MEASURING THE QUALITY OF THE BUSINESS REGISTER: METHODOLOGY AND RESULTS

P. Lorenz and N. Laniel, Statistics Canada
P. Lorenz, 11-M, R.H. Coats Bldg, Ottawa, K1A 0T6, CANADA

KEY WORDS: Content Error, Coverage Error, Sample Measurement, Small Businesses

1. INTRODUCTION

The **Business Register (BR)** is a list frame for sub-annual business surveys at Statistics Canada. It is mainly built from two administrative sources from Revenue Canada Taxation (RCT): the **Tax returns** of corporations and individuals, and the **Payroll Deduction (PD)** accounts. All employers in Canada must have a PD account in order to remit monies for Pension Plans, Unemployment Insurance, and other deductions.

As any business frame, the BR is not perfect. Two types of error that occur within the frame are content and coverage errors. Content errors are errors in the data elements of a unit, and coverage errors are errors in the assignment of a unit to a survey population.

The Business Register Division at Statistics Canada has decided to undertake a semi-annual survey to measure errors in the BR. Some benefits of error measurement are that the cost of correcting the most serious problems can be ascertained. Also, with the estimates of each error, the impact on survey estimates can be determined. The benefits of measuring these errors have been discussed by Laniel and Finlay (1991).

The purpose of this paper is to present the concepts and definitions of the survey, the design of the survey, and some of the results from the first run of the survey. We also compare some of the results of this survey with the results obtained from a study of the old version of the BR.

2. CONCEPTS AND DEFINITIONS

In this section, the BR concepts and definitions of frame errors are presented from the point of view of its use by sub-annual surveys.

2.1 BR Concepts

The units on the BR can be categorized as to whether they are accessible or inaccessible to economic surveys. Accessible units have complete industrial classification code and complete geographic information, and are in-scope for economic surveys.

The accessible units can be assigned to one of two categories depending upon revenue. This assignment is based on the 2 digit Standard Industrial Classification (SIC) code and province. If the revenue for a unit is above the specified threshold for its industry by province group, it is placed into the **Integrated Portion (IP)** category. The IP units are complex or large businesses, and have the PD account data and the tax return data linked.

The remainder of the accessible units are referred to as the **Non-Integrated Portion (NIP)** of the frame. For these smaller units, the BR makes use of only the PD account data. These PD accounts are the sampling units for this survey.

The accessible units are available to economic surveys via the **Statistical list**, which contains the data necessary for sub-annual survey frame delineation and construction.

The inaccessible BR units are either units out-of-scope, the **ZIP - out-of-scope (ZIP/oos)**, or not completely classified. The latter means that either an industrial classification or geographical information is not known. These PD's are referred to as the **ZIP - unclassified (ZIP/unc)**.

Note that out-of-scope accounts are: accounts with no remittances in the last twelve months, household accounts, foreign accounts, accounts owned by businesses which have ceased their economic activities, government special work program accounts, pension plan accounts, or accounts used to settle an estate.

2.2 Definition of Frame Errors

The errors, for which measures are required, are of two types: **content** and **coverage** errors. General definitions of these two types of errors can be found in both Raj (1972), and Konschnik (1988).

a) Content errors

The different categories of content error for which measures have been produced are defined below.

1. Wrong Major Group
This category includes the units which have the wrong major group value. The major group is defined by the first two digits of the SIC code.

2. Wrong Industry Group
This category includes the units which have the wrong value assigned for the industry group. The industry group is defined by the first three digits of the SIC code.

3. Wrong Industry Class
This category includes the units which have the wrong value for their industry class. The industry class is defined by the first four digits of the SIC code.

4. Wrong Province
This category includes the units which have the wrong province value, as determined by the two digit standard geographical code.

5. Wrong CMA/CA
This category includes the units which have an incorrect three digit value for the Census Metropolitan Area or Census Agglomeration.

6. Wrong Legal Name
This category includes the units which have the wrong legal name for the PD account. The legal name is the name the business is known as for tax and other administrative purposes.

7. Wrong Operating Name

This category includes the units which have the wrong operating name. The operating name is the name the business uses in everyday economic activity.

8. Wrong Location postal code
This category includes the units which have an incorrect location postal code. Incorrect postal code information makes contact via mailout surveys difficult.

b) Coverage errors
The categories of coverage error for which estimates are reported in this document are defined as follows.

1. Duplicates
A duplicate is a NIP unit which corresponds to a PD account truly representing the same business activity as an IP unit. The duplication within the NIP, which might occur due to a business which has ceased to use an account for an activity and then opened a new account for the same activity, is not assessed by this survey.

2. Extraneous units
A NIP unit is extraneous if it is truly out-of-scope to economic surveys. Reasons for being out-of-scope have been mentioned in the previous section. (BR Concepts)

3. Superfluous due to Industry Misclassification
This category includes units which are superfluous for an industry division because their SIC code was wrongly determined to be part of that industry division.

4. Missing due to Unclassified Units
A ZIP/unc unit is missing in the NIP if it is truly in-scope to economic surveys.

5. Missing due to Industry Misclassification
A NIP unit is missing from an industry division if it has wrongly been SIC coded to another industry division.

3. SURVEY DESIGN

The first run of the survey was based on a sample survey of around 4000 NIP businesses, and a census of the ZIP unclassified units. There were approximately 17,500 units in the ZIP unclassified.

3.1 Sampling Frame

The sampling frame is composed of the NIP units and the unclassified ZIP units. The data used to build the frame mainly come from the BR PD account and Statistical lists and from the Master List used for the Survey of Employment, Payroll and Hours (SEPH) for the NIP units. For the ZIP units, the data mainly come from the BR PD account list.

3.2 Stratification

The NIP and ZIP units form two separate frames. The stratification of these two frames is presented separately below.

3.2.1 NIP Units

The first level of stratification for the NIP units is based on whether they belong to SEPH strata with New Entrants or not. These strata are further stratified as described below.

a) SEPH Strata with New Entrants

First, this portion of the frame is stratified with three variables. These variables are: the industry group, the province, and the number of employees size class (0-19, 20-49, 50-199, and 200+) (Schiopu-Kratina and Srinath 1986). In addition, this is further stratified as to whether units are births (i.e. new in the NIP) or not (non-births). This is to account for the difference in sampling rates for births and non-births.

b) Remainder of the NIP

The remainder of the NIP is stratified by the 18 industry divisions. Within each stratum in the remainder of the NIP, units are further stratified as to whether they are not eligible to be contacted (recent contacts), or eligible to be contacted. A unit is not eligible to be contacted if it has been contacted in the last 12 months. This was done in order to reduce respondent burden for the units recently contacted for updating frame data. The 18 industry divisions are as follows:

A. Agriculture
B. Fishing, Trapping
C. Logging, Forestry
D. Mining
E. Manufacturing
F. Construction
G. Transportation, Storage
H. Communication
I. Wholesale Trade
J. Retail Trade
K. Finance, Insurance
L. Real Estate Operators
M. Business Services
N. Government Services
O. Educational Services
P. Health and Social Services
Q. Accommodation, Food, Beverages
R. Other Services

3.2.2 ZIP Unclassified Units

The ZIP units are not stratified before the survey is taken.

3.3 Allocation and Selection

The sample allocation and selection are presented separately for the NIP and ZIP frames.

3.3.1 NIP Units

The allocation and selection are first described for the SEPH strata with New Entrants and then for the remainder of the NIP frame.

a) SEPH strata with New Entrants

The allocation and selection for these strata are predetermined by the SEPH sampling algorithms. The SEPH New Entrant sample coverage was then examined and strata where no New Entrants were selected were assigned to the supplemental portion of the frame. (Remainder of the NIP)

b) Remainder of the NIP

A supplemental sample is allocated for the remainder

of the NIP proportional to (the remainder of) the population size in each stratum of units eligible to be contacted. No units are allocated to the strata of units not eligible to be contacted.

The selection of the sample for the strata of units eligible to be contacted is performed as follows. Each stratum is sorted by province, by industry class within province, and finally a circular systematic sample is drawn (Cochran 1977).

3.3.2 ZIP Unclassified Units

All units in the ZIP unclassified portion are subject to some form of contact and thus included into the sample.

3.4 Data Collection and Capture

The data collection and capture procedures differ for the NIP and ZIP frames. A description follows.

3.4.1 NIP Units

The data collection for these units is performed in the Regional Offices (ROs) via a telephone interview. It is then captured at the RO and sent to Statistics Canada Head Office. Finally the survey data is edited and corrected by subject matter personnel.

3.4.2 ZIP Unclassified Units

The collection of data and its capture for the ZIP unclassified units are done in two steps. First RCT mails a form to the businesses which opened a new account. Copies of these forms are sent to STC Head Office for capture. Secondly, new PD accounts for which the RCT form is not received at STC within 2 months, are interviewed in RO's by phone. The resulting data is then captured by RO staff and sent to HO. Then, the survey data is edited and corrected.

3.5 Estimation

Many methods are used to produce the estimates of content and coverage errors required for this survey, depending on the stratum in question.

3.5.1 NIP Units

As for all previous steps in the survey methodology, we have to consider where the unit was sampled from before estimating the total number of units with each error type and the variances of these totals.

a) SEPH Strata with New Entrants

The estimates of total for the SEPH strata with New Entrants are calculated separately for the birth and the non-birth portion using the expansion estimator with the sampling weights adjusted for non-response.

For the SEPH strata the estimates of variance are calculated in one of two ways depending on the number of respondents in the stratum. When there is more than one respondent in a stratum the variance is calculated using the standard formula. When there is only one respondent in a stratum the method used to estimate the variance is the "collapsed strata" method. To apply the method, the strata with one respondent are paired within industry division, according to their population size (Cochran 1977).

b) Remainder of NIP

The remainder of the NIP consists of the units eligible for contact and those not eligible for contact. For the units that are eligible for contact the estimate of total is calculated using the expansion estimator adjusted for non-response. The units not eligible for contact have the estimated total calculated by prorating at the industry division level. The prorating consists of assuming the same error rate as the one calculated for the SEPH New Entrant birth strata. This is a good approximation since the units not eligible for contact have recently been contacted, like the births.

For the remainder of the NIP, the method of variance estimation depends upon whether the stratum is eligible for contact or not. For units eligible for contact, the standard variance formula is used. For units not eligible for contact, the variance is calculated as the squared prorating factor times the variance calculated for the SEPH New Entrant births. The prorating factor is of the form $(1+k)^2$. K is the ratio of the population size of the no contact strata over the population size of the birth strata.

3.5.2 ZIP Unclassified Units

The approach to produce estimates for the ZIP/unc units is post-stratification (Cochran 1977). More specifically the units are classified as to whether the estimated gross business income is above or below the upper quartile for the entire ZIP unclassified portion, the date of first remittance is within the last 12 months or not, the collection vehicle is the pd20 form or form 800, and the business status is active or inactive. This choice of post-strata has been made to account for non-response. It is assumed that within each of the post-stratum, the respondents are equivalent to a simple random sample.

Once the post-stratification weights are determined, the totals are estimated with the expansion estimator.

The estimate of variance is calculated with the standard variance estimator using the post-stratification weights.

4. SURVEY RESULTS

The NIP sample for this survey consisted of 3992 units. The response rate for the NIP portion was 90% overall.

The ZIP unclassified units were covered by a census approach. Overall, the ZIP had a response rate of 79.61%.

Table 1 presents the coverage error results for the entire NIP population. The estimated number of extraneous units is 14.33% of the population size and is the largest frame error. As the NIP portion of the BR is made up primarily from payroll deduction accounts, a unit is not flagged as inactive until there have been no remittances for 13 months. This time period is to account for seasonal business activity. In the current economic picture, the length of time in assigning an inactive flag to a business could be causing the high number of extraneous units. This result, even though it is high, is an improvement from a 1985 study of the BR, where the number of extraneous units was found to be 19.24 percent of the NIP population. Survey programs can adjust the sample to minimize the effect of extraneous units on the variance of estimates. Duplicates of IP businesses are estimated to make up only 0.18% of the population. This is an underestimate since it is difficult to collect appropriate linkage data from respondents. Units that belong in NIP but are currently in the ZIP unclassified portion are estimated to be 0.41% of the NIP population

size.

Table 2 shows the content error results over all industry divisions as well. The percentages in this table refer to percentage of the estimated active units in the NIP. The estimated active units is the NIP population minus the estimate of extraneous units. This total is estimated at 668,730 units.

As it can be seen from the table, the wrong legal name and the wrong operating name are the most common content errors, occurring in 11.63 and 12.51 percent of the active units respectively. This relatively high estimate may be due to differences in the definition of legal and operating name variables in the source files initially used to create the BR list.

The industry code estimates have been determined to be acceptable by the BR staff. The wrong major industry group estimate is 6.98% of the active population, while the wrong industry group is 8.23%, and the wrong industry class is 8.70%. As a comparison, the industry class coding in 1985 was found to be incorrect on 21.3% of the records in the NIP.

The geographical information is also considered fairly good. The provincial information is wrong 0.20% of the time, as compared to 1985 when 1.4% of the units had an incorrect province. The Census Metropolitan Area/Census Agglomeration is incorrect 2.68% of the time. The physical business location postal code is incorrect 7.95% of the time.

Table 3 shows the industry misclassification errors within each of the 18 major industry divisions. The largest estimate of superfluous businesses is in retail trade, where 6108 businesses were estimated to be superfluous to that industry division. In terms of percentage, mining has the most superfluous units, with 20.3% of the population estimated to be superfluous.

The largest estimate of number of businesses missing from an industry division is in real estate operators, where 4518 businesses were estimated to be missing from the division. In terms of percentages, educational services has missing units that account for 34.4% of the industry division population size.

The results of the first run of the survey have been published and made available to the client divisions of the BR division through a special issue newsletter.

5. CONCLUSIONS

The first run of the survey has provided some informative results for assessing the quality of the NIP portion of the Business Register.

5.1 Assessment of Results

The content errors are considered by the BR staff to be acceptable overall, given the dynamic nature of the frame. The estimate of extraneous units for the coverage errors is a cause for some concern. It may be necessary to develop a new way to determine if a business is active or not during recessionary times, rather than waiting 13 months to record it as inactive.

The results of the first run of the survey are to be discussed with the clients of the BR. Their input into interpreting the results as they pertain to their surveys will aid in determining the areas of the BR that are in most need of improvement.

As this survey is going to be a semi-annual undertaking, it will be possible to determine whether the Business Register is improving or deteriorating over time. The next two runs of the survey are scheduled for June 1992, and November 1992. The results from these runs will enable further assessments of the quality of the Business Register.

5.2 Proposed Changes and Studies

The initial run of this survey has given rise to some new ideas on improving the survey design itself. One of the proposed changes is to provide estimates of errors in terms of the number of employees and estimated gross business income. This will provide a better indication of the impact of the errors in the NIP population on the economic survey estimates.

To provide the best estimates possible, improvements to the survey design are being considered. The stratification of the population is being considered as one possible improvement, perhaps stratifying the remainder of the NIP (i.e. non-SEPH) units by size of units would be useful. A second area where the methodology may be improved is the allocation of the sample in the remainder of the NIP. As the survey currently uses allocation proportional to stratum population size, it may be beneficial to use Neyman allocation. Another study that should be considered is to investigate the use of imputation for non-respondents instead of adjusting the sampling weights. This could be done using techniques such as hot deck or the nearest neighbour imputation.

REFERENCES

Business Register Division, "B.R. Quality Report", BRD Newsletter Special Issue, Statistics Canada, 1992

Cochran, W.G.,Sampling Techniques, 3rd Edition, John Wiley & Sons, 1977

Konschnik, C.A. (1988), "Coverage Error in Establishment Surveys", Proceedings of the Section of Survey Research Methods, American Statistical Association

Laniel, N. and Finlay, H.(1991),"Data Quality Concerns with Sub-Annual Business Survey Frames", American Statistical Association

Raj, D. (1972), The Design of Sample Surveys, Chapter 8: Sampling from Imperfect Frames, pp 126-138, McGraw-Hill

Schiopu-Kratina, I. and Srinath, K.P. (1986), "The Methodology of the Survey of Employment,

Payroll and Hours", Working Paper no. 86-010E, Methodology Branch, Statistics Canada, Ottawa

TABLE 1: COVERAGE ERRORS OVER ALL INDUSTRY DIVISIONS

NIP POPULATION SIZE: 780,549					
Category	Estimated Count	Estimated Percentage of Population	Lower 95% Bound for Percentage	Upper 95% Bound for Percentage	Coefficient of Variation
Duplicates	1,398	0.18	0.01	0.35	0.48
Extraneous	111,819	14.33	12.40	16.25	0.07
Missing due to Unclassified Units	3,205	0.41	0.40	0.42	0.01
Adjusted Population	670,537	85.91	83.97	87.84	0.01

TABLE 2: CONTENT ERRORS OVER ALL INDUSTRY DIVISIONS

NIP ESTIMATED ACTIVE POPULATION SIZE: 668,730					
Category	Estimated Count	Estimated Percentage of Active Population	Lower 95% Bound for Count	Upper 95% Bound for Count	Coefficient of Variation
Wrong Major Group	46,703	6.98	39,577	53,829	0.08
Wrong Group	55,061	8.23	47,170	62,952	0.07
Wrong Class	58,190	8.70	49,843	66,537	0.07
Wrong Province	1,348	0.20	0	2,597	0.56
Wrong CMA/CA	17,907	2.68	12,582	23,232	0.15
Wrong Legal Name	77,800	11.63	66,019	89,580	0.08
Wrong Operating Name	83,671	12.51	73,192	94,149	0.06
Wrong Postal Code	53,175	7.95	41,779	64,570	0.10

TABLE 3: SUPERFLUOUS AND MISSING ESTIMATES FOR EACH INDUSTRY DIVISION

Industry Division	Superfluous to Industry Division	Percent of Industry Popn size	Missing from Industry Division	Percent of Industry Popn size
Agriculture	4,181	7.8	744	1.3
Fishing, Trapping	611	10.3	601	10.1
Logging, Forestry	1,120	15.5	402	5.5
Mining	870	20.3	636	14.8
Manufacturing	2,768	6.1	3,788	8.4
Construction	2,835	2.7	3,855	3.7
Transportation, Storage	821	2.7	1,824	6.0
Communication	236	5.6	4	0.0
Wholesale Trade	5,445	12.3	4,067	9.1
Retail Trade	6,108	4.6	4,385	3.3
Finance, Insurance	1,457	7.1	3,845	18.8
Real Estate Operators	653	2.0	4,518	14.1
Business Services	3,816	5.1	2,300	3.0
Gov't Services	0	0.0	0	0.0
Educational Services	267	4.8	1,910	34.4
Health and Social Services	856	1.3	616	0.9
Accommodation, Food, Beverages	490	0.8	1,402	2.4
Other Services	4,986	5.0	2,620	2.6