

L'ÉCHANTILLONNAGE AVEC UNE APPROCHE UNIFIÉE : LE CAS DE L'ENQUÊTE UNIFIÉE AUPRÈS DES ENTREPRISES

M.-N. Parent et M. Simard¹

RÉSUMÉ

La nouvelle Enquête unifiée sur les entreprises (EUE) intègre différentes enquêtes annuelles de Statistique Canada. Certains concepts et méthodes employés pour l'échantillonnage de l'enquête et les principes qui sont demeurés sensiblement les mêmes depuis le début de l'EUE ont été définis. Ceci inclut l'utilisation d'une seule base de sondage, une approche à deux phases, la stratification, l'instauration des numéros aléatoires communs, la répartition et la sélection de l'échantillon. De plus, une nouvelle méthode de détermination du coefficient de variation pour enquête à buts multiples a été développée. Finalement, les stratégies à adopter envers les différentes demandes des industries en termes de précision ont été préparées.

MOTS CLÉS : Unités d'échantillonnage; cellule; méthode itérative du quotient.

ABSTRACT

The new Unified Enterprise Survey (UES) integrates many Statistics Canada's annual surveys. Some survey concepts and methods applied in the different sampling cycles remain essentially the same from year to year. These include the use of a single frame, a two-phase approach, the stratification algorithm, the use of common random numbers, the sample allocation and selection. Furthermore, a new method of determining coefficients of variation for multi-purpose surveys has been developed. Finally, strategies to meet different industry specifications in term of precision have been prepared.

KEY WORDS: Sampling unit; Cell; Raking ratio.

1. INTRODUCTION

Le Projet d'amélioration des statistiques entreprises provinciales (PASEP) débute sa quatrième année. Ce projet a débuté en 1996, au moment où trois provinces signaient un accord avec le gouvernement canadien pour permettre l'harmonisation de la collecte de leur taxe de ventes provinciales avec la taxe nationale (la Taxe sur les produits et services (TPS)). Le gouvernement canadien collecte cette Taxe de vente harmonisée (TVH) pendant l'année et à la fin de chaque année fiscale, les revenus de la taxe sont réalloués à chaque province selon sa part correspondante. Statistique Canada a obtenu le mandat de produire des estimations provinciales

fiables pour permettre de dériver les parts de chacune des provinces.

Depuis, Statistique Canada est en pleine restructuration de ces programmes économiques annuels. Une des améliorations majeures est la mise en place d'une nouvelle enquête annuelle qui deviendra le véhicule ultime pour produire les estimations annuelles au niveau requis pour tous les secteurs industriels. Cette enquête, l'Enquête unifiée sur les entreprises (EUE), a été développée selon ces besoins. Un des objectifs de cette enquête est d'intégrer la plupart des enquêtes annuelles à partir d'une même plate-forme, en utilisant des méthodes et systèmes normalisés en plus de produire des statistiques provinciales fiables. Il existe d'autres objectifs importants : la réduction du fardeau de réponse, une utilisation normalisée et maximale des

¹ Marie-Noëlle Parent (paremar@statcan.ca), Michelle Simard (simamic@statcan.ca), Immeuble R.H. Coats, 3-R, Parc Tunnel, Ottawa, Ontario, K1A 0T6, Canada

données administratives, des méthodes et systèmes unifiées pour tous les processus de traitement et l'utilisation d'une seule base de sondage pour toutes les industries.

Avec l'arrivée des enquêtes provenant de programmes réguliers dans la plate-forme de l'EUE, le plan de sondage, notamment la répartition de l'échantillon est devenue une étape cruciale du processus d'échantillonnage. La formule du partage de la TVH requiert une précision des estimations provinciales égales et détaillées pour toutes les provinces. Par contre, des estimations détaillées à différents niveaux industriels sont également requises par les programmes réguliers. La répartition de l'échantillon se doit donc de rencontrer ces deux spécifications de précision tout en essayant de réduire le fardeau de réponse.

L'objectif de cet article est de décrire le processus d'échantillonnage de l'enquête. La section 2 présente différentes définitions et concepts qui seront utilisés au cours de l'exposé. La section 3 décrit la stratification avec l'utilisation de seuils optimaux. La procédure de sélection et la technique de réseautage seront brièvement décrites dans la section 4. Finalement, la section 5 décrit la méthode utilisée pour allouer l'échantillon selon les contraintes provinciales et industrielles. La méthode, qui est basée sur un algorithme d'itération du quotient avec le coefficient de variation, constitue la principale nouveauté du processus d'échantillonnage de 1999.

2. QUELQUES DÉFINITIONS

2.1. Fichier Univers de l'Enquête (FUE)

Le processus d'échantillonnage débute avec la création de la base de sondage : le FUE. Ce fichier est créé par le registre des entreprises de Statistique Canada et couvre la population cible de l'enquête. Tous les établissements sur le FUE sont classifiés par un code SCIAN : le système de classification industriel de l'Amérique du Nord. Le code SCIAN à 6 chiffres

(SCIAN-6) décrit l'activité économique de l'établissement. Le secteur industriel, simplement appelé secteur, correspond généralement à un SCIAN à deux chiffres.

2.2 Cellule et unité d'échantillonnage

Dans l'EUE, la cellule est un concept essentiel. Deux dimensions importantes gouvernent le concept de cellule. La première est la dimension provinciale alors que la deuxième correspond aux différentes industries dont les activités d'enquête appartiennent à l'EUE. Plus précisément, une cellule est constituée de tous les établissements qui opèrent dans la même province et qui sont dans la même agrégation de codes SCIAN-6 utilisée pour la stratification industrielle donnée tel que montré dans le tableau 1.

L'unité d'échantillonnage est étroitement liée au concept de cellule. Une entreprise peut avoir des établissements dans différentes cellules. Autrement dit, une cellule sépare une entreprise en grappes d'établissements qui sont appelées des unités d'échantillonnage. Une unité d'échantillonnage est donc constituée de tous les établissements d'une entreprise appartenant à la même cellule.

2.3 Entreprises simple et complexe

Une entreprise peut être simple ou complexe. Pour remplir les critères de complexité, une entreprise doit respecter au moins un des critères suivants:

- Multi-provinciale: avoir des établissements dans plus d'une province
- Multi-SCIAN: avoir des établissements dans plus d'une activité économique
- Multi-légale: être propriété de plus d'une entité légale

Une entreprise qui ne rencontre aucune de ces conditions est dite simple.

Tableau 1. Représentation des cellules de l'EUE

SCIAN/Prov.	Prov.1	Prov.2	Prov.3
Agrégation A de SCIAN	Cellule ₁₁	Cellule ₁₂	Cellule ₁₃
Agrégation B de SCIAN	Cellule ₂₁	Cellule ₂₂	Cellule ₂₃
Agrégation C de SCIAN	Cellule ₃₁	Cellule ₃₂	Cellule ₃₃

3. STRATIFICATION

Il existe trois niveaux de stratification : les deux premiers coïncident avec les dimensions de la cellule, i.e. la province et l'agrégation industrielle. Ces deux variables proviennent du FUE. Il est à noter qu'elles correspondent au concept de cellule tel que décrit dans la section précédente. La troisième dimension est une variable de taille. En général, cette variable est le revenu annuel, également disponible sur la base. Deux algorithmes de calcul de seuils sont utilisés. Le premier sert à obtenir un seuil qui distinguera les unités éligibles à recevoir un questionnaire, i.e. les seuils de Royce-Maranda (1998). Le deuxième, l'algorithme de Lavallée et Hidioglou (1988), permet d'obtenir les bornes qui délimitent les unités appartenant aux différentes strates selon la variable de taille. Pour plus d'information sur la stratification dans l'EUE, lire Simard et Laniel (1998).

3.1 Seuils de Royce-Maranda

Afin de diminuer le fardeau de réponse, certaines unités à faible revenu ne seront pas éligibles à être enquêtées par questionnaire. Par contre, des données administratives seront utilisées pour l'estimation des caractéristiques de cette population. La détermination d'une borne pour identifier les unités à enquêter dans chaque cellule est donc nécessaire. Dans la première année de l'EUE, une seule borne arbitraire était utilisée pour toutes les cellules ce qui ne s'est pas révélé optimal puisque cela ne tenait pas compte de l'importance de chaque cellule. Un comité de travail mené par Don Royce et François Maranda a donc établi 6 bornes distinctes. Pour chaque cellule, l'idée est d'identifier, parmi les 6 bornes possibles, celle qui exclura le plus d'unités sans exclure plus de 5 % de l'activité économique.

3.2 Algorithme de Lavallée-Hidioglou

La stratification selon la variable de taille se fait par l'algorithme de stratification de Lavallée-Hidioglou dans chaque cellule. Pour l'application de cet algorithme, un coefficient de variation (CV) est nécessaire en entrée. La méthode qui permet de dériver ces CV est présentée dans la section 5. Une fois que ces CV initiaux sont établis, il importe d'identifier toutes les unités présélectionnées. Ces unités sont spécifiées dès le départ soit par les clients, ou soit selon d'autres critères spéciaux. Les établissements restants sont ensuite soumis à l'algorithme qui calcule les bornes pour stratifier chacune des cellules en trois strates: une strate à tirage complet et deux strates à tirage partiel. Dans un même

temps, l'algorithme identifie les tailles d'échantillon de chacune des strates. Dans certains cas où la cellule ne contient pas suffisamment d'unités pour être stratifiée, toutes les unités appartenant à ces dernières sont recensées.

Comme dans toute enquête, l'EUE fait aussi face aux imperfections de la base de sondage ainsi qu'à la non-réponse. Afin de contrer ces facteurs inévitables, les tailles sont augmentées par l'application d'un taux de décès ainsi que d'un taux de non-réponse, déterminés par les gérants de chaque enquête selon leurs expériences précédentes, si possible.

4. SÉLECTION DE L'ÉCHANTILLON

4.1 Numéros aléatoires communs et sélection de l'échantillon

La première étape menant à la sélection est l'assignation d'un numéro aléatoire commun (NAC) à chaque unité d'échantillonnage. Les unités d'échantillonnage sont ordonnées selon leur NAC dans chaque strate et sont ensuite sélectionnées selon le nombre requis en commençant par le plus petit NAC. Tous les établissements appartenant aux unités sélectionnées feront partie de l'échantillon. Ce processus est identique à un échantillonnage aléatoire simple à l'intérieur de chaque strate.

4.2 Réseautage

La dernière étape est le réseautage ou échantillonnage par réseau. Ce concept a été appliqué en 1997, puis a disparu en 1998, et a finalement été modifié et de nouveau appliqué en 1999. Le principe du réseautage est que lorsqu'une unité est sélectionnée dans l'échantillon, toutes les unités de la même entreprise et du même secteur d'activité seront incluses dans l'échantillon. Ce processus augmente donc légèrement la taille initiale de l'échantillon. Simard et Hidioglou (1999) donnent plus de détails sur l'échantillonnage par réseau.

5. OBTENTION DES CV INITIAUX PAR LA MÉTHODE ITÉRATIVE DU QUOTIENT

5.1 Origine de la méthode

En 1997 et 1998, un même CV était visé dans toutes les cellules, peu importe la province ou l'industrie considérées. Cette méthode est simple mais elle ne tient pas compte de l'importance relative des cellules. De plus, les CV provinciaux obtenus n'étaient pas égaux. Une méthode permettant d'obtenir des CV

différents par cellule, tout en étant contraint par les CV provinciaux et les CV des industries, a donc été développée et utilisée en 1999. Une approche similaire à un calage aux marges, i.e. une méthode itérative du quotient appliquée sur les CV permet d'obtenir des CV cellulaires variables tout en respectant les CV marginaux requis. L'application de cette méthode a été faite en deux étapes.

5.2 Application de la méthode itérative du quotient

Premièrement, le processus d'allocation est basé sur la spécification d'un CV national pour toutes les industries mises ensemble. La figure 1 montre les deux étapes qui sont expliquées dans cette section. Essentiellement, la première étape consiste à obtenir des CV sectoriels à partir du CV national, toutes industries, toutes provinces confondues. La deuxième étape consiste à prendre ces CV sectoriels obtenus à la première étape et d'en dériver des CV cellulaires, i.e. au niveau de chaque cellule (voir Figure 1).

La première étape distribue le CV national, CV_t , à travers les secteurs. Il n'est pas très efficace de contraindre les CV sectoriels à avoir la même valeur puisque les secteurs ne revêtent pas la même importance. Une allocation de puissance (*exp*) a donc été utilisée afin de minimiser les différences entre les CV des secteurs plus et moins importants. Cette allocation est décrite dans Bankier (1988). Pour un secteur s donné, CV_s est calculé de façon à être inversement proportionnel à son revenu, défini par GBI_s , tel que décrit par (1) :

$$CV_s = \frac{CV_t GBI_t}{GBI_s^{s \exp} \sqrt{\sum_s GBI_s^{2-2s \exp}}} \quad (1)$$

où $s \exp$ est la puissance utilisée pour le secteur s et $0 < s \exp < 1$ et GBI_t est le revenu total.

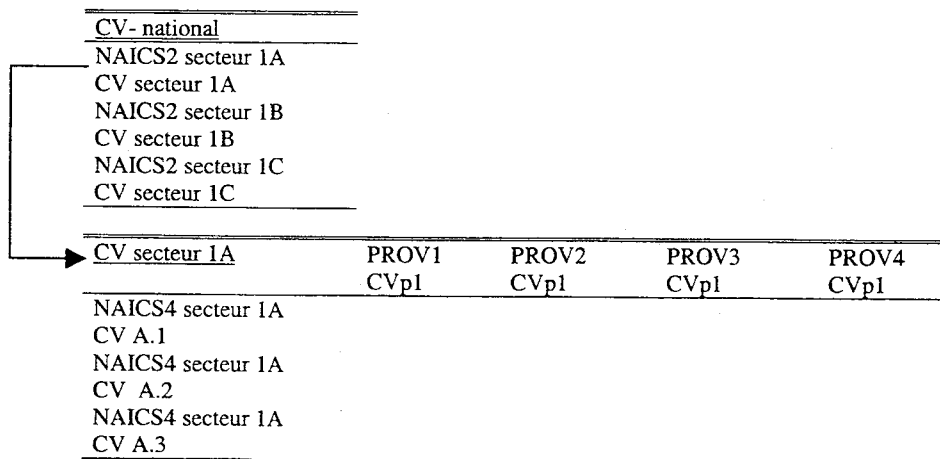
Une fois les CV sectoriels obtenus, ils sont mis en entrée dans un deuxième tableau où ils sont, de nouveau, soumis à une allocation en puissance. Cette fois-ci, par contre, la dimension industrielle est beaucoup plus détaillée, i.e. au niveau du SCIAN à 4 chiffres. Chacun des secteurs subira ainsi une deuxième allocation en puissance qui sera indépendante de la première.

Au niveau des provinces, les CV provinciaux, CV_p , sont tous égaux. Cela a été fait pour satisfaire les contraintes provinciales de la TVH. Pour une province p donnée, CV_p est donné par (2) :

$$CV_p = \frac{CV_t GBI_t}{\sqrt{\sum_p GBI_p^2}} \quad (2)$$

Une fois ces CV des marginales du tableau obtenus, la deuxième étape consiste à compléter la méthode itérative pour obtenir les CV de chacune des cellules de ce dernier. Le deuxième tableau correspond exactement au tableau 1 de la section 2, car il représente le concept de cellule, tel que défini par l'EUE. Les cellules représentent les cases du deuxième tableau de la figure 1. En effet, les provinces sont en colonne et la stratification industrielle est en rangée. En théorie, tous les CV cellulaires ainsi obtenus sont prêts à être entrés dans l'algorithme de Lavallée-Hidiroglou.

Figure 1



5.3 Application des plafonds par cellule

Les CV ainsi calculés sont plus optimaux et tiennent mieux compte de l'importance de la cellule. Par contre, dans certains cas où le revenu total de la cellule est relativement petit, le CV visé est particulièrement élevé. Il est important de différencier les cellules importantes des cellules de moindre importance mais il est tout de même souhaitable d'obtenir de bonnes estimations pour chaque cellule. Dans cette optique, tous les CV visés qui étaient supérieurs à 15%, peu importe l'enquête ou le secteur auxquels ils appartenaient, ont été ramenés à 15%.

Dans un même ordre d'idée, certaines cellules se sont avérées particulièrement importantes à l'intérieur d'une enquête pour diverses considérations. Pour ces quelques exceptions, identifiées par les analystes de chaque enquête, il était permis de fixer un plafond plus bas. Ces plafonds varient entre 5% et 10%. Notons que puisqu'il y a environ 1000 cellules appartenant à l'EUE, un nombre égal de CV a été calculé. De ce nombre, environ 40% ont subi des modifications suite à l'application des divers plafonds. Ce pourcentage de changements est particulièrement élevé.

5.4 Résultats

Bien que les résultats globaux aient été satisfaisants, la méthode a été et est présentement en révision pour le prochain cycle de l'EUE. En effet, puisque les provinces ont été contraintes à avoir le même CV marginal, certaines provinces ont été pénalisées. Conséquemment, l'importance relative de certaines provinces n'est pas reflétée par cette approche. De plus, l'utilisation des plafonds a été quelque peu sur-utilisée.

5.5 Développements futurs

Il est impossible selon l'objectif principal du PASEP d'oublier complètement la contrainte d'égalité des CV provinciaux marginaux. Par contre, une certaine flexibilité peut être changée lors de la deuxième étape, i.e. l'allocation par secteur. En effet, une allocation de puissance peut y être considérée au niveau de la province. De plus, l'allocation de puissance qui a été utilisée au niveau du secteur et au niveau de l'agrégation industrielle peut aussi être modifiée si cela mène à de meilleurs résultats.

Il s'agit donc de trouver la meilleure combinaison de CV total ainsi que des coefficients de l'allocation de puissance au niveau du secteur dans la première étape et au niveau de l'agrégation industrielle et de la province dans la deuxième étape. Une meilleure combinaison de ces facteurs permettrait aussi de réduire la nécessité des plafonds.

6. CONCLUSION

Les différentes étapes du processus d'échantillonnage sont, pour la plupart, déjà en place et ne seront pas modifiées pour l'EUE 2000. La seule étape pour laquelle des changements mineurs seront effectués est celle qui détermine l'allocation des CV. Il semble que la meilleure option à envisager pour 2000 est d'ajouter une petite allocation en puissance au niveau des provinces dans la deuxième étape de la méthode. Cette option est aussi intéressante d'un point de vue stabilité puisqu'elle ressemble beaucoup à la méthode actuelle.

REMERCIEMENTS

Les auteurs tiennent à remercier Jocelyn Smith pour son travail sur la méthode d'allocation des CV pour l'EUE 2000 et Claude Girard pour son aide et son support.

RÉFÉRENCES

- Bankier M. (1988). Power allocations: Determining Sample Sizes for Sub-national Areas. *Document interne*. Statistique Canada.
- Lavallée, P. et Hidirolou, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, no.14, pp 33-45.
- Royce, D. et Maranda, F. (1998). Task Force on Data Acquisition from Businesses. *Internal Document*. Statistique Canada.
- Simard, M. et Hidirolou, M. (1999). Estimation For Annual Business Surveys Based On Two-Phase Network Sampling. *SSC Proceedings of the Survey methods Section*, pp 11-19.
- Simard, M. et Laniel, N. (1998). Échantillonnage et estimation pour l'enquête unifiée sur les entreprises. *SSC Proceedings of the Survey methods Section*, pp 77-82.