

## CHALLENGES IN MEASURING ELECTRONIC COMMERCE

Patricia Whitridge<sup>1</sup>

### ABSTRACT

Interest in electronic commerce has been increasing over the last few years. Statistics Canada undertook a survey to measure the amount of sales taking place over the Internet for reference year 1999. It was decided to take advantage of the sample design of an existing economy wide establishment survey. A questionnaire was developed with questions about the use of technology in businesses. Information was collected on the availability of computers, the use of e-mail and access to the Internet. Establishments were also asked if they had a Web page, if it could be used for sales, and the amount of those sales. The data was released in the summer of 2000. Now, many issues are being discussed concerning the most effective methods to use to estimate the total sales over the Internet. For example, should we be asking questions about Web pages to establishments, or rather to the head offices? Is a general survey of businesses the best method? Could we draw a sample of businesses and then research the Web pages ourselves?

The presentation will begin with some background information about the 1999 survey, including its challenges. Then, current plans for the 2000 survey, which is underway, will be discussed. Some issues that have already surfaced will be identified, and different scenarios for the longer-term future will be examined. The presentation will conclude with some general observations from Statistics Canada's experience measuring electronic commerce.

KEY WORDS: Internet sales; Weight share method.

### RÉSUMÉ

L'intérêt pour le commerce électronique avait augmenté au cours des dernières années. Statistique Canada a entrepris une étude pour mesurer la quantité de ventes ayant lieu sur Internet pendant l'année 1999. On a décidé de profiter du plan d'échantillonnage d'une enquête déjà existante auprès des établissements de tous les secteurs économiques. Un questionnaire a été développé avec des questions au sujet de l'utilisation de la technologie dans les entreprises. L'information a été recueillie sur la disponibilité des ordinateurs, sur l'utilisation du courriel et sur l'accès à l'Internet. On a également demandé aux établissements s'ils avaient une page Web, si elle pouvait être utilisée pour des ventes, et la quantité de ces ventes. Les données ont été disponibles à l'été 2000. Maintenant, plusieurs possibilités sont discutées au sujet des méthodes les plus pertinentes à utiliser pour estimer toutes les ventes réalisées par Internet. Par exemple, devrions-nous poser des questions sur les pages Web aux établissements, ou plutôt aux sièges sociaux? Une étude générale des entreprises est-elle la meilleure méthode? Pourrions-nous tirer un échantillon d'entreprises et rechercher alors nous-même les pages Web?

La présentation commencera par de l'information de base au sujet de l'étude de 1999, incluant ses défis. Puis, les plans actuels pour l'étude de 2000, qui est en cours, seront discutés. Quelques possibilités qui se sont déjà présentées seront identifiées, et différents scénarios d'avenir à plus long terme seront examinés. La présentation se conclura par quelques observations générales sur l'expérience de Statistique Canada dans la mesure du commerce électronique.

MOTS CLÉS : La méthode de partage de poids; ventes sur Internet.

### 1. INTRODUCTION

Often cited as a key contributor to the sustained period of economic growth through the 1990's has been the widespread use of information and communications technologies (ICTs), the Internet in particular. In addition to providing a new channel for businesses to sell their products, it also helps businesses reduce costs

by expanding the pool of potential suppliers (i.e. electronic marketplaces) and by improving the performance of their supply chains. In response to requirements by policy departments in the federal government, Statistics Canada implemented the first economy-wide survey of electronic commerce for reference year 1999. The survey used an ambitious questionnaire covering not only purchases and sales

<sup>1</sup> Patricia Whitridge, Business Survey Methods Division, 11-F R.H. Coats Building, Statistics Canada, Ottawa, Ontario, K1A-0T6, whitpat@statcan.ca.

over the Internet, but also questions on computer and Internet use, the presence of Web sites and maintenance costs. Since then, an annual survey program has been implemented. This paper describes the 1999 survey, the way in which it was developed, the particular challenges encountered, and the solutions implemented. A discussion of the 2000 survey follows, highlighting the changes that were implemented. In conclusion, some general comments are offered.

## **2. 1999 SURVEY OF INFORMATION AND COMMUNICATION TECHNOLOGIES AND ELECTRONIC COMMERCE**

### **2.1 Sample Design**

For the reference year 1999, we decided to make use of an existing economy-wide survey: the Capital Expenditures Survey (CAPEX). CAPEX is an establishment-based economy-wide survey with a sample of 27,000 establishments (Statistics Canada, 2000). Businesses in both public and private sectors are surveyed using a sample drawn principally from Statistics Canada's Business Register (private sector) and from Public Institution Division (public sector).

The Capital Expenditures Survey uses stratified simple random sampling with one take-all and one take-some stratum in each province by industry grouping. For sample allocation, there are two steps. A first algorithm takes the overall CV into cell (province by industry) CVs. Then the Hidiroglou (1986) method is used to draw the take-all boundaries and establish the sampling fraction for the take-some stratum such that the sample size is minimized while the target cell CV is achieved. A minimum stratum sampling fraction of 1% and a minimum sample size of three per stratum are used to ensure sufficient sample for estimation and variance calculation.

Due to the time lag between sample selection (November 1998) and mail-out (November 1999) and to the fact that CAPEX had used this sample in the field for two other survey occasions, there were already a large number of units identified as inactive before mail-out. No questionnaires were sent to these units and they remained in the sample with their original weights and zero data, representing other units not sampled in the population that are inactive.

### **2.2 Questionnaires and Collection**

The E-COM survey questionnaire was physically part of the CAPEX questionnaire. An appendix was added to explain definitions and concepts. The questionnaire

itself was comprised of six sections with 85 questions in total. Most of the questions were categorical: usually Yes/No with sub-categories to explain the No answers. The remaining questions were percentage distributions and a few numerical financial questions. Although the questionnaire was very long, skip patterns were used to help respondents navigate, and there were only two sections out of the six that had to be completed by all respondents.

CAPEX usually achieves response rates above 70% in terms of both counts and GBI coverage. However, after closing collection, we had achieved a response rate of about 50% for the E-commerce Survey. In order to improve this, we developed a short form questionnaire with only 14 questions, which was then faxed to all non-respondents. All of the 14 questions retained were categorical with one exception: Sales over the Internet. This short follow-up form contributed to an increase in the response rate to 65%. From this point on, only the short form questionnaire was considered.

Because the original response rate was so low, questions were raised about the fact that non-respondents might not be randomly distributed. Establishments that do not use electronic technologies might be less likely to respond. To assess the quality of the estimates and help design a valid non-response treatment method, a small sample of 250 non-respondents was chosen across all industries and they were asked three main technology questions: Use of computer, Use of Internet, Having a Web site. The answers of that small sample were compared with the results from the completed questionnaires. A  $\chi^2$ -test was done to verify if there was a difference between the two distributions. A p-value of 0.27 confirmed that respondents and non-respondents had similar behaviour.

### **2.3 Edit and Imputation**

Once the questionnaire was reduced to the 14-question short form, the task of edit and imputation was greatly simplified. During collection, checks were made to ensure that all categorical responses were internally consistent. Respondents that indicated that they had made sales over the Internet without actually specifying the actual value were followed-up by subject matter staff. All respondents who specified a non-zero value in this field were examined manually and validated.

As the questionnaire had a large number of categorical variables, we had decided to use an auxiliary variable as an economic weight in order to assess the

magnitude of the unit being surveyed. The auxiliary variables of choice were Gross Business Income (GBI) and the number of employees (NE). Unfortunately, they were not always available, either because the units were not on the Business Register (BR) or there was only default information on the BR. Values of GBI and NE were imputed when necessary.

A three-part strategy was used to resolve the remaining edit failures and missing values. First, no imputation was done for the 8,000 total non-respondents. The weights of the respondents were recalculated in such a way to re-distribute the weights of the non-respondents. Secondly, based on the questionnaire, it was possible to determine a pattern of each respondent, based on the questions: Use of computers, Use of Internet, making Purchases on-line etc. Where a pattern was incomplete and there was more than one valid choice of pattern to complete it, donor imputation was used to assign the final pattern. Thirdly, once a pattern was assigned, Statistics Canada's Generalized Edit and Imputation System (GEIS) was used to complete the records. Matching variables were chosen from the pattern variables and the two auxiliary variables (Gross Business Income and Number of Employees). We looked for donors at the lowest level of aggregation possible, based on industry and geography.

Based on the data, we observed that selling over the Internet in 1999 was a rare event. Of 27,000 sampled establishments, only 850 actually reported a non-zero value for Sales over the Internet. The final patterns, after imputation, indicated that an almost equal number of establishments should have had sales but did not report the amount. Three different imputation methods were attempted for this variable: mean, median, and nearest neighbour donor. Upon analysis, imputation using simple means at the Canada and Industry group was used, since this seemed to produce more acceptable data than either donor or median. It was also difficult to decide which observations were actually outliers and could not be used as donors, given the paucity of actual respondent data. For the numerical variable, Sales over the Internet, outliers were detected and excluded from being used as donors during imputation using the Hidiroglou-Berthelot (1986) algorithm available in GEIS.

#### **2.4 Transition from Establishments to Enterprises**

When the survey was designed, it was decided to use the establishment level of the enterprise structure as the unit of interest. However, further discussion and thoughts identified that Web-related concepts might be difficult to answer below the enterprise level. For

example, consider a retail enterprise with establishments in different provinces. Each individual store is unlikely to have a Web site, yet the overall enterprise may well have one. Therefore the survey strategy was revised, after the fact, to produce estimates at the enterprise level only. In order to accomplish this, we had to redo three survey steps: the analysis of the sample for data collection, the conversion of establishment data into enterprise data, and the calculation of estimates at the enterprise level.

Our first step was to analyze the sample in terms of complex and simple enterprise structures. An enterprise is complex if it has more than one establishment and operates in more than one province, in more than one NAICS or is linked to more than one legal entity. If a selected establishment belonged to a simple enterprise, then the data collected were those of the enterprise. However, if a selected establishment belonged to a complex enterprise, then the data collected would not necessarily correspond to those of the enterprise.

Secondly, we assessed the responses from establishments belonging to complex enterprises to see if they were likely already at the enterprise level, specifically for questions relating to the use of the Web. To do this, we searched the list of establishments to see if the enterprise head-office was part of the sampled units. Or, if one of the establishments had a Web site, we could assume that the enterprise had a Web site even though some other establishments could have answered differently. The cases where none of the establishments selected were head-office and none of them said they had a Web site, as well as cases where at least one establishment had declared a Web site but no Internet sales were identified and contacted by telephone to obtain data for the short questionnaire, but with an enterprise contact point. For the unambiguous units, the subject matter experts developed a set of rules in order to re-create the enterprise responses to the categorical questions. For the numerical variable, an estimate was derived based on the total number of establishments, both those selected and those not selected in the sample, linked to the enterprise: a weighted average was used.

Finally, since the sample design was establishment-based and we needed to produce estimates of enterprise data, we decided to use the weight share method (Lavallée, 1995). The information between enterprise and establishments is transferred using links. If an enterprise is linked to  $n$  establishments, then  $n$  links are assigned to the enterprise. The enterprise data are disaggregated to all of its establishments, even to the non-sampled ones. Each

variable (categorical and numerical) is simply divided by the number of links to form equal values for each establishment. Once these values are calculated, the sampled establishments are used in exactly the same manner as they were originally used under the establishment sample design. The variance and estimation formulae remain unchanged, but we use disaggregated enterprise data in the calculations, as if it was establishment data.

The calculation of the final weights was done through calibration and re-weighting. The calibration of the first phase sample was done in order to match to the GBI total. Once this was done, we went to the second phase where only the set of respondents and inactive units was weighted. Again it was calibrated using GBI to create the final weight.

One of the differences between using an establishment or an enterprise point of view is that the domain of estimation is now based on enterprise NAICS (instead of establishments NAICS). This leads to changes in the distribution of the estimates by industry. As well, although provincial estimates were part of the original requirements for estimation, they are not usually produced for enterprise data; since complex enterprises quite often operate in more than one province, it is difficult to measure their activities in each separate province.

## 2.5 Analysis of Results

The results were published in August 2000. They showed that computer technology was used in the private sector in 82% of enterprises, and in 100% of the public sector. However, the use of the Internet was far behind: only 53% of private enterprises had access, compared to 95% of public enterprises. Of all businesses, only 10% made sales over the Internet. Those Internet sales accounted for 0.2% of all operating revenue (Bakker, 2000).

Prior to the 1999 ECOM survey, there was no real idea of how much sales activity was taking place over the Internet in Canada. As it turned out, our survey estimates were well received by the media. The only negative aspect they pointed out was that the data would have been more useful if it had been published earlier! The United States had published current estimates of sales over the Internet by retailers, based on questions added to the Monthly Retail Trade Survey. Additionally, questions were added to existing annual production-based surveys; results for the 1999 reference year were released in 2001. Other OECD countries (Denmark, Australia, Finland, Sweden,

EUROSTAT) have followed a strategy similar to Canada's (OECD, 2000).

## 3. THE 2000 SURVEY OF ELECTRONIC COMMERCE AND TECHNOLOGY

Due to the tight time constraints, the 2000 survey used the same sample design as the 1999 survey: the CAPEX economy-wide sample of 27,000 establishments. However, before the sample went into the field for collection, it was analyzed. Any establishments belonging to a complex enterprise were grouped together and an enterprise contact was identified. The original 27,000 establishments became 21,000 enterprises. For certain key industries, such as retail and wholesale, additional enterprises were sampled. Since one of the objectives was to publish the estimates earlier, the mail-out was done separately from CAPEX, and took place in November 2000.

The most significant change was to identify the enterprise contacts for each sampled establishment before mail-out. The questionnaire that was used was similar to the short 14-question form that was faxed to non-respondents for the 1999 survey, although several questions pertaining to innovation were added. Extra care was taken with the wording of the questionnaire to be clear what concepts we were using. As well, historical data was now available for some sampled units, which helped during collection, edit and imputation, and at estimation.

To alleviate the non-response problems experienced in 1999, certain non-respondents were targeted for priority follow-up during collection. These priority units included units that had reported Sales over the Internet in the 1999 survey and other units where preliminary Web searching indicated that the enterprise had a Web site that was able to produce sales. Lastly, we tried to ensure that there was at least a certain level of coverage by industry and size of enterprise. At the end of collection, the final response rate was 77%, accounting for 93% of total GBI. A non-response bias study was carried out and it showed no differences between the respondents and non-respondents for certain key factors, grouped by industry and size.

A new system for edit and imputation was developed, based on the new short questionnaire. Patterns of response were used to validate where imputation was needed. Again, the financial variable of primary interest was Sales over the Internet. It was imputed using four different methods: historical, mean, median, and nearest neighbour donor imputation. The results were analyzed and in the end historical imputation was

used where the respondent showed the same pattern and had good 1999 data, and donor imputation for the other cases.

The weight-share method was used again at estimation, since we still had an establishment-based sample design, yet answers at the enterprise level. Once the weights had been distributed, they were not calibrated to the auxiliary variable (GBI): we felt the gain in 1999 was marginal and the correlation was not strong enough to justify the technique. Again, the estimates refer to enterprise level attributes.

### 3.1 Results

The results of the 2000 Survey of Electronic Commerce were published in early April 2001 (Statistics Canada, 2001). The 2000 survey yielded some important results. In 2000, private sector sales over the Internet amounted to \$7.2 billion. This represented a 73% increase from 1999. However, expressed as a percentage of total operating revenue, sales were 0.4%, increasing from 0.2% in 1999. The value of sales over the Internet was still small.

Also, between 1999 and 2000, the proportion of businesses selling on-line fell from 10% to 6%. For the enterprises that were in the sample for both 1999 and 2000, for every two businesses that started to sell on-line in 2000, five stopped. Finally, businesses that sold over the Internet in 2000 accounted for 25% of economic activity, an increase from 17% in the previous year. Electronic commerce appeared to become concentrated into fewer, larger businesses (Peterson, 2001).

When the results were released, they were well received by the media and analysts. While comparisons with the United States' numbers were still difficult owing to different methodologies, when compared to other OECD countries, the results showed that Canada was among the leaders in terms of ICT adoption.

### 4. CONCLUDING REMARKS

The design of the 2001 Survey of Electronic Commerce is now underway. A new economy-wide sample design was created, based on the population of enterprises. We expect that a similar questionnaire to the 2000 survey will be used. Similar methodology is planned for collection, follow-up, edit and imputation. Investigation into alternative estimation methodologies to improve the quality of the published estimates will take place, including the use of post-stratification and calibration.

We would like to spend some time analyzing the phenomenon that is electronic commerce. It would be helpful to have a better understanding of how businesses are organized for sales over the Internet. One potential study would identify some complex enterprises that are selling over the Internet, then ask one set of questions to the enterprise level contact, and a different set of questions to contacts at the establishment level.

It would also be useful to link these survey results with other data, such as corporate income tax data, to relate the behaviour of businesses in terms of their ICT use with outcomes such as their financial performance measures.

When Statistics Canada decided to develop a survey of electronic commerce for the 1999 reference year, we had no real idea of what to expect. It was difficult to design a sample to measure electronic commerce, since we had no frame data to use to predict sales over the Internet. Electronic commerce remains a rare event, and we had little understanding of how businesses operate when they are selling over the Internet. Over the last few years we have gained a tremendous store of knowledge about electronic commerce, yet many challenges.

### REFERENCES

- Bakker, C. (2000) "Information and Communications Technologies and Electronic Commerce in Canadian Industries," Statistics Canada, Catalogue number 88F0006XPB-0004.
- OECD (2000) "Defining and Measuring Electronic Commerce: A Provisional Framework and Follow-up Strategy," Paris.
- Hidiroglou, M.A. (1986) "On the Construction of a Self-Representing Stratum of Large Units in Survey Design", *The American Statistician*, Vol. 40, No. 1, 27-31.
- Hidiroglou, M.A., and Berthelot, J.-M. (1986) "Statistical Editing and Imputation for Periodic Business Surveys", *Survey Methodology Journal* Vol. 12, 73-83.
- Lavallée, P. (1995) "Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method", *Survey Methodology* 21, 25-32.

Peterson, G. (2001) "Electronic Commerce and technology Use", Connectedness Series, Statistics Canada, Catalogue number 56F0004MPE, no. 5.

Statistics Canada (2000) Private and Public Investment in Canada - Revised Intentions (Catalogue No. 61-206-XIB).

Statistics Canada (2001) The Daily, April 3. <http://www.statcan.ca/Daily/English/010403/d010403a.htm>

Whitridge, P. and Beaucage, Y (2000) "Statistics Canada's Electronic Commerce Survey", presented to the Advisory Committee on Statistical Methods, Statistics Canada, October 6.