Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2004: Innovative Methods for Surveying Difficult-to-reach Populations

2004

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2004
Innovative methods for surveying difficult-to-reach populations

CHALLENGES IN THE SURVEY OF ELECTRONIC COMMERCE AND TECHNOLOGY

Marie-Claude Duval and Edith Latendresse[1]

ABSTRACT

The Survey of Electronic Commerce and Technology (SECT) collects information about the use of information and communications technologies such as the Internet, e-mail, intranets and Web sites and about the use or non-use of the Internet for electronic commerce. It is an annual survey of a sample of 20,000 Canadian businesses from every industrial sector of the economy. Since the survey was launched in 1999, data needs have continued to grow, as has expertise concerning computer technology and electronic commerce. This led to an exploration of the survey methodology used, which was based on a standard economic survey model. To date, there have been many studies of various aspects of the survey, including coverage of the survey population, the sample design, collection, and edit and imputation. Further studies will also be needed on rare-event estimation. This article describes the survey, its main problems and challenges, study findings, and past and future actions.

KEYWORDS: Dichotomous Variables; Economic Survey; Rare Event; Sample Design.

1. INTRODUCTION

The Survey of Electronic Commerce and Technology is an annual survey that collects information about the use of technology and electronic commerce from Canadian businesses in every industrial sector. The main variables of interest are Internet use, Web site ownership, and purchase and/or sale of goods and services over the Internet. Businesses that do not buy or sell over the Internet are asked about the reasons for not doing so. There is also a question on the technological improvements businesses have made over the years. The survey was launched in 1999.

Unlike ordinary economic surveys, which usually measure numerical economic variables such as the components of income and expenditure, the SECT measures qualitative variables, specifically dichotomous variables. We want to know whether a characteristic is present or not: for example, “Do you use the Internet (yes or no)?”. Over the years, we have found that the current sample design has certain weaknesses in view of the nature of the survey, the type of variables being measured, and the analytical requirements.

The survey also measures just one numeric variable, but it is an important one: total sales made through electronic commerce. Since the survey’s inception in 1999, we have opted for a sample design that optimized the measurement of total sales for the sample frame available to us. We have found, however, that this variable is unstable over time despite the sample design used. First, we are dealing with a rare event. Few businesses engage in electronic commerce, and some industries have no market for Internet sales. Second, we have a variable whose concepts do not always seem to be well understood by respondents or for which the information is unavailable. For example, we have noticed that some businesses that sell on the Internet know their total sales but cannot say specifically what portion of those sales were made via the Internet.

This article describes the standard economic survey methodology used for this survey, the reassessment of needs over the years, the weaknesses identified, the proposed methodology for meeting the needs, and future challenges.

[1] Marie-Claude Duval, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney’s Pasture, Ottawa, Ontario, Canada, K1A 0T6 (marie-claude.duval@statcan.ca), Edith Latendresse, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats Building, Tunney’s Pasture, Ottawa, Canada, K1A 0T6 (edith.latendresse@statcan.ca)

2. THE SURVEY'S METHODOLOGY

The survey has an optimum sample design for the measurement of business income variables, a design commonly used in Statistics Canada's economic surveys. The goal is essentially to minimize the sample size for a desired level of precision in the estimates. In this section, we will describe the population covered by the survey, sample stratification and allocation, sample selection, edit and imputation, and estimation.

The population is drawn from Statistics Canada's Business Register (BR) (Cuthill, 1997), which includes nearly 2 million enterprises. The first step is to remove the smallest units from the population. Such units have little or no impact on economic activity. For each domain of interest, exclusion limits are set so as to exclude no more than 5% of the domain's total income. The income used is the gross business income available for all units in the BR. This yields a survey population of about 650,000 units; the target population. The domains of interest are determined according to the North American Industry Classification System (NAICS) at different levels of detail for each industrial sector.

The survey population is then stratified by size within the NAICS domains using the Lavallée-Hidiroglou algorithm (1988). The algorithm groups the units into take-some strata and a take-all stratum and allocates the sample across the strata in such a way as to minimize the sample size for a given coefficient of variation (CV). We specify two take-some strata and one take-all stratum for each domain of interest. To calculate the CV, we use as an auxiliary variable, the gross business income available for all units in the BR. The available budget of about 20,000 units for the survey is preset. For that reason, our approach is to test various CVs and select the one that produces the permitted sample size. We also want a larger sample in sectors that are more likely to engage in electronic commerce in order to improve the quality of the estimates of total Internet sales. To that end, in 2003, we used a target CV of only 2% in the manufacturing, wholesale, retail and transportation and warehousing sectors, where electronic commerce was more likely, and a target CV of 5% for all other industrial sectors. Finally, we increase the sample size in some strata in order to ensure a minimum sampling fraction of 1% and no fewer than five sample units per stratum.

All units in the take-all strata are sampled. In addition, units that are expected to show very high Internet sales are identified and forced into the sample with a sampling weight of 1. In the take-some strata, a random sample is selected; at the same time, overlap with the previous year's sample is maximized. The method proposed by Kish and Scott (1971) is used; in 2003, this produced a total overlap of 65% with the previous sample. Each year, a new population is generated and a new sample is selected.

A questionnaire is then mailed out to each unit in the sample. During data capture, each questionnaire has to pass a number of edits, such as consistency and historical comparison edits. Follow-up letters and faxes are sent to businesses that have not responded. Some telephone follow-up is also done to improve the response rate, obtain unreported data, or validate or correct extreme or inconsistent data provided by respondents. Some units are given priority in the follow-up process; this is based on stratum response rate, industrial sector, business size, the importance of the missing variables, and the type of inconsistencies in the questionnaire. The response rate is about 70% each year.

Following collection, records that are still incomplete or inconsistent are imputed. On average, only 6% of the fields are imputed because they are missing, and less than 0.5% because they are inconsistent. Various imputation methods are used, including deterministic imputation, historical imputation and donor imputation. Only partial questionnaires are imputed. No imputation is done for total non-response. Instead, the respondent units are reweighted. Outliers in total Internet sales data are detected with a variant of the sigma gap method (Nobrega, 1998).

Statistics Canada's Generalized Estimation System (GES) is used to estimate all variables. For information about the GES, see Estevao (2005). Two main types of estimates are produced.

(1) For **dichotomous variables (*C*)**, a quotient is used.

$$\hat{C}_d = \frac{\sum_s w_i z_i c_i(d)}{\sum_s w_i z_i} \quad \text{where } c_i(d) = \begin{cases} 1 \text{ if } i \in d \text{ and } i \text{ has the characteristic} \\ 0 \text{ otherwise} \end{cases}$$

(2) For **numerical variables (*Y*)**, a total estimator is used.

$$\hat{Y}_d = \sum_s w_i y_i(d) \quad \text{where } y_i(d) = \begin{cases} y_i \text{ if } i \in d \\ 0 \text{ otherwise} \end{cases}$$

The variable w_i represents the adjusted sampling weight of unit i after reweighting for non-response. The variable z_i is an auxiliary variable: income, number of employees, or some other variable, depending on the variable being estimated. Estimates are produced with and without this auxiliary variable. The auxiliary variable generates what we call "economically weighted" estimates by increasing the weight of large units. For example, in 2003, the proportion of businesses that had a Web site was estimated at 34% without economic weighting and 85% when economically weighted by income. This means that the 34% of businesses with Web sites account for 85% of total income in Canada.

The GES also computes a variance and a CV for all estimated variables.

3. NEEDS REASSESSMENT

Over the years, data needs have grown, as has expertise on the subject.

First, comparing estimates over time is very important to users. They want to measure not only developments and changes in technology use but also company behaviour in technology use over time. For example, by how much did the percentage of businesses with Web sites increase between 2001 and 2003? In 2003, what was the behaviour in 2003 of businesses that made little use of technology in 2000?

Second, over the last two years, there has been interest in technology use by business size (small, medium and large) where size is based on number of employees. We currently use the number of employees reported in the questionnaire and generate a domain estimate.

Business size is defined as follows (except in manufacturing):

Small:	Fewer than 20 employees
Medium:	20-99 employees
Large:	100 employees or more

For the manufacturing sector, the following definition is used:

Small:	Fewer than 20 employees
Medium:	20-499 employees
Large:	500 employees or more

Analysts are interested in the use and the evolution of technology for the small, medium and large businesses based on these definitions.

Third, one of the most important variables of interest in the survey is total Internet sales. However, Internet sales is a rare event, and we have little auxiliary information available to help us target businesses selling over the Internet. In 2003, it was estimated that only 7% of private enterprises were making sales over the Internet. In many cases, the estimates produced up to now have been of lower quality, and we would like to find ways of improving the quality for this variable.

4. NEEDS REASSESSMENT VS METHODOLOGY USED

In section 2, we noted that we were generating a new population and selecting a new sample each year, while excluding the smallest units, which accounted for less than 5% of total income for each domain of interest. We also mentioned that stratification was based on an algorithm that determined the strata limits for each domain of interest and the sample allocation for each stratum so as to minimize the required sample size for a given CV. This is an optimum method when we are measuring totals correlated to the auxiliary variable. However, when we are measuring dichotomous variables and the probability of having a characteristic varies by business size, this method can sometimes produce inconclusive results.

First, the target population can vary considerably over time depending on changes in the units' incomes from year to year. For example, between 2001 and 2002, the overall target population grew by 2%, which is plausible. However, in one particular industry, the target population grew by 41%. The increase is attributable to a downward adjustment in one unit's income between 2001 and 2002. Thus, even though the total population was similar from one year to the next, the target population changed dramatically, since in order to cover 95% of the population on an income basis, we had to add many companies. We are covering the same population in income terms, but a much larger population in terms of number of businesses. One problem is that when businesses are added in order to reach the 95% threshold, they are added in decreasing order of size, so that each unit added is smaller than the one before. Since the smallest companies make less use of technology, the concern is that increasing the number of businesses in the target population may result in an estimate that is smaller than the previous year's estimate. Such a decline would be due to a change in the target population and not to a real evolution in the sector's technology use; this is not a desirable conclusion.

Second, the current sample design is not representative by size based on number of employees. The Lavallée-Hidiroglou algorithm (1988) sets arbitrary stratum limits for each domain of interest and allocates the sample for each stratum. The only constraint on it is that the sample size must be minimized for a target CV based on an auxiliary variable (in our case, gross business income). Hence the stratum limits and the sample allocation will vary depending on the domain, income and target population. There is no guarantee that the sample will be representative by size based on number of employees. A representativeness problem was detected in one industry by comparing the estimates for 2002 and 2003. The estimated percentage of businesses with a Web site declined from 68% to 61%. In technology, decreases in estimates are somewhat surprising since we would expect technology use to grow over time. Consequently, we looked at results for each business size (Table 1).

Table 1: Estimated percentage of businesses that have a Web site, and distribution of the estimated population by business size for a particular industry

Size	% of businesses with a Web site		% of the estimated population	
	2002	2003	2002	2003
Small	60%	54%	72%	83%
Medium-sized	87%	97%	23%	13%
Large	84%	98%	4%	4%

The table clearly shows the change in the distribution of the estimated population by business size between the two years. The estimated percentage of small businesses was 83% in 2003 and 72% in 2002. The estimated percentage of businesses with a Web site was 54% to 60% for small businesses and more than 85% for other businesses. With so many small businesses in the population, it is not surprising that the 2003 estimate was smaller than the 2002 estimate. Is this because the sample was non-representative in one of those years? Is it because the population changed as a result of changes in the income of certain units? We considered that our sample design might not be very robust for what we wanted to estimate. After some analysis, however, we realized that such cases are not very common, as the number of employees is partially correlated with income.

While the current design is flawed for certain types of analysis, it is an optimum design for computing economically weighted estimates. The economically weighted estimator uses income to weight the characteristic's value (0 or 1). Since the target population and the sample design are based on income and cover the same population in terms of income (95%) from year to year, the estimators are consistent.

Finally, steps were taken to address the faults in the current design. First, an outlier detection technique (Nobrega, 1998) is applied to the population's income over time. As a result, significant changes in income are checked and, if necessary, corrected. This stabilizes the target population over time. With regard to distribution by size, we make sure that the response rate is uniform across the strata. In contrast to economic surveys, where the goal is good coverage of respondent units in terms of income or some other auxiliary variable, the aim in this survey is to have a good response rate for both low-income and high-income businesses (Lohr, 1999). Hence we try to improve representativeness by size since there is some degree of correlation between number of employees and income.

Another of the sample design's limitations lies in the estimation of Internet sales. As noted in the previous section, Internet sales is a rare event, and we have little auxiliary information about the sample frame to help us target businesses with Internet sales. The estimates for this variable are not very precise. At the NAICS2 level, which consists of 22 sectors in Canada, more than half of the estimates have a CV in excess of 25%. In 2001, nearly a third of the sectors had a CV of more than 50%. There was a slight improvement in 2002 and 2003, as CVs exceeding 50% were present in only one-tenth of the 22 estimates produced.

Over the years, some steps have been taken to slightly improve the estimation of Internet sales. For example, we oversample in sectors that are more likely to engage in electronic commerce, such as manufacturing, wholesale, retail, and transportation and warehousing. We also ensure that the largest known contributors to Internet sales and the largest enterprises in Canada are included in the sample. We assume that the largest firms are more likely to engage in electronic commerce and to have significant Internet sales. In addition, we apply edit rules that compare current data with historical data so that we can perform appropriate follow-up and correct or validate the information supplied by respondents.

5. PROPOSED METHODOLOGY

To better meet the survey's needs and users' requirements, we propose using a new sample design. This design should help improve two aspects of the survey: comparability across time and representativeness by size.

First, we wanted to stabilize the target population in time, that is, make it less sensitive to fluctuations and/or corrections in the income data taken from the Business Register. To that end, we proposed setting a fixed exclusion limit instead of 95% income coverage. We considered various options that would keep the population comparable to the population used in previous years, with various income limits for each industry. The choice came down to two income limits: $250,000 and $100,000. For each industrial sector, the limit that provides 95% income coverage is selected. For example, for the manufacturing sector, when we remove all units with an income of less than $250,000 from the target population, the population we are left with easily covers 95% of the sector's total income. In the accommodation and food services sector, on the other hand, a limit of $100,000 is required to achieve 95% coverage. To stabilize the target population, we proposed retaining the limits for the next two or three years. This method keeps the population size similar to what it was the previous year (about 650,000 units).

Another analytic requirement is estimates by business size (small, medium and large). We want to measure technology use by small, medium and large businesses and its evolution over time. We therefore wanted to change the sampling method to better reflect the definitions of small, medium and large businesses used by survey analysts.

First, Statistics Canada's Business Register contains data on the number of employees for all businesses. We checked the correlation between this variable and the number of employees measured in the survey. Since the correlation was very strong within classes of interest, we used the number of employees in the Business Register in developing the sample design. We tested a number of scenarios, with the constraint that large businesses must be included in the sample, since they are more likely to have significant Internet sales. The scenario we selected consists of four strata within the domains of interest.

Stratum 1: businesses with fewer than 20 employees
Stratum 2: businesses with 20-99 employees
Stratum 3: businesses with 100-499 employees
Stratum 4: businesses with 500 employees or more

Strata 1, 2 and 3 are take-some strata, and stratum 4 is a take-all stratum. On the basis of historical estimates, stratum 4 alone would account for 45% of all Internet sales. For strata 1, 2 and 3, we allocated the sample size using the Neyman allocation method and the income variable. The advantage of using income data is that we obtain a larger sample in the strata that are less homogeneous in terms of income, and are therefore able to generate better economically weighted estimates.

In allocating the sample, we selected more units in sectors that were more likely to have Internet sales. After testing a number of scenarios, we decided to select half the sample from sectors more likely to engage in electronic commerce, even though those sectors make up only a third of the total population.

In addition, as in the old sample design, we selected the sample with a view to maximizing the overlap with the previous year's sample. The method proposed by Kish and Scott (1971) was also used, providing an overlap of nearly 60% with the previous year's sample.

6. NEEDS REASSESSMENT VS PROPOSED METHODOLOGY

In this section, we summarize how implementing the proposed methodology should satisfy the survey requirements defined in section 3.

First, with respect to the comparability of estimates over time, the new design, with its much more stable population and fixed stratification limits for all industries, will clearly improve the analysis of the qualitative variables measured in the survey.

The same is true for analysis of the estimates by business size based on number of employees, since stratification was based on number of employees with the same size definitions.

As for the estimates of total Internet sales, we do not believe that this variable will improve under the new design. We are using the same frame (Business Register), and so have the same limitations as before (no information about Internet sales except for industrial sector and income). On the other hand, we do not expect the estimates' quality to be worse, and we are considering other ways of improving this variable in the future (see the next section).

Finally, even though the new design is not an optimum design for economically weighted estimates, we believe that it is quite robust. The largest units are in self-representing strata, and the sample size was allocated by stratum on the basis of business income variability in the strata (Neyman allocation).

For next year, we have decided to use a sample under the new design in parallel with a sample under the old design. This will enable us to assess the new design's impact on the estimates and to adjust historical data if necessary.

7. FUTURE CHALLENGES

One of the biggest challenges for the future will be to improve the quality of the total Internet sales estimates. As noted in earlier, Internet sales is a rare event, and we have little auxiliary information to help us target businesses with Internet sales. The estimates for this variable are not very precise. In addition, following historical comparisons and follow-up, we have noticed that this variable has not always been well understood by respondents. In both cases, steps were taken, such as including businesses with substantial Internet sales in the sample, oversampling in sectors more likely to engage in electronic commerce, performing respondent follow-up, and adding clarifications to the questionnaire.

One of the challenges for this variable is the use of a new frame or a complementary frame containing information about businesses with Internet sales. We wondered if such businesses could be identified by running a search query on the Internet. Clearly, if businesses are making sales over the Internet, they must be present on the Internet. The question is whether a database extracted from the Internet would be of good quality and capable of covering our population of interest. Data mining techniques, specifically text mining, might be considered (Chauchat and Morin, 2004). Text mining uses matching techniques to find associations between words or between documents. In our case, the idea would be to prepare a common vocabulary used on the Web sites of companies with Internet sales (for example, using survey respondents that reported Internet sales) and find the common vocabulary through analysis. Then the common words can be used as search terms on the Internet.

Another challenge is to assess how well respondents understand this variable. Do they understand and report exactly what we want to measure (inclusions/exclusions)? Do they have the information on hand? Even though focus groups were used to test the questionnaire prior to the survey's launch, it would be a good idea to hold more focus groups with businesses It is important to keep in mind that since 1999, expertise has grown and many changes have been made in the survey (new sample unit, new questionnaire, feedback obtained from respondent follow-up, etc.). The discussions could be on the subject of electronic commerce but also on the use of technology in general by businesses (for example, how do businesses launch a Web site and how do they maintain it? At what level in the company are decisions made whether to use certain technologies or not?).

8. CONCLUSION

Economic surveys are always evolving. Needs change and expand; expertise grows; the number of available information sources increases; and businesses evolve. As a result, surveys have to change over time to keep pace with users' needs.

The Survey of Electronic Commerce and Technology is a fairly recent survey concerning a field that is constantly evolving. The information has to be released quickly or it will be out of date. Requirements for more detailed data are growing, and historical comparison is critical in order to properly measure changes in technology use over time. We believe that a new sample design was necessary in order to meet those needs. The new design will undoubtedly improve comparability over time and analysis by business size.

One of the remaining challenges is to improve total Internet sales, a key variable in the survey. Our next step will be to measure that rare event more effectively. The Internet contains some of the information we need. It's up to us to assess its potential!

REFERENCES

Chauchat, J.-H. and Morin, A. (2004), "Data Mining and Text Minimg", Workshop presented during the *XXI International Methodology Symposium: Innovative methods for surveying difficult-to-reach populations*, Gatineau, Canada.

Cuthill, I. (1997), "Le registre des entreprises de Statistique Canada", unpublished report, Ottawa, Canada, Statistics Canada.

Estevao, V. M. (2005), "GES V4.3 User Guide", unpublished report, Ottawa, Canada, Statistics Canada.

Kish, L. and Scott, A. (1971), "Retaining Units after Changing Strata and Probabilities", *Journal of the American Statistical Association*, 66, pp. 461-470

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

Lohr., S. L. (1999), "Non-response", *Sampling Design and Analysis*, Press, pp. 255-282.

Nobrega, K. (1998), "Outlier Detection in Asymmetric Samples: A Comparison of an Inter-quartile Range Method and a Variation of a Sigma Gap Method", *1998 Proceedings of the Survey Methods Section*, Statistical Society of Canada, pp. 137-141.