

ESTIMATION FOR ANNUAL BUSINESS SURVEYS BASED ON TWO-PHASE NETWORK SAMPLING

Michelle Simard and Michel Hidiroglou¹

ABSTRACT

Business survey's frames will most likely contain *business entities* (members) that are of a complex nature. The complexity of the business entity is reflected via a hierarchy of levels that corresponds to its operating and financial representation. The sampling of a business universe may occur at any of these levels. At Statistics Canada, a business entity can be sampled at any of four nested levels. These levels, arranged in order of importance are the *enterprise* (being the highest level), the company, the *establishment*, and the location (being the smallest identifiable level of a business entity). The linkage between enterprise and establishments is either one to one, or one to many. It has usually been the practice to collect the required data from the sampling units at the same or higher level than the target statistical units. If the associated sampling mechanism is simple, i.e. simple stratified random sampling (SRS), then the associated estimation process is straightforward. However, if the sampling occurs at lower levels than the target universe, then there is an additional complexity in terms of computing the associated inclusion probabilities. This type of sampling is known as *network sampling* (Sirken 1970). Thompson (1992) has extensively studied it, and termed it *adaptive cluster sampling*. It is this procedure that was used to select the first-phase sample of a new annual survey known as '*Unified Enterprise Survey*' (UES). The sampling, estimation and variance estimation procedures are discussed in this paper.

KEY WORDS: Two-phase sampling; Network sampling; Adaptive cluster sampling; Auxiliary information; Calibration method; Weight-share method; Post-stratification estimator; Regression estimator; Variance estimation.

RÉSUMÉ

Les bases de sondages des enquêtes économiques sont en général de nature complexe. La complexité est due au fait qu'il existe une hiérarchie de niveaux qui correspondent à différentes représentations de la même entité, i.e. entité de production, entité financière, etc. L'échantillonnage peut ainsi être effectué à n'importe lequel de ces niveaux. À Statistique Canada, il existe quatre niveaux possibles, i.e. quatre unités d'échantillonnage possibles, pour effectuer l'échantillonnage des enquêtes économiques. Ils sont, par ordre d'importance: l'entreprise (le niveau le plus élevé), la compagnie, l'établissement, et l'emplacement (le plus petit niveau identifiable). Les liens possibles entre l'entreprise et le(s) établissements sont soit un pour un (structure simple) ou un pour plusieurs (structure complexe). Il est monnaie courante d'obtenir les données requises de l'unité d'échantillonnage au même niveau ou plus haut que l'unité statistique cible. Si le mécanisme d'échantillonnage est simple, i.e. *échantillonnage aléatoire simple* (EAS), alors le processus d'estimation associé est simple. Cependant, si l'échantillonnage est effectué à un niveau inférieur que l'unité statistique cible, alors une complexité supplémentaire est introduite quant au calcul des probabilités d'inclusion. Ce type d'échantillonnage est connu sous le nom d'*échantillonnage par réseau* (Sirken 1970). Thompson (1992) l'a étudié largement, et l'a appelé *échantillonnage par grappe adapté*. Cette procédure a été utilisée pour sélectionner la première phase d'échantillon d'une nouvelle enquête annuelle l'*Enquête Unifiée sur les Entreprises* (EUE). L'échantillonnage ainsi que les procédures d'estimation et d'estimation de variance sont discutées dans cet article.

MOTS CLÉS : Échantillonnage à deux phases; échantillonnage par réseau; échantillonnage par grappe adapté; information auxiliaire; méthode de calage; méthode du partage des poids; post-stratification; estimateur GREG; estimation de variance.

¹ Michelle Simard and Michel Hidiroglou, Business Survey Methods Division, Statistics Canada, Ottawa (Ontario), Canada K1A 0T6.

1. INTRODUCTION

The main objective of most surveys are to produce reliable estimates of the population characteristics for a *single* targeted statistical level, i.e. the individual members of the population whose characteristics are to be measured. The frame is then delineated into these target statistical levels (units). At Statistics Canada, a business entity can be sampled at any of four nested levels. These statistical levels, arranged in order of importance are the *enterprise* (being the highest level), the company, the *establishment*, and the location (being the smallest identifiable level of a business entity). Data are collected from these units to produce estimates of interest at the appropriate level. For some business surveys the objectives are multi-purpose and more than one statistical level needs to be estimated within the same survey. The target statistical units are linked on the frame across the different levels, and the resulting sampling needs to consider this. This is the case of the Unified Enterprise Survey (UES), conducted at Statistics Canada. Two of the objectives of this multi-purpose survey are to: 1) ensure coherence analysis, i.e. to compare and to reconcile the enterprise level estimates with its corresponding establishments' estimates; and 2) reduce response burden. The first objective was satisfied by using adaptive cluster sampling. That is, if an enterprise had been selected into the sample, then all corresponding establishments were also selected. The second objective was achieved by using auxiliary data with survey data for the simple structure statistical units in a two-phase approach. The simple structure being defined as an enterprise having only one sampling unit.

The first-phase sample consisted of all simple establishments selected in the adaptive sampling cluster. The second-phase sample is a sub-sample of the first. Information for this second phase is obtained by direct survey. Tax information is available for all selected first-phase units. It is used to improve the efficiency of the estimation process. This resulted in a sampling design whose estimation process is more complicated than the one used for traditional business surveys.

The paper is structured as follows. The objectives of the UES, the Business Register (Statistics Canada' business frame) are described in Sections 2 and 3. A description of the sample design process as well as the data acquisition strategy is next provided in Section 4. Two estimators for adaptive cluster sampling are

proposed in Section 5. We provide the associated estimated variance expressions in Section 6.

2. PROJECT TO IMPROVE PROVINCIAL ESTIMATES AND THE UNIFIED ENTERPRISE SURVEY

The Project to Improve Provincial Estimates Statistics (PIPES) is one of the most important projects at Statistics Canada in recent years. In 1996, three Canadian provinces signed an agreement with the government of Canada to harmonize their provincial sales taxes with the national sales tax (the Goods and Services Taxes (GST)). The Canadian government collects the Harmonized Sales Tax (HST) during the whole year. At the end of each fiscal year, the tax is reallocated to each province with its appropriate share. Statistics Canada obtained the mandate to produce reliable provincial estimates, as to be able to derive provincial share based on the provincial HST allocation formula. The HST allocation formula is complex; it is based on all money transactions, i.e. salary and wages, expenses, revenues, sales, etc, occurring during a year within and between provinces.

Previously, business surveys conducted at Statistics Canada did not allow the production of detailed provincial estimates. There were no common methodologies for business surveys, as they had different frames, reference periods, designs and estimation procedures. There was neither integration nor co-ordination between them. Furthermore, some industries had not been surveyed for several years. Conversely, some multi-industrial businesses were surveyed several times during the year. There was, therefore, a need to redesign annual business surveys.

Statistics Canada has been carrying out a major redesign of its annual business surveys. Some of the improvements included: (i) increased sample size for some of the on-going surveys; (ii) the use of a single business frame, i.e. the Business Register (BR); (iii) restructuring the System of National Accounts (SNA); and (iv) implementation of a new annual business survey that ultimately would become the vehicle for producing annual estimates for all industries at the required details.

The UES was created for producing reliable provincial estimate and integrating annual survey methods. Its objectives are to produce financial information as well as production information for the business universe. The first cycle of the UES was first carried out in 1997. Seven sectors of the Canadian industries that

had not been surveyed for quite some time were included in the pilot survey. These seven pilot industries were Aquaculture, Taxis and Limousines, Couriers and Messengers, Real Estate Agents and Brokers, Lessors, Construction and Food Services. *Establishment data*, i.e. the production level, was required for those seven pilot industries. *Enterprise data* was also required for a major financial survey, Industrial Organisation & Finance. These financial data were only available at the enterprise level. The financial survey is an economy-wide survey covering all industries, including the seven, of the incorporated businesses in Canada. These two statistical levels were integrated within UES for 1997. The plan is to gradually integrate on-going annual surveys into the UES platform. More details can be found in Laniel and Royce (1998).

3. BUSINESS REGISTER

The Business Register (BR) was re-designed in the mid-eighties. It is a complex system based on information provided by administrative files. It covers the universe of Canadian businesses, employers as well as non-employers. It contains information about each of the businesses as well as links to other statistical or administrative levels. Some of the available information on the BR includes: provincial and industrial coding, size variables (revenue and number of employees), address, legal name, unique statistical identifier, Business Number (BN) for linking the business entity to the administrative files, birth and death dates, complexity structure indicators, coverage information, etc. Any one of the four levels of the statistical entities (enterprise, company, establishment and location) residing on the BR can be used as sampling units. The creation of the statistical units or entity level is based on a set of standard rules and criteria. The first two levels are often used when the objective of the survey is to produce financial information, i.e. consolidated revenues and expenses, as well as the financial statement of the business, to name a few. The latter two are most often used when the objective of the survey is to produce operational or production type of information, i.e. type of activity, salary, number of widget produced, etc. An important unit for UES is the legal entity structure. This level is defined by Revenue Canada (in collaboration with Statistics Canada) and it is derived, based on the legal structure in which the business defines and operates itself. The legal entity is central in the context of UES for the data acquisition and estimation strategy because it is at this level that the auxiliary informations, (i.e. the tax data records) are available.

The two statistical levels used in UES are the enterprise and the establishment. The establishment level is defined as a physical production entity operating in one province and in one industry. Its industrial code is assigned using the North American Industrial Classification System (NAICS), at the 6-digits level. The enterprise is defined as the administrative entity managing the establishment(s). It is at this level that financial statements are produced. The enterprise does not have a provincial classification assigned to it. However, it is assigned the dominant industrial classification of its corresponding establishments. The dominance rules are based on a standard algorithm based on the revenue variable.

To simplify matters, the BR can be viewed as being delimited into two portions: the Integrated Portion (IP) and the Non-Integrated portion (NIP). The NIP contains units with a simple structures and small revenues. All statistical levels of a business entity within this portion (NIP) are the same. For those businesses, one legal entity covers one enterprise, with one establishment. Such units are updated automatically by the administrative files. Figure 1 shows this type of structure. The IP portion contains the more complex structured business types and some large simple businesses. It can be defined as consisting of business entities that have multi-establishments, are multi-provincial or multi-legal entities. Figure 2 displays these units. For more details, see Castonguay (1998).

Figure 1: Simple Structured Enterprise
Legal entity ↔ *Statistical entities*

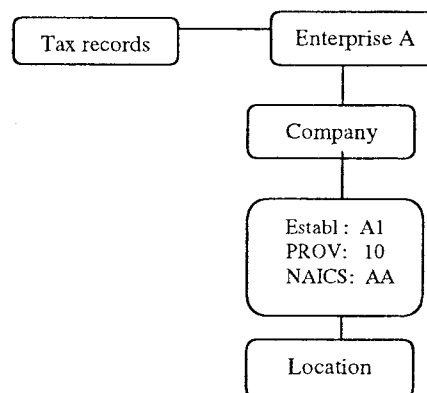
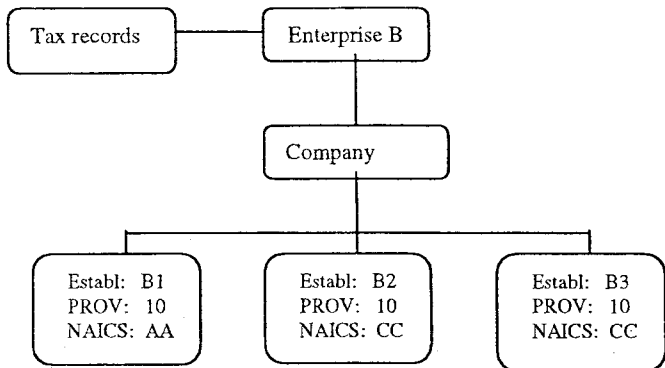


Figure 2: Complex Structured Enterprise

Legal entity ↔ Statistical entities

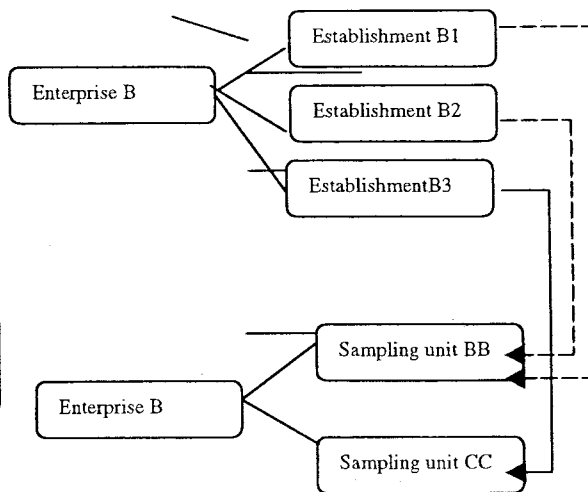


The IP units are updated by profiling or by survey feedback. The BN is used to link the legal entity to the statistical entities. The BR assigns statistical numbers to the statistical entities. These numbers are also used for linkage within the BR statistical entities. As Revenue Canada initially introduced the BN in 1994, not all units on the BR have been linked at the legal structure. The 1997 UES in-scope population is made up of about 1 200 000 records. The IP contains some 100 000 of these records and the remaining are NIP records.

4. SAMPLING UES 1997

The enterprise and the establishments were the only two statistical levels considered for sampling purposes for UES 1997. Establishments within an enterprise were first regrouped into clusters operating within the same 6-digits NAICS and province. These clusters, referred to as *establishment clusters*, are used as sampling units for UES. The following definitions of a *complex* and *simple* enterprise are also used for UES sampling purposes. An enterprise is defined as complex, if there is more than one establishment cluster, or if there is more than one tax data record linked to it. An enterprise is defined as simple, if it has a one-to-one relationship with the associated tax data record and its establishment cluster. An example is given in Figure 3, using the enterprise structure given in Figure 2. Enterprise B is linked to three establishments B1, B2 and B3. Since two of the establishments, B1 and B2, operate within the same 6-digits NAICS and province, they belong to a single establishment cluster (sampling unit) BB. Similarly B3 yields the establishment cluster CC.

Figure 3: Establishments Clusters



4.1 Sample design

A two-phase sample design was chosen to reduce response burden. The first-phase sample was mainly used to obtain data at the enterprise level and for the establishments belonging to a complex enterprise. The second-phase sample mainly provided data at the establishment level for the simple enterprises.

4.1.1 Phase 1

The targeted level of reliability or coefficient of variation (c.v.) was set at 5 % for each block of 6-digits NAICS and province (referred to as cell). These two dimensions, i.e. provincial and industrial, represented the first two stratification variables. This level of reliability was based on the revenue variable. It used the *establishment cluster* as the sampling unit. Each cell was further stratified into size groups (take-all, large take-some, small take-some) using the Lavallée-Hidiroglou (1988) algorithm. This algorithm simultaneously yields the optimal stratification boundaries and the required sample size, n_h , within the size strata. The sampling units (establishment cluster) within each stratum were assigned a permanent random number (PRN). After sorting units by their PRN, the first n_h PRN within each size strata were selected. It is equivalent to simple random sampling (SRS).

Recall that one of the objectives of the UES is to ensure data coherence between the enterprise and its associated establishments. This objective was achieved as follows: If any of the enterprises had at least one of its establishment clusters included in the sample, then the enterprise was selected in the sample of enterprise. In addition, all the associated establishment clusters were included in the sample of

establishments' clusters. This type of sampling, referred to as network sampling, is well documented in Thompson (1992), and in Thompson and Seber (1996). Using the example given in Figure 3, suppose that, initially, only sampling unit CC had been selected into the sample (Figure 4). Network sampling resulted in having units BB, CC and the associated enterprise B included in the sample (Figure 5).

Figure 4: Before network sampling

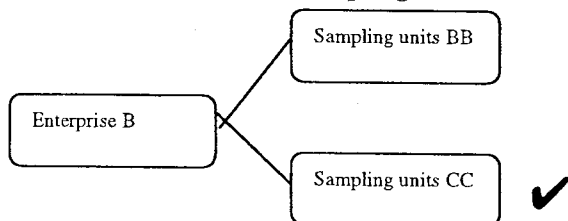
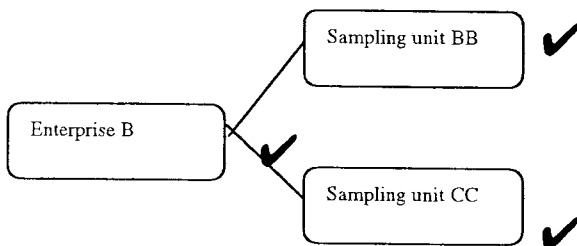


Figure 5: After network sampling



The impacts of applying network sampling for UES 1997 are provided in Figure 6.

Data acquisition for the first-phase sampled units was as follows. Units that were surveyed directly by questionnaire included: (i) selected complex enterprises and their associated establishment clusters; and (ii) simple establishments selected in the take-all stratum. Tax data was obtained for selected simple enterprises and their associated single establishment cluster in the two other strata. This was done by linking their BN on the BR to their corresponding tax data record. These simple structures were selected using a two-phase selection process.

Figure 6 : Establishments counts before and after network sampling

	Type of units	Before	After	Population
Complex	Take-all	3 730	4 411	5 727
	Take-some	195	218	
	Total	3 925	4 629	
Simple	Take-all	9 264	9 264	9 264
	Take-some	2 386	2 386	265 284
	Total	11 650	11 650	274 548
Total		15 575	16 279	280 275

Note: a. Adaptive cluster sampling did not significantly increase the overall sample size. The reason is that most establishments clusters had been selected as take-all.

b. The number of selected take-some complex is very small compared to the total sample size.

4.1.2 Phase 2

A two-phase procedure was implemented to reduce response burden for small and medium simple enterprises. The second phase was designed to produce reliable estimates at a more aggregated industrial level than the first phase. The c.v. based on the revenue variable, was set to 15 % for each grouping of 6-digits NAICS and province (referred to as *block*). The subsequent estimation (regression via the GREG estimator) uses tax data obtained in the first-phase. The GREG estimator yields a more efficient estimator than strictly weighting up the second-phase. For the second phase, a fourth level of stratification was added: the structure of the enterprise. Since tax data records are only available at the enterprise level; it is only for simple units that survey data are available at the same level as tax data. For this reason, only the simple units selected in the first phase were subject to the second-phase estimation process. Note that complex establishments² are included in the target c.v. for the second phase. Their associated second-phase sample is a complete overlap of the first-phase. Each block was separated into two groups, one for simple and one for complex. For the blocks of complex units, the size

² Complex establishments are establishments that belong to a complex enterprise. The same comment applies to simple establishments.

stratification was not considered, as the sampling fractions are always one.

Each block of simple units was stratified into two size groups (large take-some, small take-some) using a mixture of the first-phase size groupings. Second-phase sample sizes (m_g) were determined, proportional to the square root of the revenue variable within each size strata. The establishments within each stratum were assigned the same PRN as the first phase to ensure overlap. After sorting units by their PRN, the first m_g PRNs were selected within each size stratum. This procedure is equivalent to simple random sampling (SRS).

Data acquisition for the second-phase was as follows: Each selected simple unit was directly surveyed by questionnaires. Units that have both survey and tax variables were used to estimate the regression vectors required by GREG.

5. ESTIMATION

As seen in the sample selection section, the estimation procedure is more difficult for the first-phase units where enterprises and some complex establishments were selected indirectly. For the enterprise, the selection was done via the establishments. This type of selection corresponds exactly to the adaptive selection procedure described in Thompson (1992) and Thompson and Seber (1996) or Lavallée (1995). As in the sampling process, the weighting and estimation can be partitioned into two estimation procedures: 1) A one-phase approach for the complex establishments and the enterprise level statistics; and 2) A two-phase approach for the simple establishments level statistics. These two estimation procedures are described in Sections 5.1 and 5.2.

The following notation will be used:

- Y : (y_1, y_2, \dots, y_n) the variable of interest or survey data for unit receiving a questionnaire;
- X_1 : auxiliary information available at the population level used for the post-stratification adjustments;
- X_2 : ($x_{21}, x_{22}, \dots, x_{2p}$): auxiliary information available for first-phase sample unit used to calibrate the second-phase units;
- N_h : total number of establishment cluster in stratum h ;
- n_h : number of establishment clusters selected in stratum h ;

- N_I : total number of distinct enterprises in population;
- n_I : number of distinct enterprises selected in sample;
- δ_{hi} : indicator value; = 1, if some of the selected establishment cluster within an enterprise i belong to stratum h , = 0 otherwise;
- L_i : total number of links of an establishment cluster to enterprise i ;
- M_i : total number of establishment clusters in enterprise i ;
- m_i : number of selected establishment clusters in enterprise i ;
- h : stratification of phase 1;
- g : stratification of phase 2;
- i : enterprise ($i = 1, 2, \dots, N_I$);
- j : establishment cluster within an enterprise ($j = 1, 2, \dots, M_i$);
- k : entity k is the enterprise and all its establishments
- d : domain of interest;
- mg : calibration group or model group;
- a_k : calibration adjustment factor, which used the auxiliary information available;
- w_k : inverse of the selection probability π_k^{-1}

5.1 Estimation for first-phase units: enterprise level and complex establishments

Two estimators are proposed for that portion of the universe that was sampled via network sampling. Section 5.1.1 describes the first one, which uses the initial probabilities of selection for network sampling. Section 5.1.2 presents the second, which uses the expected number of initial probabilities as used in the weight-share approach described by Lavallée (1995).

5.1.1 Estimation using the initial probabilities of selection.

The UES enterprise sample selection is similar to adaptive cluster selection. The probability of selection of a given enterprise k is:

$$\pi_k^* = 1 - \prod_{h=1}^L \left(\frac{N_h}{n_h} \right)^{-1} \left(\frac{N_h - m_{hk}}{n_h} \right), \text{ where } m_{hk} \text{ is the}$$

number of clusters that belongs to enterprise k in stratum h . In UES, m_{hk} is always equal to one. Hence

$$\pi_k^* = 1 - \prod_{j \in \text{complex}} (1 - \pi_j) \text{ with } \pi_j = \frac{n_h}{N_h}.$$

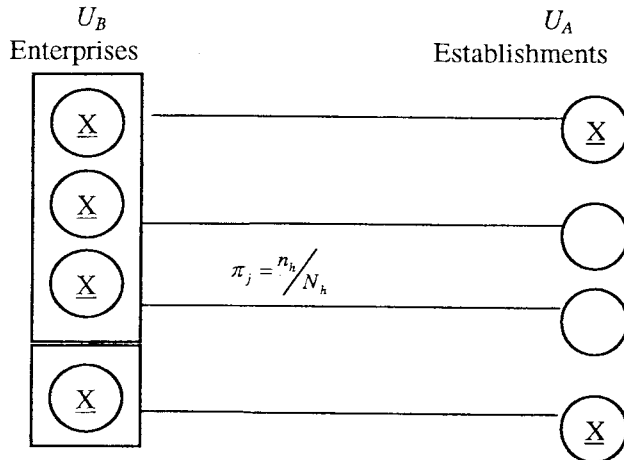
The resulting design weight $w_k^* = \pi_k^{*-1}$ is assigned to the selected establishment cluster and enterprise pair. All establishments within each selected cluster have the same design weight as the enterprise. The design

weight can then be used in a design consistent estimator such as Horvitz-Thompson type estimator $\hat{Y}_{HT} = \sum_S w_k^* y_k$ or calibration-type estimators: $\hat{Y}_{aux} = \sum_S w_k^* a_k y_k$. Note that both types are approximately unbiased estimators.

5.1.2 Estimation using the expected number of probabilities (weight-share method)

The second estimator proposed uses the weight-share method. This technique is often used in longitudinal surveys. It is used when population of interest needs to be sampled via a frame, which refers to a different population, but that is linked to it. Figure 7 illustrates this case for the UES. Here, the frame of sampling units is the establishment cluster in U_A . However, the population of interest is the universe of enterprises U_B .

Figure 7: Linkage between sampled Enterprise and its sampled Establishments



Following Lavallée (1995), the weight-share estimator

$$\text{is } \hat{Y}_{ws} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{N_h}{n_h} Y_{hi}^* \text{ where}$$

$$Y_{hi}^* = \frac{Y_i}{M_i} \delta_{hi}, \text{ if the } i^{\text{th}} \text{ enterprise intersects with stratum } h.$$

5.2 Estimation for simple units

For the simple enterprise, since $j=1$, the inclusion of probability of unit k is: $\pi_k^* = \pi_j = \frac{n_h}{N_h}$. The estimation can then be done using a standard approach. The simple units are estimated using a two-phase estimator, as given in section 4.1.1, using two different calibration techniques, one for each phase.

Thus, $\hat{Y}_{simp} = \sum_{k \in \text{sample}} w_{1k} a_{1mg1} w_{2k} a_{2mg2} y_k$, with a_1 being

the calibration factor based on the post-stratification for the calibration group $mg1$ (post-stratum) at the first-phase. In addition, a_2 , is the calibration factor based on the regression estimator for model group $mg2$ at the second-phase. It can be rewritten as:

$$\hat{Y}_{simp} = \hat{Y}_{ph2} + \hat{\beta}_{ph2} (\hat{X}_{simp} - \hat{X}_{ph2}), \text{ with}$$

$$\hat{Y}_{ph2} = \sum_{k \in S_2} w_{2k} y_k, \quad \hat{X}_{ph2} = \sum_{k \in S_2} w_{2k} x_k,$$

$$\hat{X}_{simp} = \sum_{k \in S_1} w_{1k} a_{1k} x_k = \sum_{k \in S_1} \frac{N_h}{n_h} \frac{N_{mg1}}{\hat{N}_{mg1}} x_k,$$

$$\hat{N}_{mg1} = \sum_{i \in mg1} w_{ik}.$$

Similarly, for the second estimator, the computation can also be simplified since $M_i = 1$. Hence,

$$Y_{hi}^* = \frac{Y_i}{M_i} \delta_{hi} = Y_i. \text{ This results in the standard}$$

estimation technique as described above.

More details about the estimation procedures can be found in Simard and Lanier (1998)

6. VARIANCE ESTIMATION

6.1 Variance estimation using initial probabilities of selection.

To obtain the variance estimate for \hat{Y}_{aux} as given in section 5.1.1, the joint probabilities of selection need to be computed. Let π_{ij} be the joint probabilities for selecting clusters i and j in the first-phase sample.

$$\pi_{ij} = 1 - P(\text{neither } i \text{ and } j \text{ are part of the sample})$$

$$= 1 - P(A \cup B)$$

$$= 1 - (P(\bar{A}) + P(\bar{B}) - P(\overline{A \cap B}))$$

$$= 1 - P(\bar{A}) - P(\bar{B}) + P(A \cap B)$$

$$= 1 - (1 - \pi_A) - (1 - \pi_B) + \prod_{h=1}^L \left(\frac{N_h}{n_h} \right)^{-1} \binom{N_h - 2}{n_h}$$

This can be used in the general calibration form as described in Deville and Särndal (1992), that is

$$v(\hat{Y}) = \sum \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} (g_i e_i) (g_j e_j)$$

These results give an approximately unbiased estimator of the variance. However, no existing software can easily compute the variance approximation in this context.

6.2 Variance estimation from the weight-share method.

As presented in Lavallée (1995) and in Birnbaum and Sirken (1965), the variance is straightforward and given by:

$$v(\hat{Y}) = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{N^2}{n_h} (1 - f_h) \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi}^* - \bar{Y}_h^*)^2$$

Note that Y_{hi}^* will be repeated as many times as the i^{th} enterprise intersects with stratum h .

7. VARIANCE ESTIMATION IN THE CONTEXT OF UES

Comparing results from sections 5 and 6, the weight-share estimator is unbiased and its variance computation is straightforward. The network estimator is approximately unbiased and its variance estimation computationally cumbersome. Logically the first one seems to be the best option. However, in the operational context of UES, since there was no existing generalized software ready to be used for a two-phase approach, it was not feasible to use the weight-share approach.

Another variance approximation was used for UES-97, based on the following findings:

- 1) As described in Sections 5.1.1 and 5.2, the weighting and estimation can be completed using w_k^* . However, no generalized estimation system could be used. However, for UES, the two-phase prototype program of Statistics Canada's Generalized Estimation System (GES) was modified to take into account different design weight for units within the same stratum. For more details on the prototype programs, read Arcaro (1997).
- 2) The assumption of independent selection between strata no longer holds with network sampling. For each complex enterprise, a network is created by its establishment cluster that belongs to a unique stratum (by design). Once one cluster is selected, other clusters in other stratum are forced into the sample.
- 3) As soon as one cluster is defined as take-all in one stratum, the sampling probabilities of the entity become one, i.e., a take-all entity as well. Most of the complex units can be found in this category as seen in Figure 6. Only 218 establishments out of the 16 279 units in the sample are take-some complex. The variance estimation is

computationally cumbersome from these units only.

- 4) The resulting sample can be viewed as an unequal probability sample, in which within a given stratum where there are complex units, their probabilities (of the take-some complex) are unequal to the other units as shown in Figure 8. Most units in the sample units are simples (71%), which the selection is a SRS and most selected units are take-all units (84%).
- 5) The error in the variance approximation in not using the pure variance formula is dependent on the number of take-some complex enterprise and on the length of network. The number of strata (or cluster) touched by one enterprise defines what is referred to as its length. Figure 9 shows the different lengths of network in the UES-97 sample.

Figure 8: Network (length =2)

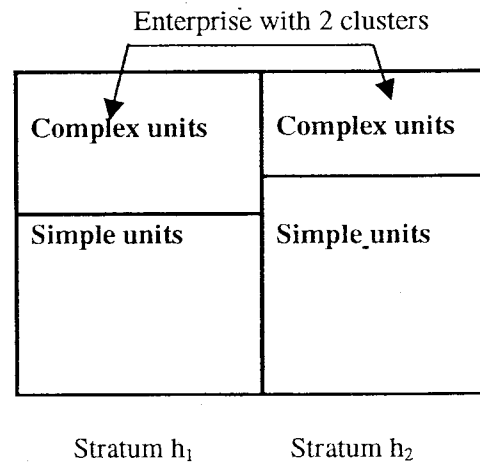


Figure 9: Length of the network for complex enterprise in the pilot industries

Length	1	2	3	4	5	6
Frequency	425	199	56	33	12	13
Length	8	9	10	11	15	16
Frequency	3	2	3	1	1	1

Note: The frequency represents the number of enterprise. The 425 enterprises with network of length 1 are multi-legal enterprises with one single sampling unit, read section 4 for the definition.

There are 756 enterprises: 331 have network longer than 1; 57 overlap with more than one

pilot industry; and 274 are multi-provinces that do not overlap with any pilot industries.

The methodology used for UES-97 was as follows: Estimates were produced with the exact probability of selection using initial probabilities as described in section 5.1.1 for complex and 5.2 for simple units. Variance estimation uses the original stratification so that existing generalized software could be used. The assumptions were: (i) The small number of take-some complex units implied that they represented only a small part of the total variance estimates; and (ii) Most of the network lengths were small. It was based on those assumptions that it was concluded that not using the network variance would have a minimal impact on the variance computations. The advantages and disadvantages of both estimators are summarised in Table 2.

Table 2: Comparison of the two estimators

	<i>Estimation</i>	<i>Variance</i>
Estimator 1	+ use generalized method + modified existing software + easy to explain to user	+ use existing software if assumption holds If not: no existing software; and computationally heavy
Estimator 2	+ use generalized method - No existing software - Difficult to explain to users	+ Straightforward + no condition required - no existing software

ACKNOWLEDGEMENTS

The authors would like to thank David Binder for his help and support, Normand Laniel for his ideas, Pierre Lavallée for his explanation of network sampling with the weight-share approach and Claude Girard for producing the tables.

REFERENCES

Arcaro, C. (1997). Specification for the two-phase prototype estimation system. Statistics Canada Internal document.

Birnbaum, Z.W. and Sirken, M.G. (1965). Design of sample surveys to estimate the prevalence of rare disease: three unbiased estimates. Vital and Health statistic, ser. 2, no. 11, Washington, DC: Government Printing Office.

Castonguay, E. (1998). Le Régistre des entreprises à Statistique Canada. *SSC Proceedings of the Survey Methods Section*, pp 65-69.

Deville, J.-C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, no 87, pp 376-382.

Laniel, N. and Royce, D. (1998). Projet d'amélioration des statistiques économiques provinciales: Objectifs et Enquête-pilote. *SSC Proceedings of the Survey Methods Section*, pp 59-63.

Lavallée, P. and Hidiroglou, M.A. (1988). On the stratification of skewed population. *Survey Methodology*, no.14, pp 33-45

Lavallée, P. (1995). Cross-sectional weighting of longitudinal survey of individuals and households using the weight-share method. *Survey Methodology*, no.21, pp 25-32.

Simard, M. and Laniel, N. (1998). Échantillonnage et Estimation pour l'enquête unifiée sur les entreprises. *SSC Proceedings of the Survey methods Section*, pp 77-82.

Thompson, S. K.(1992). *Sampling*. New York, John Wiley and Sons.

Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling*. John Wiley and Sons.