for the other groups of surveys it is the reverse. Also, Annual Surveys of Service Industries (called UES Services in Figures 5 and 6) is the group of surveys that spent the most amount of time in pre-contact.

Figure 6 shows the average time spent per unit by collection step and group of surveys. For pre-contact and failed edit follow-up, only the collection units involved in those steps are considered. The average time spent per unit varies from 4 to 13 minutes for pre-contact, from 18 to 32 minutes for non-response follow-up and from 18 to 60 minutes for failed edit follow-up. This confirms that huge amounts of time are required for failed edit follow-up. There is a need to improve collection procedures to decrease the number of cases sent for failed edit follow-up.

Figure 5
Distribution of Time Spent by Surveys

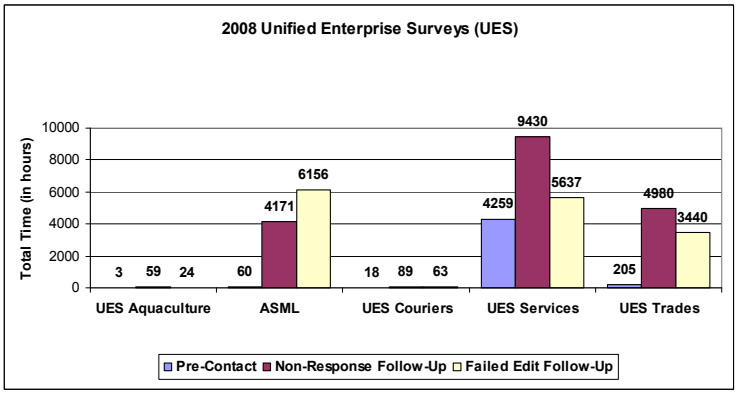
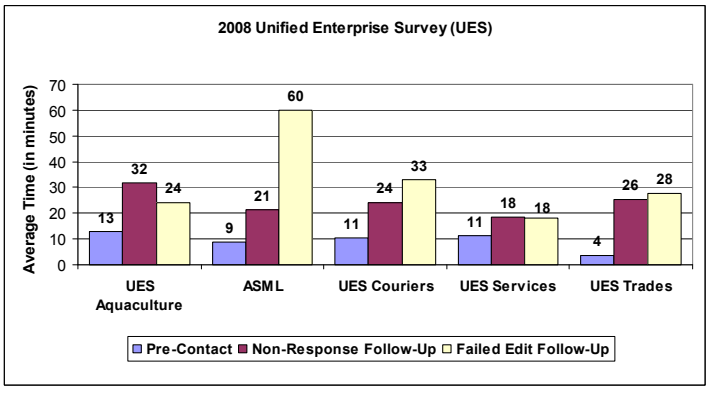


Figure 6
Average Time Spent per Unit



4. FUTURE WORK

We want to continue to analyse the BTH files to gain a better understanding of the collection process. We believe that more investigation is needed to find ways to reduce follow-up calls for failed edits. A review of collection follow-up procedures is necessary to identify how we can modify them to reduce the cost of collection without affecting the quality of the data collected. If it is not possible to reduce the time spent per unit for failed edit follow-up calls, we will need to examine if some units with failed edits could be sent directly to processing (i.e. imputation and estimation) instead of being sent to the follow-up call process. We need to evaluate if we can reduce the cost of collection by making more efficient use of the pre-contact process and identifying which units really require a pre-contact. We also want to determine if implementing a cap on calls would have an effect on estimates.

5. CONCLUSION

Paradata is an amazing source of information that helps us to evaluate collection processes. The study of Blaise files permits us to better understand the collection process for business surveys. Particularly, to understand how time is divided between the collection process steps and identify which steps are the most time consuming. Currently, Statistics Canada is undertaking a general restructuring of its business statistics programs. One of its goals is to let electronic data collection become the principal mode of collection for business surveys. For that reason, we aim to build an experimental design for electronic collection to compare different non-response follow-up methods. Analysis of paradata would permit us to compare different options of collection processes for electronic questionnaires.

For all theses reasons, research in paradata is important to help us monitor data collection changes in business surveys and to improve the quality of data collected. We have only scratched the surface. Every result leads us to more questions.

ACKNOWLEDGEMENTS

The authors would like to thank the following people for their comments and suggestions, as their contributions improved greatly the final version of this paper: Serge Godbout and Tracy Tabuchi. The views expressed in the paper are those of the authors and do not necessarily reflect the official position of Statistics Canada. All remaining errors are those of the authors.

REFERENCES

Brodeur, M., Koumanakos, P., Leduc, J., Rancourt, É. and Wilson, K. (2006), "The Integrated Approach to Economic Surveys in Canada", Statistics Canada, Catalogue No. 68-514-XIE

Evra, R.C. and DeBlois, S. (2007), "Using Paradata to Monitor and Improve the Collection Process in Annual Business Surveys". *Proceedings of the 2007 International Conference on Establishment Surveys,* Montreal, Quebec.

Laflamme, F. (2008). "Data Collection Research using Paradata at Statistics Canada". *Proceedings from the 2008 International Symposium on Methodological Issues,* Statistics Canada.

Statistics Netherlands, Blaise system, URL http://www.blaise.com

Beaucage, Y. and Yung, W. (2008), "Frame Improvements to Statistics Canada Business Register". *Proceedings of the 2007 Joint Statistical Meetings,* Salt Lake City, USA.