

The implementation of tools to support the data quality of the Business Register at Statistics Canada

Mario Ménard, Canada

Abstract

Business Registers contains a vast array of information concerning enterprises. In Canada, it provides a complete picture of the relationship between the different operating entities that make up a business and its characteristics such as business status, geographic location, industry coding and size measure. Keeping all this information up to date and error free is a very challenging task.

The purpose of this paper is to present three initiatives that have been developed to report on the quality of the business structure and its characteristics. It will describe the implementation of the Monthly Quality Assurance Survey (QAS) which measures the quality of North American Industrial Classification System (NAICS) coding for all establishments on the Register, while at the same time determining the proportion of dead units on the frame. In addition, this paper will describe the Interceptor module which reviews survey feedback updates as well as the Survey Frame Assessments Tool (SFA) to report on frame changes from one reference period to another.

1. Introduction

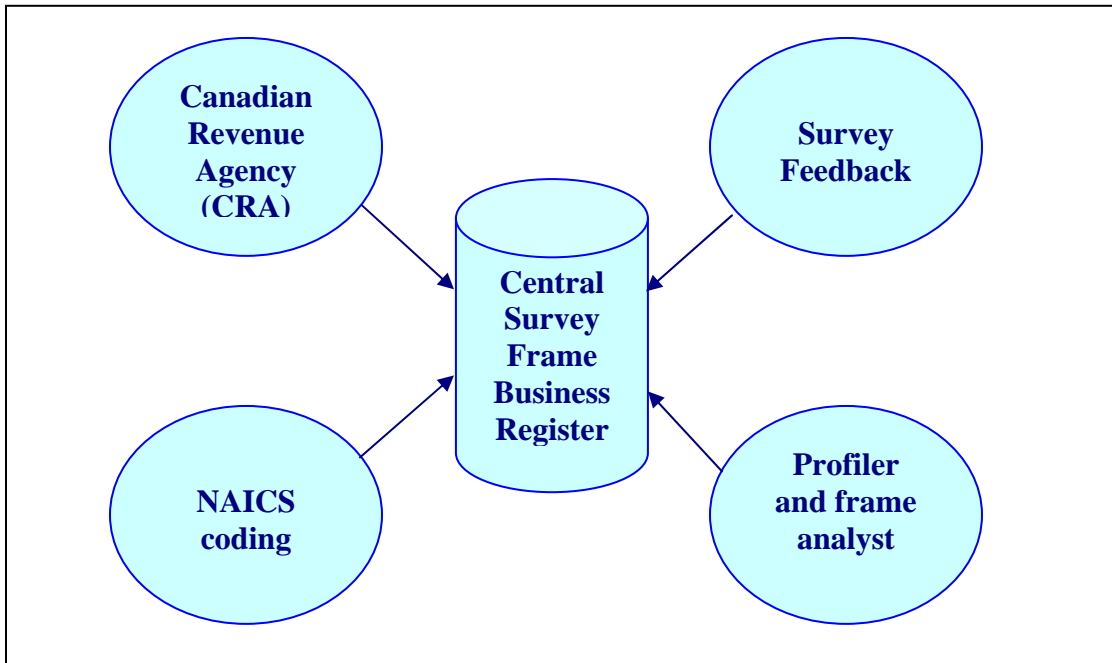
The survey frame is at the foundation of every survey program. At Statistics Canada, more than one hundred business surveys use it in various fashions to support their activities which include establishing a survey frame, sampling, collecting and processing data, and producing estimates. More than 1,500 employees throughout the Department use the Register to carry out their daily activities. Part of the numerous challenges faced by the maintenance of the survey frame is to ensure that the business structure and its characteristics are up to date and error free. Although there will never be enough resources to achieve this, numerous efforts have been made to report on the changes to the enterprise structure and its characteristics.

Errors or out dated data found on the survey frame have repercussions throughout the survey processing cycle. In order to further improve the survey frame, we need to identify weak areas and inform methodology and subject matter divisions so that they can take appropriate action in order to reduce the impact of these errors. It is also important to inform BRD managers so they may determine where additional energy/work needs to be allocated in order to eliminate, or at least reduce, the errors.

2. Populating and maintaining the survey frame at Statistics Canada

The survey frame is populated and maintained primarily from four different sources as demonstrated in Figure 1. Any one of them has the potential to introduce errors to the survey frame. Therefore, it is important to closely monitor the processes surrounding each of these four sources.

Figure 1 - Data sources used to maintain the survey frame



The next paragraphs briefly describe each data source and the element that needs to be monitored. Following these descriptions, we will demonstrate how the QAS, the Interceptor module and the SFA tool are used to identify and quantify these errors.

2.1. Canadian Revenue Agency (CRA)

The BR relies heavily on the use of administrative data to maintain its business population. In Canada, every business must register with CRA. As part of the registration process, CRA collects information such as legal name, business address and the major activity of the business. This information is sent to Statistics Canada on a monthly basis. The majority of these BNs consist of only one operating entity and they comprise 99 % of the Business Register Population, as indicated in Table 1.

Table 1 - Business Register Population Profile, January 1st 2008

	# of Businesses	% of BR Population	# of Operating Entities	% Total Revenue
Simple (one operating entity)	2 476 120	99%	2 476 120	55%
Complex (multiple operating entities)	21 300	1%	187 000	45%

Most of the updates from CRA are applied directly to the survey frame. However, the challenge faced by the BR is, once a BN is added to the survey frame as an active business, it is very difficult to determine when the same business is no longer active. Strong administrative signals exist to identify a closure and those signals are automatically applied. However, in many other cases, a BN is considered active by the taxation data authorities based upon a set of rules developed for the administration of their program. These rules do not always produce a timely indicator to remove the BN as an active business on the survey frame. The Quality Assurance Survey (QAS), described in section three, provides quantitative measures of this situation.

2.2. NAICS Coding

Each year, we must assign an industrial classification to more than 400,000 new businesses (business numbers) that have registered with the Canada Revenue Agency (CRA) in order to obtain a business number. Nearly 45% of these businesses can be coded automatically on the basis of the description of the principal activity which was obtained when the business registered with CRA. There remain some 220,000 businesses for which the description received is not specific enough and which require a manual intervention on our part. Survey managers depend on that information to delineate their population of interest. Having too many units misclassified will lead to higher variance, higher sampling and collection costs and frustration for survey respondents and/or collection personnel. It is therefore important to report on the quality of industrial coding on the survey frame. The Quality Assurance Survey (QAS) was developed to provide a quantitative measure of this problem.

2.3. Survey Feedback

Survey programs using the Business Register are committed to provide survey feedback. Data collection centers have procedures and support systems in place to transmit survey feedback to the BRD on a daily basis. Feedback is automatically applied to the base when it concerns such changes as updating the contact name for a particular questionnaire or the business address. However, changes to the business status, the industrial classification and the legal name are directed to a module called the Interceptor. This module, described in section four, is used by the subject matter analyst to review the feedback and decide whether it should be applied or not to the survey frame

2.4. Profiler /Frame Analyst

As previously stated, 99% of the BR population is composed of simple businesses and they are almost entirely updated through the use of administrative data. Most of the updates performed by the profiler and frame analysts in the organization are made to complex business (i.e., businesses that have more than one operating entity) which compose 1% of the BR population and represent 186 000 production units (see table 1). Although the Business Register uses data from CRA to identify the top legal entity, administrative data cannot provide the range of data required to maintain complex businesses on the frame. It is primarily profiling activities and survey feedback that maintain this population.

Data from financial statements and information about the business from their web site are valuable sources of information to delineate the operating entities that form an enterprise. In many cases, contact with the business is required in order to collect more detailed information such as the industrial activity of each operating entity and their accounting practices. These changes modify the picture of the business structure and their characteristics. For survey managers, it is important to know the changes that could potentially have a substantial impact on their survey estimates. The Survey Frame Assessment (SFA) tool, as described in section five, provides information to help subject matter be aware of the changes to their survey universe.

3. Quality Assurance Survey

The Quality Assurance Survey (QAS) is a monthly telephone survey of approximately 600 establishments carried out by BRD. Its purpose is to measure the quality of NAICS coding for all establishments on the Register, while at the same time determining the proportion of dead units. This survey used to be conducted every three years, it was recently decided that a monthly survey would improve both the survey's relevancy and accuracy. Taking this approach provides on-going quality indicators and regular feedback on the impact of changes made to improve these indicators.

3.1. QAS Methodology

The target population for the QAS is all establishments that are alive on the BR except for units within the public administrative sector (NAICS91). The survey population consists of the target population minus the largest complex businesses and those units that have given at least a partial response to a Statistics Canada survey in the previous twenty-four months. These units are excluded due to response burden concerns and a high level of confidence in the coding quality of these units. When producing estimates for the target population both groups of units are included and assumed to be active with a correct NAICS code

The sample design is a stratified simple random sample, using revenue and industrial sector as stratification variables. The industrial sector was chosen as a stratification variable to give a preliminary idea of the coding quality on the BR while keeping the survey at a reasonable cost. Revenue was chosen as the secondary stratification variable to allow for sampling among different sized strata. Monthly samples are drawn for each sector, with a monthly rotation feature to exclude units previously sampled by the QAS in the past 24 months.

3.2. QAS collection

Data collection for the QAS is carried out each month. Interviewers contact the selected establishments to obtain the required information. An establishment is considered to be dead in either of the following situations:

- The respondent confirms that the establishment is dead;
- The establishment cannot be located when researched.

If the establishment is in operation, then the interviewer first validates the name and the address of the establishment to ensure that he or she will be discussing the correct one. The interviewer then obtains the description of the main business activity and the NAICS code associated with this description. If the NAICS code provided by the respondent is different from the one available on the BR, the respondent is asked to identify which one most accurately describes this establishment's current main business activity. If the NAICS supplied by the respondent is more accurate, the respondent is asked if the NAICS currently listed on the BR was ever the main business activity of the establishment. If the answer is "yes", the respondent is asked to identify when the change in the main business activity occurred (which is referred to as volatility) and the change is attributed to volatility rather than to error.

3.3. QAS results

The QAS results are used to identify problem areas in NAICS coding so that training can be improved. The results will also be used by methodology and subject matter divisions to identify weak areas in the industrial classification process and make any adjustments required to their sampling strategies to compensate for these errors.

Table 2 indicates an annual estimate of 18.5% of businesses on the BR which have ceased operations (as determined by direct contact) but for which we received no signal from administrative data sources indicating that operations had ceased. Based upon these results, a review of our inactivation rules was initiated and the rules were modified. In July 2008, we were able to remove an additional 70 000 businesses. This change should be reflected on future QAS results.

At the 2 digit NAICS code level, the QAS indicates that 13.2% of the businesses are in error. This is indicative of the difficulty we encounter in determining the main business activity of an entity and assigning the appropriate description. A problematic area for statistical agencies in general, this error rate concerns us and we have initiated a project

to identify potential areas for improvement and potential solutions (review of problem cases, information sharing amongst staff etc.). The volatility aspect is small at the aggregate level, representing only 2.5% of the units on the base. The table also provides the breakdown by industry.

Table 2 - Quality Assurance Survey Results from June 2007 to July 2008, by industry

Sectors	Death Rate %	NAICS 2 Digit Error Rate %	Volatility Rate %
All sectors	18.5	13.2	2.5
11 Agriculture, Forestry, Fishing & Hunting	14.8	8.2	5.2
21 Mining, Oil and Gas Extraction	19.0	23.0	6.5
22 Utilities	14.5	7.6	-
23 Construction	17.4	7.3	0.1
31-33 Manufacturing	16.0	20.2	2.2
41 Wholesale Trade	21.8	24.2	5.7
44-45 Retail Trade	16.4	8.3	2.1
48-49 Transportation & Warehousing	21.6	12.8	5.0
51 Information & Cultural Industries	22.7	13.6	3.1
52 Finances & Insurance	19.2	22.7	2.2
53 Real Estate, Rental & Leasing	24.9	16.6	2.2
54 Professional, Scientific and Technical Services	18.4	10.5	0.1
55 Management of Companies and Enterprises	23.0	33.5	3.8
56 Administrative & Support	17.3	29.7	5.5
61 Educational Services	16.0	16.9	1.6
62 Health Care & Social Assistance	16.1	3.5	-
71 Arts, Entertainment & Recreation	20.2	12.0	2.6
72 Accommodation & Food Services	17.9	3.5	0.1
81 Other Services except Public Administration	18.8	9.9	1.7

The QAS results can be used by BRD to identify areas where coders are having difficulty coding, as indicated by high error rates in the results. This could lead to targeted training for BRD coders by BRD supervisors or non-BRD personnel in order to improve the coding quality within these areas. As well, the reviewing of sectors that are often mistakenly coded to one instead of the other can be used to improve training procedures in BRD by making coders aware of these frequently made errors.

The collection tools and methodology of the QAS can also be applied to obtain estimates for domains, other than those for which the survey was designed. For instance, they

could be used to produce domain estimates by tax remittance method (such as T1 or T2) or by different NAICS levels (i.e. 3, 4, 5, or 6 digit level) for various NAICS codes

4. Interceptor module

This module has been developed to allow subject matter analysts to review updates to legal name, business status or industrial classification that have been submitted by the data collection centers through survey feedback. It is estimated that these changes could have a substantial impact on the survey results and subject matter analysts want to have the opportunity to review the update before it is applied to the frame. In this module, they can cancel the requested update, indicate that the update has already been applied or allow the system to process the update as requested.

As indicated in table 3, the Interceptor module received 13 400 request for updates during the first 9 months of 2008. Out of these requests, subject matter specialist cancelled 4 900 of them, representing 36 % of the total request

Table 3 - Number and status of update requests received by the Interceptor module from January 2008 to September 2009

	# of update requests	# of update requests cancelled	% of cancelled requests	# of update requests accepted	% of accepted requests
Total # of update requests	13 500	4 900	36 %	8 600	64 %
Services Surveys	7 797	1 941	25 %	5 856	75 %
Distributive Trade Surveys	1 571	899	57 %	672	43 %
Manufacturing Surveys	1 533	919	60 %	614	40 %
Others	2 599	1 141	44 %	1 458	56 %

This table also shows the number of messages by type of surveys. In an ideal world, data collection staff would only send valid changes and these changes would be applied directly on the frame. However, with the high percentage of requests being cancelled, it proves that the Interceptor module is a useful tool that needs to be kept in place until we are able to ensure that the updates sent by collection staff are valid. We will be initiating discussions with the data collection centers and subject matter analysts to determine how we might reach this goal. One observation that has already been noted is that we need to improve training programs in the collection areas so they are better able to interpret information received by businesses during data collection and thus reduce the number of erroneous updates.

5. Survey Frame Assessment Tool

Numerous changes occur on the survey frame from one month to the next. Our profiler and frame specialists are updating the business structure and its characteristics on an on-going basis. Subject matter divisions are interested in understanding both the importance and source of these changes. A large increase or decrease in the number of establishments and any large change in the total revenue requires Methodology and Subject matter specialists to review the change and determine how it should be incorporated within their specific survey program. This is particularly important for the monthly survey program as any large variation could have an impact upon the estimates that they produce.

We have developed an analytical tool that displays all changes to the survey frame including the source and the nature of the updates. This tool accesses the list of all micro records that brought about these changes. The analysis is performed just before the release of the monthly GSUF by BRD staff. This is to ensure that the largest contributors to the change identified from one month to the other are legitimate changes and if this is not the case, take action to remedy the problem. This analysis also provides us with some indication of the quality of work being performed by the people making updates to the frame. In the cases where errors are found, this analysis allowed us to provide constructive feedback to the appropriate individuals. This analysis also provides a good indication of any additional training that may be required.

Table 4 display an example of the type of information provided by this tool. From this table, people can easily ascertain the impact of changes on the GSUF. For example, there were 349 establishments on the July GSUF that were not flagged as an establishment in the previous GSUF. This means that the business structure must have changed in such a way that a given operating entity is now flagged as an establishment. This added around \$9 400 billion to the revenue. In total, 24 973 establishments were added to the July GSUF and they increased the revenue by \$27 265 billion which represents about 1% of the total revenue on the BR. Similar tables are produced for units that were removed from the GSUF during the same period. For each type of change, the complete list of micro records is available and can be sorted according to different characteristics such as size measures or sources of the change.

Table 4 - Number of additional establishments and their revenue contribution by type of updates between the June 1st and July 1st G-SUF, 2008

Type of updates	# of Establishments	Total Revenue (000,000)
New establishments:		
1) From profiling activities	349	\$ 9 400
2) From assigning a NAICS code	11 800	\$ 5 699
3) From obtaining a size measure from administrative source	11 329	\$ 6 480
Changes to existing establishment:		
4) From profiling activities	1 100	\$ 5 600
5) From the Quality Assurance Survey	125	\$ 10
6) From a new description received from CRA	270	\$ 76
Total	24 973	\$ 27 265

6. Conclusion

Reporting on data quality is a key element in the effective use of the Register. Errors are inevitable. Therefore, developing the tools that report on and allow effective analysis of errors let BRD and its clients identify potential solutions and make effective use of the survey frame.