

DEALING WITH INDUSTRY MISCLASSIFICATIONS IN THE UNIFIED ENTERPRISE SURVEY

David MacNeil and Stuart Pursey¹

ABSTRACT

The Unified Enterprise Survey brings together many of the industry surveys of Statistics Canada that were formerly isolated from each other. This integration provides an opportunity to use businesses that are discovered, during data collection, to be “misclassified by industry”. This paper describes the amount of industry misclassifications that have been found and what can be done during data collection and data processing after the correct classification is determined. It also proposes several methods that can be used during estimation to improve industry estimates.

KEY WORDS: collection, estimation, adjustment

RÉSUMÉ

L'Enquête unifiée sur les entreprises rassemble plusieurs enquêtes sur des industries faites par Statistique Canada et qui étaient autrefois isolées les unes des autres. Ce jumelage fournit une occasion d'utiliser les entreprises pour lesquelles on réalise, lors de la collecte de données, qu'elles sont “classées dans la mauvaise industrie”. Ce document décrit la quantité de mauvaises classifications industrielles qui ont été notées et ce qui peut être fait lors de la collecte et l'analyse des données après que la bonne classification a été déterminée. Il propose aussi plusieurs méthodes qui peuvent être employées pendant l'estimation pour améliorer les estimations dans chaque industrie.

MOTS CLÉS: collecte, estimation, ajustement

1. INTRODUCTION

The Unified Enterprise Survey is a result of Statistics Canada's Project to Improve Provincial Economic Statistics (PIPES). The objective of PIPES was to provide the data needed to support the formula of the Harmonized Sales Tax (HST). The HST developed from an agreement between the Government of Canada and the provincial governments of Newfoundland, Nova Scotia and New Brunswick to harmonise the federal Goods and Services Tax with the provincial sales tax. More details on PIPES can be found in Royce, Hardy and Beelen (1998) and Smith (1998).

The UES began first as a pilot for the reference year 1997. As of the reference year 2002, it brings into a single survey about 20 annual industry surveys that were isolated from each other in the past. By revenue, the UES covers about 65% of the Canadian economy. The total sample size of the UES is about 68,000 units.

Industry misclassification occurs when the industry classification that existed on the frame at the time of sampling differs from the industry classification reported during data collection. When a unit is discovered to have an incorrect industry classification during data collection, it is coded out-of-scope and no data are collected. This approach can cause a downward bias in the estimates because the out-of-scope units may not be accounted for in the final estimates of their correct industry classification. Before describing the amount of industry misclassification and steps that can be taken to improve the estimates, this paper will provide a brief description of the sampling frame and survey steps in the next section. It will then present the magnitude of the problem, ways to reduce it, and estimation methods to correct it.

¹ David MacNeil and Stuart Pursey, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6, Dave.Macneil@statcan.ca.

2. METHODOLOGY OF THE UES

Although there are some exceptions, where possible the methodology and systems of the UES are common for all of the industry surveys that belong to it.

2.1 Frame

The frame for the UES is the Business Register (BR) which is a list of Canadian businesses that is maintained by Statistics Canada. For every unit on the BR, we have an industry classification, a province code and an estimate of the total revenue of the business, the gross business income (GBI). The industry classification used is the North American Industrial Classification System. Links to tax are available and it is possible to obtain tax data for most units on the BR.

2.2 Sampling

The UES sample excludes the smallest units on the frame. Tax data is used to estimate their contribution to the economy. This helps to reduce costs as well as the response burden of the smallest businesses. The remaining units are stratified by province and industry group. Within a given province and industry group, three strata are defined using the gross business income variable. The boundaries between these three strata are set optimally using the Lavallée-Hidiroglou (1988) algorithm. More details on the sample design can be found in Simard, Girard, Parent, and Smith (2001).

2.3 Data Collection

Collection of data is by mail-out / mail-back questionnaires with telephone follow-up for non-response and edit failures. Collection edits built into the CATI system (Computer Assisted Telephone Interview) are used to detect and resolve data inconsistencies during follow-up. There is a separate collection instrument for each survey and collection periods can differ by survey.

2.4 Data Processing

Imputation is used to produce a complete record for every unit in the sample, even those with complete non-response.

More details on imputation can be found in Martin and Laroche (2001).

2.5 Estimation

Before estimation, outlier detection and correction is done to prevent extreme size stratum jumpers. An example of a size stratum jumper would be a unit that has a small GBI value on the frame and is selected in a small size stratum with a large weight, but when the data for the unit is collected it reports a very large revenue value. The original sampling weights are used in estimation. However they are modified and improved using post-stratification techniques with additional frame information that becomes available at the time of estimation. Calculation of the domain estimates and variances is done using the Generalized Estimation System developed by Statistics Canada.

3. AMOUNT OF INDUSTRY MISCLASSIFICATION

Table 1 below shows that close to 10% of the UES sample units in both 1999 and 2000 were found, at collection, to be industry misclassifications. Just over half of these were re-coded to an industry within the UES and the others re-coded to an industry outside the UES. There is no information from non-UES industry surveys – that is, we are not able to determine the number of businesses in non-UES industries that should be within the UES.

The estimation of total revenue by industry is a key estimate. Table 2 extends Table 1 by showing the percentage of total revenue that is associated with sample units that are industry misclassifications. The problem is less significant than an examination of the simple counts in Table 1 would imply.

Tables 1 and 2 show results for the entire UES. Estimates are published by province and industry and so a more detailed analysis is needed. The distribution of the problem is not uniform. Certain industries are more prone to industry misclassifications than others. Also what an industry “gives” to other industries is not balanced by what it “receives” from others. The problem is clustered among industries. Table 3 shows an example of clustering among three 1999 UES industries.

Table 1 – Industry Classification at the time of Collection (Percentage of the sample size)

	Sample Size	Determined at collection to be the same as at sampling	Found to be different at collection and re-coded <u>outside</u> of the UES	Found to be different at collection and re-coded elsewhere <u>within</u> the UES
1999 UES	50,587	91.1	3.3	5.6
2000 UES	51,433	93.1	3.4	3.5

Table 2 – Industry Classification at the time of Collection (Percentage of the sample weighted revenue)

	Sample Size	Determined at collection to be the same as at sampling	Found to be different at collection and re-coded <u>outside</u> of the UES	Found to be different at collection and re-coded elsewhere <u>within</u> the UES
1999 UES	50,587	96.3	1.2	2.5
2000 UES	51,433	97.2	1.2	1.6

Table 3 – Pattern of movement between some UES industries (1999 UES)

		Classification at Collection			
		Retail Non-Store	Retail Store	Wholesale	Other UES
Classification at Sampling	Retail Non-Store	2,311	254	352	152
	Retail Store	21	11,320	221	89
	Wholesale	2	210	10,459	81
	Other UES	4	240	185	

4. DATA COLLECTION AND DATA PROCESSING

The method used to improve the industry estimates will depend on the data that is available about the reclassified units (those re-coded to an industry within the UES) after data collection and data processing. In addition to the correct classification, the data available might be total revenue, a few key variables or all variables.

4.1 Data collection

The total revenue variable could be collected with minimal impact on the data collection process. This is because the total revenue variable already exists on all industry surveys. If a few key variables could be identified that exist on all industry surveys, then these too could be collected without adding any significant burden to the data collection process.

It is not currently possible to collect all variables for all reclassified units. To do this, all industry surveys would have to exist in the same collection instrument so that units

could easily switch from one industry survey to another. If all industry surveys existed in the same collection instrument, the performance of the collection instrument would be greatly reduced. More importantly, this would mean that data collection would not finish for any industry survey until all industry surveys were complete, something that is unacceptable when considering the timeliness of the estimates.

One compromise might be to integrate certain industry surveys together in the same collection instrument when there is significant movement between the industries. By integrating just a few industry surveys, it might be possible to collect complete data for the majority of the reclassified units. If only a few industry surveys existed in the same collection instrument then the performance of the collection instrument should not be significantly affected. The integrated surveys would need to have similar collection periods so that the timeliness of the data was not affected.

During data collection, it is very important that the correct classification of these out-of-scope units is transferred to

the Business Register so that it is corrected for the next survey cycle.

4.2 Data processing

The imputation process could be used to create complete micro records for all reclassified units. However, if only total revenue or a few key variables are collected, this would represent a significant burden on the imputation process. The number of units requiring imputation would increase while the number of eligible donors would remain the same. In the case of integrated surveys, where all variables would be collected, it is assumed that the ratio of the number of units requiring imputation to the number of eligible donors would not change. This would impose much less of a burden on the imputation process.

5. ESTIMATION

The true estimate of a given UES industry consists of the following three components:

- 1) units with a correct industry classification on the frame
- 2) units originally misclassified to another UES industry on the frame
- 3) units misclassified to a non-UES industry on the frame

The current UES estimates are based on the first component only, resulting in undercoverage. It is possible to improve the estimates and reduce the undercoverage by incorporating the second component. Unfortunately, it is difficult to incorporate the third component because these units are not part of the UES. Non-UES surveys and administrative sources cover these units. The options available to incorporate the second component are applying macro adjustments, using micro records directly or some combination of the two.

5.1 Macro adjustments

Currently, for every reclassified unit, the following information is available:

- 1) The correct classification determined during collection
- 2) The revenue of each reclassified unit as reported on tax
- 3) The probability of selection and therefore the design weight

With this information, macro adjustments can be calculated to account for the units that were reclassified to other UES industries during data collection. The macro adjustments would be applied to the current estimates that are calculated using only the in-scope units, units that were classified to the correct industry survey at the time the sample was selected. One of the main objectives of the UES is to obtain accurate provincial estimates, therefore it

is necessary to calculate a separate macro adjustment for each province within a given industry or sub-industry.

Three different macro adjustments were considered:

Adjustment 1: The sum of the weights of the in-scope units plus the sum of the weights of the reclassified units all divided by the sum of the weights of the in-scope units.

Adjustment 2: The sum of the weighted tax revenue of the in-scope units plus the sum of the weighted tax revenue of the reclassified units all divided by the sum of the weighted tax revenue of the in-scope units.

Adjustment 3: The sum of the weighted survey revenue of the in-scope units plus the sum of the weighted tax revenue of the reclassified units all divided by the sum of the weighted survey revenue of the in-scope units.

Table 4 shows the percentage increase in the total industry estimates resulting from applying the three different macro adjustments. Reference year 1999 data was used.

For almost every industry, adjustment 1 resulted in an overestimate of the second component of the estimates, units originally misclassified to another UES industry on the frame. This is because the mean of the reclassified units was typically smaller than the mean of the in-scope units. For some industries, adjustment 2 also resulted in an overestimate of the second component of the estimates. This occurred when the tax estimate of the in-scope units was underestimated, causing the impact of the reclassified units to be overestimated.

Adjustment 3 provides the best estimate of the second component. Table 4 shows that the adjustment is significant ($> 1.5\%$) for 7 of the 20 industries. A detailed analysis would show that almost every industry has some domains where the impact is significant.

It is important to consider that the results here are preliminary and do not include any testing for outliers or validation by subject matter specialists. Outlier detection would be done on the weighted tax values to ensure there are no extreme combinations of large revenue and large weights that would unduly influence the estimates. The subject matter specialists would be required to validate the tax data as well as whether or not the unit has been correctly reclassified.

An approximation of the variance that the reclassified units contribute could be obtained using the tax data of the reclassified units.

Table 4 – Percentage increase in the total industry estimates of revenue

Industry	Adjustment 1	Adjustment 2	Adjustment 3
Accounting & Bookkeeping	2.3%	2.8%	0.9%
Aquaculture	0.1%	0.0%	0.0%
Construction	3.0%	0.6%	0.4%
Couriers	5.0%	1.8%	1.8%
Database Publishers	1.5%	0.9%	0.6%
Food Services	5.5%	0.1%	0.0%
Geomatics Services	7.0%	4.3%	3.7%
Lessors	1.0%	0.2%	0.2%
Management Consultants	9.8%	4.0%	3.9%
Newspaper Publishers	5.7%	2.6%	2.3%
Retail Non-Store	0.8%	0.7%	0.3%
Real Estate Agents	3.1%	1.1%	0.9%
Retail Store	3.8%	0.4%	0.4%
Rep. & Maint. Auto.	11.6%	2.4%	2.3%
Rep. & Maint. Non-Auto.	38.6%	5.0%	4.0%
Specialized Design	3.6%	1.8%	1.0%
Taxis & Limousines	8.9%	1.9%	1.2%
Testing Labs	0.0%	0.0%	0.0%
Translation Services	0.8%	0.0%	0.0%
Wholesale	8.3%	6.1%	1.9%

If the UES could collect the total revenue variable for the misclassified units, the tax revenue variable could be replaced with the collected revenue variable in adjustment 3. This should lead to an improvement in the quality of the macro adjustments. If additional variables (i.e. total expenses) could be collected then different macro adjustments could be calculated for different sections of the questionnaire.

5.2 Micro records

In section 4, it was explained that it would be difficult to obtain complete micro records for every reclassified unit. It is not currently possible to collect all of this information, and it would put a significant burden on the imputation process to impute this information. However, if complete micro records were available for every reclassified unit then domain estimation could be used to produce estimates that include the contribution of units originally misclassified to another UES industry on the frame.

To process Industry A, all of the units that were initially classified to Industry A would need to be included, as well as all of the units that belong to an industry that contains

units that were reclassified to Industry A. All units belonging to an industry that does not contain units that were reclassified to Industry A would be excluded from the domain estimates. In theory this is the best approach but it is not necessarily the most practical. Currently, every industry is processed separately and for some industries the amount of processing time required for the variance calculation is very high. If many industries were processed at the same time, this could impact the timeliness of the data.

5.3 Macro adjustments and micro records

In section 4.1, the possibility of integrating certain industry surveys together in the same collection instrument was mentioned. If this was implemented then the estimates could be improved with a combination of using micro data directly for the integrated surveys and applying macro adjustments for any reclassified units coming from the remaining industry surveys that were not part of the integration. This would be a significant improvement over only applying macro adjustments.

6. SUMMARY

Misclassification is a source of error in the UES estimates. There are various methods available to improve the estimates. Currently, macro corrections based on tax revenue can be applied. The quality of the macro corrections can be improved by collecting and using survey revenue. It is possible that further improvements may be realised by integrating certain industry surveys with significant movement back and forth between classifications.

It is not currently possible to account for all misclassification because there is no good measure of the impact of UES units that are misclassified on the frame to non-UES industries. However, this will become less of an issue when more industries are added to the UES. The problem of undercoverage itself should become less significant over time, as the frame is improved due to repeated surveying.

REFERENCES

- Lavallée, P. and M. Hidioglou (1988). *On the stratification of skewed population*. Survey Methodology Journal, no.14, pp 33-45
- Martin, C. and J. Laroche (2001). *A New Approach to Processing for the Unified Enterprise An Edit and Imputation Example*. Internal Statistics Canada Report, PIPES Technical Series Paper #65
- Royce D, Hardy, F. and Beelen, G. (1998). *An overview of the Project to Improve Provincial Economic Statistic*. Internal Statistics Canada Report, PIPES Technical Series Paper #6
- Simard M., C. Girard, M. Parent and J. Smith (2001). *Sampling Designs for the Unified Enterprise Surveys: The Early Years*. Internal Statistics Canada Report, PIPES Technical Series Paper #69
- Smith P. (1998). *Realising and Measuring Quality Improvements in Provincial Economic Statistics*. Internal Statistics Canada Report, PIPES Technical Series Paper #15