Measuring the Complexity and Importance of Businesses in Order to Better Manage our Data Collection Efforts

Serge Godbout and SungJin Youn

Serge Godbout, Statistics Canada, serge.godbout@statcan.gc.ca
SungJin Youn, Statistics Canada, sungjin.youn@statcan.gc.ca

1. Introduction

The Unified Enterprise Survey (UES) is an establishment-based program and covers a large portion of the Canadian economy (Brodeur et al., 2006). The domains of interest are defined by a two-dimensional grid from industrial (using the North American Industrial Classification System, NAICS) and geographical (Canadian provinces and territories) components. To provide estimates that meet the users' requirements, the UES combines data collected from a sample and tax data.

Both data sources bring their challenges. Data collection is resource consuming, relationships with respondents need to be carefully managed and compromises have to be made on the required level of desegregations to reduce response burden. On the other hand, tax data are cheaper and reduce response burden but the processing is different and data are available at a higher aggregated level so they need to be allocated into the structure using a profiling process. The obstacles related to both data sources get more complicated with the complexity of a business structure. This highlights the significance of being able to measure the complexity of a business structure and to identify the ones that are the most critical to estimates.

The focus of the paper is on describing the methodological framework developed to help the business profiling and the response management of the UES. In section 2, we describe the challenges coming from the profiling process on the Business Register, the UES collection and response management process. In sections 3 and 4, we present the complexity metric developed to rank the largest and most complex businesses and the rules applied to identify the businesses that are the most important to a survey in order to better manage the data collection efforts. In section 5, we show some results based on 2008 UES production cycle. Finally, a conclusion is given in section 6.

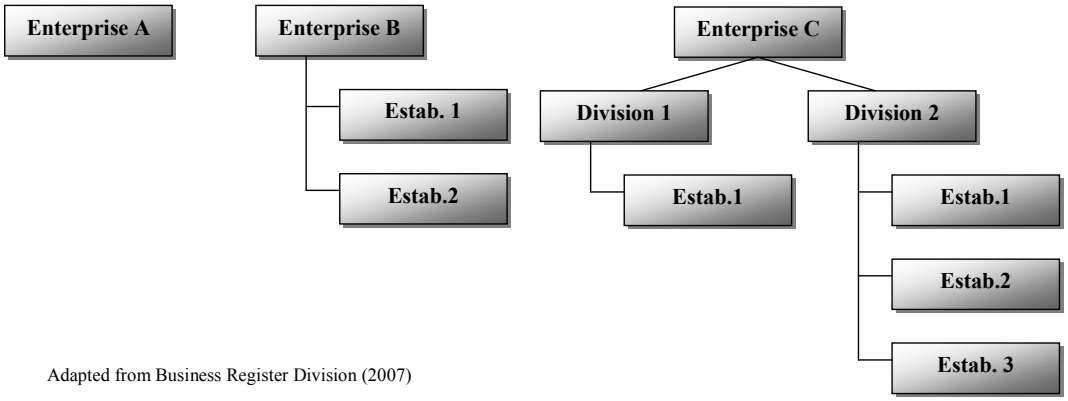
2. Profiling and collection challenges

2.1 Business structure and profiling on the Canadian Business Register

Statistics Canada's Business Register (BR) contains key information on businesses in Canada, like the activity status, industrial and geographical classifications, business structure, many size variables, contact information etc. All entities that generate economic activity in all Canadian economic sectors are covered by the BR. For more details, see Business Register Division (2009).

All businesses are classified into simple or complex. A simple business is involved in only one economic activity in only one location. The other businesses are called complex; they have economic activities in more than one industry, province or location (Beaucage, Hunsberger and Pursey, 2005). In figure 1, business A is simple and businesses B and C are complex, although C's structure looks much more complex than B's.

Figure 1: Some examples of simple and complex businesses



Adapted from Business Register Division (2007)

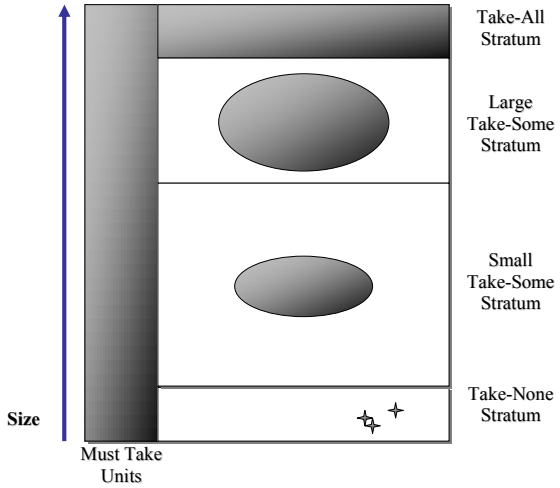
Among the size variables provided by the BR, we are more interested in the total revenues and the number of employees. These variables are available from tax data at the business level. For a business with a complex structure, tax data need to be allocated into the operating structure in order to meet the survey requirements. This is done by applying a ratio derived from profiling information provided by a profiling specialist. A business profile needs to be updated periodically because of economic fluctuations, business restructuring etc. Since this process is costly, it is important to prioritize the complex businesses to help managers choosing the best updating plan for each complex business.

In the current BR process, there is a subset of the complex businesses called the BR profiling critical units. This subset includes roughly 5000 businesses identified by subject-matter specialists for profiling updates. The identification process is done under broad guidelines provided by the BR and it is not consistent between the industrial sectors since each specialist is attached to one industry sector. Another weakness is that the importance of the multi-sector businesses might be underestimated compared to the large and complex ones under one sector.

2.2 The Unified Enterprise Survey sampling, collection design and response management

In the UES sampling design, all the establishments of a given business that are in the same domain of interest are clustered to form a sampling unit (SU). First, some SUs are automatically placed in the sample as must-takes (MT). Then, the rest of the UES population is stratified in take-all (TA), take-some (TS) and take-none (TN) strata using the Lavallée-Hidiroglou algorithm (Lavallée and Hidiroglou, 1988) as shown in figure 2.

Figure 2: Unified Enterprise Survey stratification



The sample is selected through a network sampling design, i.e. a preliminary simple random simple (SRS) is selected in each stratum first, then all non-selected SUs belonging to a business for which at least one of its SUs – within the same industrial

grouping as the non-selected one – were selected in the SRS sample are brought in the final sample. For more details on the UES sampling design, see Simard, Girard, Parent and Smith (2001).

After the sample selection stage, the selected SUs of a business are reorganized into collection entities (CE) in order to help managing the data collection process and reducing the response burden. A UES CE is defined as being one or a group of SUs from one business that belongs to the same industrial class, using preset default rules or customized according to the respondents' requirements.

Figure 3: Sampling units and collection entities for complex businesses

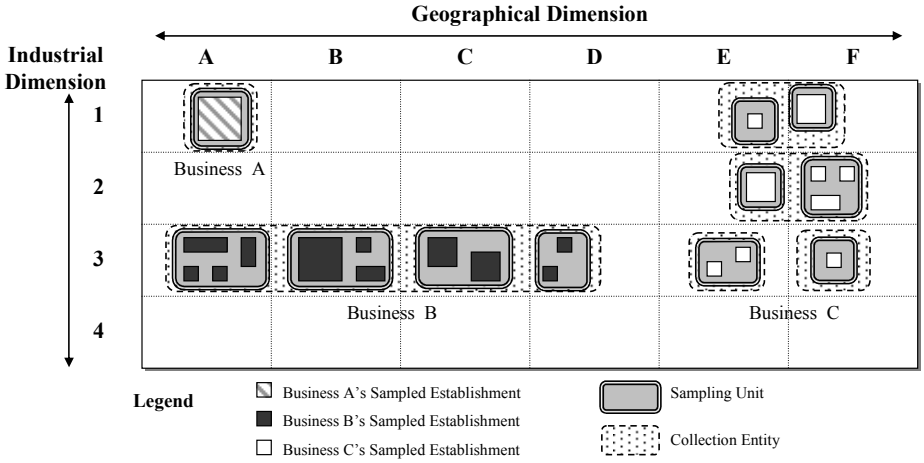


Figure 3 illustrates three examples of businesses.

– **Business A** has only one sampled establishment so this entity corresponds to the SU and the CE.
– **Business B** has 11 sampled establishments divided into 4 domains of interest (4 provinces but only one industrial class), 4 SUs and only one CE.
– **Business C** has 9 sampled establishments divided into 6 domains of interest (2 provinces and 3 industrial class), 6 SUs and 4 CEs.

For response management purposes, each business in the UES sample is assigned to one of four tiers depending on its size and its significance (Sear, Hughes and Lalande, 2007 ; Sear, Hughes, Vinette and Bozzato, 2007 ; Godbout, 2009).

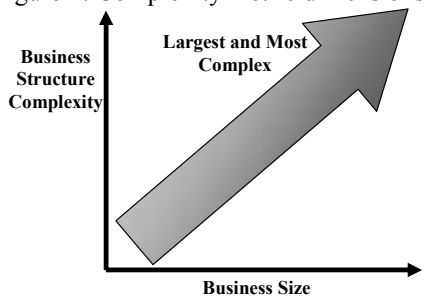
– **Tier I**: The largest, most complex businesses in Canada. Tier I is managed by the Enterprise Portfolio Management Program (EPMP), which updates the business profiling, manages the data collection and maintains proactive and ongoing relationships with them (for more details, see Enterprise Statistics Division, 2008).
– **Tier II**: Other businesses important for the UES but simpler or less complex than Tier I businesses.
– **Tier III**: Mid-size businesses carrying less individual significance than units in Tiers I and II. Tier III usually forms the bulk of the sample, and the sampled units (SUs) of Tier III businesses are considered "replaceable", in that there are many other SUs that represent the same homogeneous population as well as these ones.
– **Tier IV**: Small businesses usually excluded from the sample, because they have low impact on the estimates. Tax information is used for Tier IV businesses.

The Tier I businesses are selected by EPMP managers among the largest and most complex ones. Currently, there are roughly 350 businesses managed by the EPM program. Once the Tier I is set, the sample remaining portion is split into Tiers II, III and IV based on their importance to UES. The proposed methodological framework for the business profiling prioritization and the tier definition uses two main tools: a complexity metric and a priority flag, as described in sections 3 and 4.

3. Identifying the largest and most complex businesses

The proposal includes developing a metric to identify the largest and most complex businesses. This metric considers two dimensions: the business size (measured in revenue or in number of employees) and its structure complexity (based on the same size variable's distribution) as shown in figure 4.

Figure 4: Complexity metric dimensions



The proposed complexity measure for a business *i* has the general form of $\kappa_i = f(y_i, \eta_i)$ where y_i and η_i are respectively the business size and its structure complexity. A good and simple choice for $f(y_i, \eta_i)$ is the product of both dimensions.

$$\begin{aligned} \kappa_i &= f(y_i, \eta_i) \\ &= y_i \eta_i \end{aligned} \tag{1}$$

The complexity factor η_i is defined using the Shannon's entropy theory as shown in section 3.1.

3.1 Complexity factor and Shannon's entropy

The profiling and the data collection problems are very similar since we are in both cases interested in collecting a variable y_i and in its distribution in the business structure. The business structure can be represented in statistical terms using, for a partition P_i of a complex business *i*, a vector $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{in})$ with $0 \le p_{ik} \le 1$ for every *k* and $\sum_k p_{ik} = 1$ so we get for the size variable y_i its allocated size vector

$$\begin{aligned} \mathbf{y}_i &= (y_{i1}, y_{i2} \cdots y_{in}) \\ &= y_i (p_{i1}, p_{i2} \cdots p_{in}) \\ &= y_i \mathbf{p}_i \end{aligned} \tag{2}$$

For a business *i*, its profiling vector $\mathbf{p}_i$ is not fixed in time because of restructuring, changes in revenues, employment etc. The updated or realized allocation $\tilde{\mathbf{p}}_i$ of the variable y_i is expected to be close to $\mathbf{p}_i$ without being necessarily equal. The information theory could help studying the uncertainty related to the profiling vector, specially the Shannon's entropy theory.

The entropy (Shannon, 1948 ; Reesor and McLeish, 2002 ; Wikipedia, 2009) is a measure of the uncertainty associated with a random variable X having *n* possible values $x_1, x_2 \ldots x_n$ using a distance between 2 probability distributions. Shannon's entropy measures the minimal amount of information needed to complete a set without any losses. It's defined as the expected value of the logarithm of the inverse probabilities:

$$\begin{aligned} H(X) &= E\left[\log P^{-1}(X = x_k)\right] \\ &= \sum_{k=1}^{n} - p_k \log p_k \end{aligned} \tag{3}$$

with $p_k \log p_k \equiv 0$ when $p_k = 0$. We have $0 \le H(X) \le \log n$ with $H(X) = 0$ when X takes only one value with a probability of 1 ($P(X = x_i) = 1$) and $H(X) = \log n$ when all n possible values of X follow a uniform distribution ($P(X = x_i) = 1/n$).

If a random variable X is associated to a business partition P_i with n operating entities, and takes the value set $\mathbf{y}_i$ respectively with probabilities $\mathbf{p}_i$, $H(X)$ gives the amount of y_{ik}'s required to have a complete set of information on the business total y_i so Shannon's entropy becomes a good candidate for measuring a business structure complexity.

– For a business with a simple structure (only one establishment, i.e. $n = 1$), its entropy equals 0 ($H(X) = 0$).
– For a given number n of establishments within a business, the most complex structure would be for a uniformly distributed business with $\mathbf{p}_i = (1/n, 1/n, ..., 1/n)$ since every pieces of information we can get from the business let a maximum of uncertainty for its remaining portion. Its entropy would be maximized with $H(X) = \log n$.
– If a complex business contains many establishments but one represents a very large proportion of its size, the entropy would be very low since the uncertainty is highly reduced once we know the information related to the main piece of the business.

We propose to measure the complexity factor η_i for a business i and its partition P_i related to the size variable y_i as the following:

$$\begin{aligned} \eta_i &= H(Y \mid P_i) \\ &= \sum_{k \in P_i} - p_{ik} \log p_{ik} \end{aligned} \tag{4}$$

The proposed complexity metric κ_i for a business i, becomes:

$$\begin{aligned} \kappa_i &= y_i \eta_i \\ &= y_i \sum_{k \in P_i} - p_{ik} \log p_{ik} \end{aligned} \tag{5}$$

One may consider more than one partition of a given business whether we want to measure the structure complexity from the whole business or from its partition by industry, by geography or both or combine them by putting different weights on the respective complexity factors. A general complexity factor that combines the J entropy measures for the different partitions P_{ij} is proposed as the following:

$$\begin{aligned} \eta_i &= \sum_{j=1}^{J} \alpha_j H(Y \mid P_{ij}) \\ &= \sum_{j=1}^{J} \alpha_j \left(\sum_{k \in P_{ij}} - p_{ik} \log p_{ik} \right) \end{aligned} \tag{6}$$

with $0 < \alpha_j < 1$ and $\sum \alpha_j = 1$. The α_j's are selected to balance all complexity factors one want to consider.

3.2 Example

To illustrate the proposed complexity metric, the table 4 shows five examples of businesses having the same size but different structures.

Table 4: Examples of the complexity metric

Business i	**Business Size** y_i	**Size Partition at Establishment Level** p_{ik}					**Structure Complexity** $\eta_i = H(Y \mid P_i)$	**Complexity Metric** $\kappa_i = y_i \eta_i$
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$		
A	500	1.000					0.000	0
B	500	0.500	0.500				0.301	151
C	500	0.900	0.025	0.025	0.025	0.025	0.201	101
D	500	0.500	0.125	0.125	0.125	0.125	0.602	301
E	500	0.200	0.200	0.200	0.200	0.200	0.699	349

Here are some observations:

– **Business A** has only one establishment so its complexity factor and its complexity metric are 0.

– **Business B** has two establishments equally divided. Its complexity factor is $\eta_i = \log 2$.

– **Businesses C, D and E** have five establishments. However, C is heavily concentrated on one establishment so its complexity factor is highly reduced while E is uniformly distributed among its five establishments so $\eta_i = \log 5$.

To distinguish the Tiers II, III and IV, we need to identify the units with the highest importance in the response management. This is done in section 4.

4. Assigning a response management priority to units

The second portion of the methodological framework is to set a unit response management. For that, we proposed to base the priority rule on the way each SU represents its population according to the sampling design.

4.1 Response management priority at sampling unit level

A priority level is assigned to each SU according to the selection stratum.

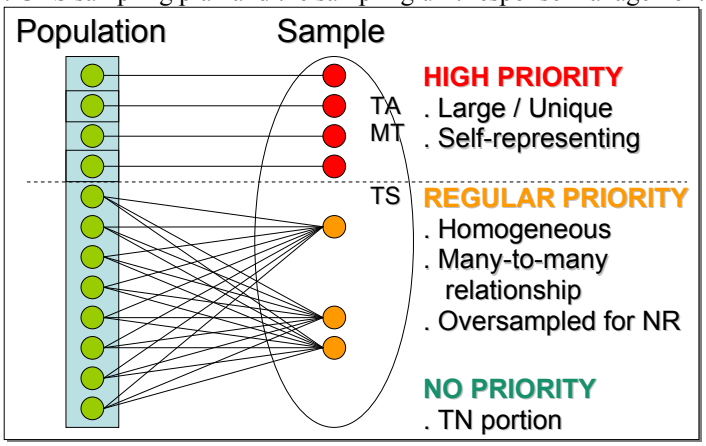
– **MT and TA** are assigned a high priority, since they are very large or unique, and they are self-representative in the sample.

– **TS** are assigned a regular priority, since they are in a randomly selected group representing a population that is assumed to be homogeneous. Therefore, they do not carry the same individual significance in the representation of the population as MT or TA do, especially since the sampling fraction of TS strata usually includes an offset for expected ignorable non-response.

–**TN** selected through network sampling are assigned a zero priority, since they have been selected for coherence purposes and to ease the reporting burden of some complex businesses and not necessarily to represent a part of the population.

The relationship between the UES sampling plan and the SU response management priority is illustrated in figure 5.

Figure 5: UES sampling plan and the sampling unit response management priority



4.2 Response management priority at collection entity level

The respondent completes a single questionnaire for the CE; therefore, assigning a priority to each CE is important for the response management. Some CEs include SUs with different priorities. The priority of a CE is defined as the highest priority of its SUs. Therefore, a high-priority CE has at least one high-priority SU. A regular-priority CE has at least one regular-priority SU, but no high-priority SUs. A zero-priority CE has zero-priority SUs only (no high or regular-priority SUs).

At the CEs creation step, a response management priority flag is attached to each, based on the survey design, and are available to all response managers through different BR tools.

5. Results for reference year 2008

The original plans were to implement the proposed methodology for the reference year 2008 as a pilot project but some operational constraints forced us to postpone the study. Nevertheless, most of the features from the proposal were tested on the 2008 UES data.

First, using the methodology described above, all the UES sampled businesses have been assigned a tier. The Tier I set was associated to the current EPM business list. All the non-Tier I businesses were assigned the remaining tiers II, III and IV based on the response management priority variable derived from the proposed model. The model and the distribution of all UES sampled businesses are given in table 6.

Table 6: Tier identification model and business counts

Tier	Identification Rule	Nb of Businesses
I	All business part of the EPM program	235
II	All non-Tier I businesses that have at least 1 high-priority SU	14,000
III	All non-Tier I & II businesses that have at least 1 regular-priority SU	33,000
IV	All the remaining businesses	200
Total	**All UES sampled businesses**	**47,435**

Also, all CEs have received a response management priority based on the highest level of priority among its SUs. Table 7 shows the distribution of all CEs by their response management priority and their business' tier.

Table 7: UES collection entities by response management priority and tier

Tier	CE Response Management Priority			Total
	High	Regular	No	
I	1,500	500	50	2,050
II	16,000	1,500	500	18,000
III	0	34,000	200	34,200
IV	0	0	250	250
Total	**17,500**	**36,000**	**1,000**	**54,500**

By definition, the 235 Tier I businesses selected by the UES are very large and complex. In average, each Tier I business has 10 CEs and 75% of these have a high priority. One can observe that 500 Tier I CEs have a regular priority. Even if they are critical for the UES program, the Tier I businesses might have some parts that are less important to surveys. This information is critical for a response manager when dealing with a reluctant respondent.

On the other hand, the number of Tier II businesses is overwhelming. There are 14,000 Tier II businesses with a low average of 1.3 CE each. Since they have at least one high priority CE, roughly 90% of their 18,000 CEs have a high priority. The response management cannot be approached the same way as the Tier I.

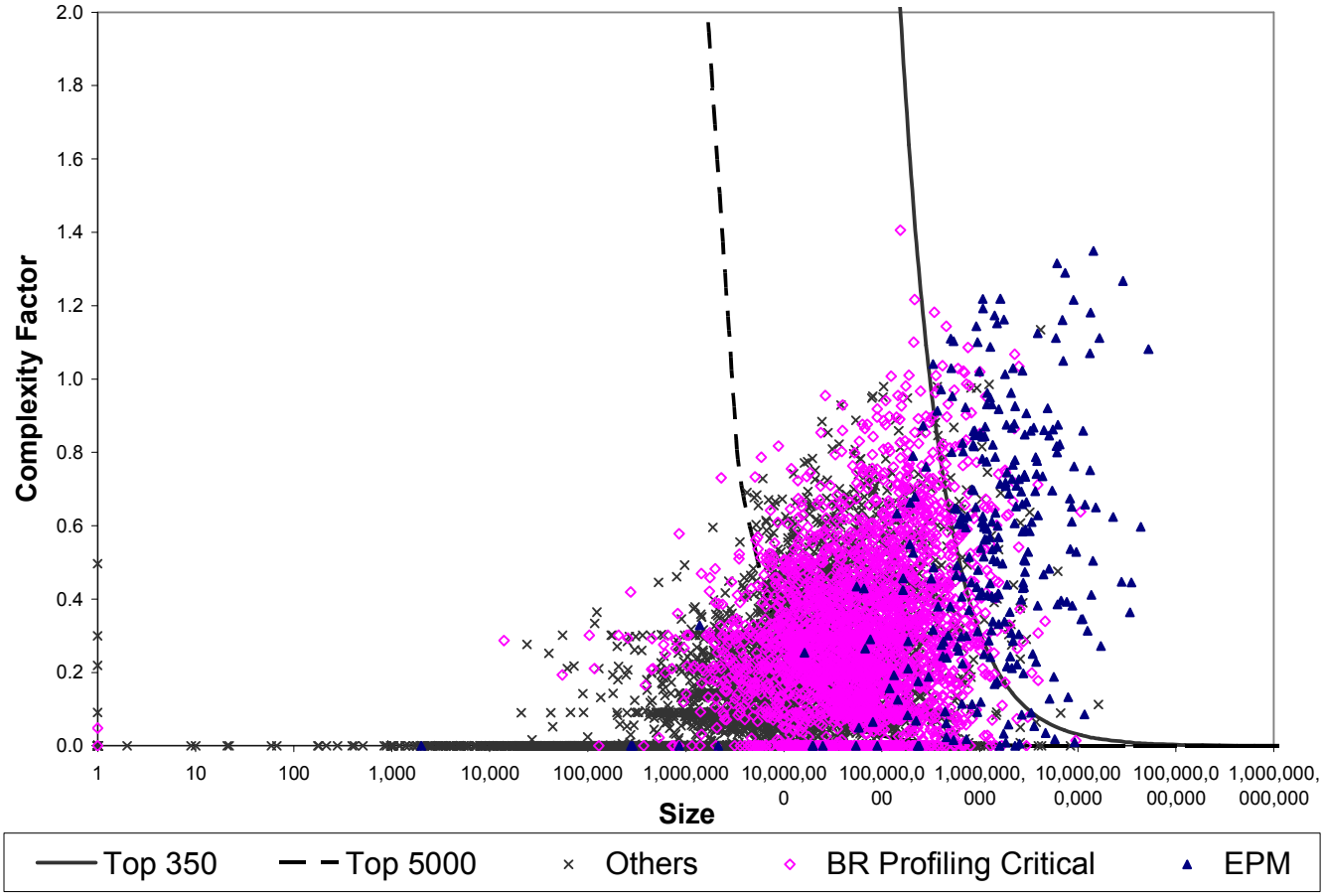
Second, the complexity metric was calculated on all complex businesses on the BR according to three overlapping partitions described in Table 8.

Table 8: Business structure partitions and weighting factors

Partition	Business Desegregation Level	Factor α_j
1	Establishment	0.3
2	4-Digit NAICS x Province/Territory	0.4
3	4-Digit NAICS	0.3

These three partitions were selected to assign a minimal complexity on the multi-establishment businesses but to put on top the multi-industry ones because they lead to more complex collection challenges. Figure 9 shows the distribution of a random subpopulation (to lighten the graph) of all the complex businesses by size and structure complexity according to their total revenue.

Figure 9: Complex businesses by size and structure complexity



All the green triangles represent a business from the current EPM program; the pink circles are the current BR profiling critical businesses while the blue squares are the other complex businesses. The continuous black line delimits the top 350 businesses according to the complexity metric and the dotted one delimits the top 5000 businesses. All the points on the horizontal axis represent complex businesses having all their revenue concentrated on only one establishment. One can observe that in most cases, the businesses from the EPM and the BR profiling critical programs are usually among the largest according to the complexity metric but the consistency could be improved. The table 10 gives the proportion of the businesses from both programs that are above their respective threshold (top 350 for the EPM businesses, top 5000 from the BR profiling critical businesses).

Table 10: Proportion of the complex businesses from both programs that are above their respective threshold

Class	Current EPM Businesses	Profiling Critical Businesses
Total	334	4669
Above Threshold - N	201	2498
Above Threshold - %	60.2	53.5

Note that in table 10, there are only 334 EPM businesses because some of the current Tier I businesses are no longer complex. Table 10 shows that roughly 60% of the EPM complex businesses and 55% of the BR Profiling Critical Businesses would respectively fall in the top 350 and the top 5000 complex ones. So, the complexity metric could help updating the lists of both groups.

6. Conclusion

In order to prioritize the businesses according to their size and their structure complexity, a metric was proposed. In addition, a response management priority flag was defined based on the sampling design. These tools can prevent managers from relying only on subjective criteria to identify important units for profiling and data collection.

The future plans are to allow the complexity metric to combine the measures calculated from different size variables (e.g. total revenue, number of employees, total assets etc.). Also, another feature is to integrate the response management priority and the complexity metric in order to allow measuring the complexity on a subset of the businesses that are important to a survey.

Finally, a study will try to see how the complexity metric might help identifying and managing the Tier I and the BR profiling critical businesses. For RY2009, the pilot will continue and the non-Tier I portion of the UES sample will be classified into the Tiers II, III and IV using the proposed framework. A study will be conducted after the 2009 survey cycle to measure the benefits of the proposed methodology on the response management.

7. Acknowledgments

The authors would like to acknowledge Claude Turmelle and Janet Hughes from Statistics Canada for their contribution to the project and the revisers for their relevant comments.

8. References

Beaucage, Y., Hunsberger, P. and Pursey, S. (2005), The Redesign of the Statistics Canada's Business Register, Proceedings of the 2005 Joint Statistical Meetings, Minneapolis (United States).

Brodeur, M. (2006) *et al*, The Integrated Approach to Economic Surveys in Canada, Statistics Canada, Ottawa (Canada). Catalogue 68-514.

Business Register Division (2009), A Brief Guide to the BR, Statistics Canada, Ottawa (Canada). http://www.statcan.gc.ca/imdb-bmdi/document/1105_D2_T1_V2-eng.pdf

Business Register Division (2007), An Introduction to Concepts, Statistics Canada, Ottawa (Canada).

Enterprise Statistics Division (2008), Unified Enterprise Survey Bulletin quarterly, Number 2, Fall 2008, Statistics Canada, Ottawa (Canada).

Godbout, S. (2009), Methodological Aspects of Holistic Response Management in the Unified Enterprise Survey, SIMP II: Economic Statistics Project, Document #36, Statistics Canada, Ottawa (Canada).

Lavallée P. and Hidiroglou M. (1988), On the Stratification of Skewed Populations, Survey Methodology. June 1988. Volume 14, no 1, Ottawa, Canada: Statistics Canada

Reesor, R.M. and McLeish, D.L. (2002), Risk, Entropy amd the Transformation of Distributions, Bank of Canada, Working Paper 2002-11.

Sear, J., Hughes, J. and Lalande, D. (2007), Controlling and Managing Response Burden in Business Surveys at Statistics Canada, SIMP II: Economic Statistics Project, Document #16, Statistics Canada, Ottawa (Canada).

Sear, J., Hughes, J., Vinette, L. and Bozzato, W. (2007), Holistic Response Management of Business Surveys at Statistics Canada, ICES-III, Montréal (Canada).

Shannon, C.E. (1948), A Mathematical Theory of Communication, The Bell System Technical Journal, Volume 27, pp. 379-423, 623-656, July, October 1948,

Simard, M., Girard, C., Parent, M.-N. and Smith, J. (2001), Sampling Designs for the Unified Enterprise Surveys: The Early Years, Working paper BSMD-2001-003E, Statistics Canada, Ottawa (Canada).

Wikipedia (2009), Entropy – Information theory, http://en.wikipedia.org/wiki/Information_entropy