

## ÉCHANTILLONNAGE ET ESTIMATION POUR L'ENQUÊTE UNIFIÉE SUR LES ENTREPRISES

Michelle Simard et Normand Laniel<sup>1</sup>

### RÉSUMÉ

Ce nouveau projet qui consiste à intégrer les différents programmes annuels à Statistique Canada avec une approche unifiée est en plein développement dans cette première année pilote. Un des plus grands défis de cette enquête a été de développer une stratégie d'échantillonnage et d'estimation basée sur des méthodes d'enquête standard, mais avec une application beaucoup plus vaste et complexe qu'une enquête entreprise traditionnelle. De plus, les échéanciers de production sont très serrés.

Les principaux aspects qui ont été considérés dans l'élaboration du plan d'échantillonnage seront présentés, entre autres: i) une unité d'échantillonnage commune pour les différentes structures d'entreprise au Canada, ii) la répartition et la sélection de l'échantillon qui sont fortement liées avec la stratégie d'acquisition des données, iii) l'utilisation d'une seule base de sondage, iv) une approche à deux-phases, v) les besoins en précision et en détails des utilisateurs, souvent différents et contrastant avec l'importance de réduire le fardeau de réponse.

Également présenté, la méthode d'estimation proposée: les estimateurs de régression généralisés avec l'utilisation de données fiscales, sélectionnés dans la première phase, utilisées comme information auxiliaire dans le calage des estimateurs.

**MOTS-CLÉS :** Échantillonnage à deux phases; échantillonnage par réseau; numéro aléatoire permanent; estimateur par calage; post-stratification; régression.

### ABSTRACT

This new project which consist of integrating different Statistics Canada annual programs with a unified approach is in full development in this first pilot year. One of the biggest challenges of this survey has been to develop a sampling and estimation strategy using standard methodologies, but with a more complex and global application than traditional business surveys. Furthermore, the production schedule is very tight. The most important aspects which were considered in the sampling design will be presented, among them: i) a common sampling unit for all of the different structure of Canadian business enterprises, ii) the sampling allocation and selection which were strongly related to the data acquisition strategy, iii) the use of one survey frame, iv) a two-phase approach, v) the users' requirements in terms of precision and details. Those aspects are often different and contrasting with the importance of response burden reduction.

Also presented will be the proposed estimation method: the generalized regression estimator with a intensive use of fiscal data, selected in the first phase, as auxiliary information in the calibration of the estimator.

**KEY WORDS :** Two-phase sampling; Network sampling; Hash number; Calibration estimators; Post-stratification, regression.

### 1. INTRODUCTION

L'Enquête unifiée des entreprises (EUE) a pour objectif de produire des estimations annuelles sur la production économique des établissements par province et branche d'activité économique. En particulier, on doit calculer des statistiques sur les revenus, les dépenses, les marchandises utilisées et produites ainsi que le type de client. De plus, des

statistiques financières (e.g. revenus et dépenses consolidées, actifs et dettes) doivent être calculées au niveau de l'entreprise. Cet article se concentre sur les aspects d'échantillonnage et d'estimation touchant les statistiques au niveau de l'établissement.

Quatre objectifs stratégiques ont été établis en ce qui concerne le plan d'échantillonnage de l'enquête. Premièrement, une seule base de sondage sera utilisée pour tiré les échantillons afin que les estimations

<sup>1</sup> Michelle Simard, DMEE, RHC 11-M, Statistique Canada, Ottawa, K1A 0T6, Normand Laniel, DMEE, RHC 11-O, Statistique Canada, Ottawa, K1A 0T6.

produites pour les différentes branches d'activité économique soient cohérentes. Deuxièmement, l'utilisation de données fiscales sera maximisée afin de réduire le fardeau des répondants, surtout pour les petites entreprises. Troisièmement, le plan devra permettre un bon contrôle du fardeau de réponse globale auprès des entreprises. Quatrièmement, le plan devra donner la possibilité d'analyser la cohérence entre les micro-données au niveau de l'entreprise et celles au niveau de ses établissements. L'échantillon de l'Enquête unifiée des entreprises de 1997 a été sélectionné à la fin de 1997 afin de permettre la collecte des données pendant le printemps et l'été de 1998. Il est prévu de produire les estimations pour le printemps de 1999.

La deuxième section présente le plan d'échantillonnage à deux phases utilisé pour l'enquête qui fait un usage important de données fiscales. Les méthodes d'estimation, quant à elles sont décrites à la section 3. La section 4 donne un aperçu des plans futurs.

## **2. ÉCHANTILLONNAGE À DEUX PHASES ET FARDEAU DE RÉPONSE**

La base de sondage de l'EUE est le Registre des entreprises (RE) tel que décrit dans Castonguay (1998). Le RE fournit la liste des unités statistiques (i.e. entreprise et établissements) avec l'information nécessaire pour concevoir et sélectionner l'échantillon. Cet échantillon est à deux phases afin de réduire le fardeau des répondants. À la première phase, on tire essentiellement un échantillon de données fiscales qui n'impliquent aucun fardeau supplémentaire aux entreprises. À la deuxième phase, on tire un sous-échantillon de la première phase pour lesquelles certaines entreprises seront enquêtées par questionnaire. Au moment de l'estimation, on utilisera l'information du grand échantillon pour améliorer celle du petit par une technique d'estimation par calage. De plus, on enquête par questionnaire qu'une partie de l'univers; celle au-dessus des seuils de fardeau de réponse. Pour compléter l'estimation totale, certaines unités seront sélectionnées sous les seuils où on utilisera que les données fiscales. On maximise ainsi l'utilisation des données fiscales de deux façons.

### **2.1 Première phase**

L'échantillon de première phase est essentiellement conçu pour produire des estimations fiables à un niveau d'agrégation très détaillé, i.e. province par SCIAN à 6 chiffres.

### **2.1.1 Unité d'échantillonnage**

Étant donné que l'objectif de l'EUE est de produire des estimations fiables selon la province et l'activité économique, cela implique que les paramètres du plan de sondage devront être contrôlés selon ces deux variables de ventilation. C'est pour cette raison que l'unité d'échantillonnage choisie est une grappe d'établissements faisant partie de la même entreprise et ayant la même province ainsi que la même activité économique au niveau du SCIAN à 6 chiffres. Ainsi une entreprise complexe peut être liée à plus d'une unité d'échantillonnage. Quant aux entreprises simples, elles ne sont liées qu'à une seule unité d'échantillonnage.

### **2.1.2 Stratification des unités d'échantillonnage**

Trois niveaux de stratification sont utilisés dans l'EUE de 1997. Les deux premiers servent à rencontrer les besoins des utilisateurs des estimations et le dernier à rendre le plan efficace. Premièrement, les unités d'échantillonnage sont groupées selon la province ou le territoire le cas échéant. Deuxièmement, à l'intérieur de chaque strate primaire ou géographique, on classe les unités selon le code SCIAN à 6 chiffres. Troisièmement, chaque strate secondaire (i.e. croisement géographique et d'activité économique) est divisée en trois groupes relativement homogènes quant à la taille des unités. Ces groupes de taille sont aussi appelés strates tertiaires. On obtiendra trois groupes ou strates soient; une strate à tirage complet et deux strates à tirage partiel. La mesure de taille utilisée pour cette stratification est la somme des revenus bruts d'entreprise de chacun des établissements de la grappe. L'origine de ces revenus bruts est expliquée dans Castonguay (1998). La méthode de division en groupe taille est celle de Lavallée et Hidiroglou (1988). Cette méthode consiste essentiellement à déterminer les (trois) groupes qui minimisent la taille totale de l'échantillon pour un coefficient de variation (CV) désiré. Cette minimisation est faite en fonction d'une méthode de répartition choisie par le statisticien. La méthode utilisée pour l'EUE 1997 est décrite à la section suivante.

### **2.1.3 Répartition de l'échantillon**

La taille de l'échantillon de chaque strate secondaire a été déterminée par l'application de la méthode de Lavallée et Hidiroglou avec un CV désiré de 5% pour la variable revenu brut. Cette détermination a été effectuée avec une répartition proportionnelle à la racine carrée du revenu brut entre les 3 strates de taille. À noter que la strate contenant les unités de plus grande taille obtient habituellement un échantillon de taille égale à celle de sa population. En d'autres mots, toutes ses unités seront autoreprésentatives. Ce sont

les strates à tirage complet. Les deux autres strates sont dites à tirage partiel.

Après avoir déterminer les tailles d'échantillons hauts, on les a rajustées à la hausse afin de tenir compte de la proportion des unités inactives sur le Registre ainsi que du taux prévu de non-réponse.

#### 2.1.4 Sélection de l'échantillon

Pour la sélection de l'échantillon, une méthode d'échantillonnage par réseau est utilisée afin d'assurer que toute entreprise complexe sélectionnée aura toutes ses unités sélectionnées. Cette sélection se fait en deux étapes. Premièrement, indépendamment dans chaque strate tertiaire, on sélectionne un échantillon par tirage aléatoire séquentiel. Pour cette étape, on assigne un nombre aléatoire sur l'intervalle  $[0,1[$  à chaque unité d'échantillonnage. Ce nombre est généré à l'aide d'un générateur de nombre pseudo-aléatoire utilisant le numéro d'identification de l'unité. Ensuite, ces nombres sont ordonnés en ordre croissant et l'on sélectionne les premières unités de sorte que la taille désirée soit atteinte. Cette méthode est équivalente à un échantillonnage aléatoire simple sans remise. La deuxième étape consiste à sélectionner toutes les entreprises ayant au moins une unité de sélectionner lors de la première étape puis à sélectionner toutes leurs unités. De cette façon, toute entreprise sélectionnée dans l'échantillon aura aussi tous ses établissements dans l'échantillon. Avec cet échantillonnage par réseau, la probabilité de sélectionner une entreprise ainsi que n'importe laquelle de ses unités s'écrit :

$$\pi_E = 1 - \prod_{i \in E} (1 - \pi_i)$$

où  $\pi_E$  est la probabilité de sélection de l'entreprise  $E$  et  $\pi_i$  la probabilité de sélection de l'unité d'échantillonnage  $i$ .

Toutes les unités sélectionnées dans les strates à tirage complet recevront un questionnaire dès leur sélection à la première phase.

#### 2.2 Deuxième phase

L'échantillon de deuxième phase est conçu pour produire des estimations fiables à un niveau d'agrégation moins détaillé que celui de première phase (i.e. province par regroupement de SCIAN à 6 chiffres). Ceci afin d'imposer un fardeau de réponse raisonnable aux moyennes et petites entreprises sélectionnées car elles seront contactées afin de remplir un ou plusieurs questionnaires. Cependant, en combinant les données des deux phases, on pourra obtenir des estimateurs plus efficaces qu'avec la deuxième seule. À la deuxième phase, une quatrième variable est utilisée: la structure ou complexité de

l'entreprise. Le but de cet ajout est de tenir compte du fait que les données fiscales ne sont disponibles qu'au niveau de l'entreprise et non des établissements. Une conséquence est que l'échantillonnage à deux phases n'est possible que pour les entreprises simples où l'entreprise est égale à l'établissement. Étant donné que l'on ne peut utiliser les méthodes de calages que si les données sont de mêmes niveaux, seules les entreprises simples seront sujettes à l'estimation à deux phases. La section 3 donne les détails à ce sujet.

#### 2.2.1 Unité d'échantillonnage

À l'estimation, pour les entreprises simples, des modèles seront construits afin d'expliquer les variables d'enquête avec les variables provenant des dossiers fiscaux. Il faut donc assurer le chevauchement entre les échantillons de première et de deuxième phase. Pour cette raison, l'unité d'échantillonnage de première phase est aussi utilisée pour la deuxième. De plus, les numéros aléatoires générés à la première phase seront réutilisés pour la sélection de la deuxième phase pour ainsi assurer un chevauchement maximum.

#### 2.2.2 Stratification des unités d'échantillonnage et acquisition des données.

Pour les entreprises complexes, les unités suivent tout simplement la stratification de première phase. Cependant, on rajoute un autre niveau de stratification ou seuil en terme d'acquisition des données. Toutes celles qui ont un revenu sous la barre du million de dollars ne seront pas éligibles à recevoir un questionnaire, mais on produira plutôt les estimations avec des données fiscales afin de réduire le fardeau de réponse. Ce sont les unités sélectionnées dans les strates dites à tirage nul en terme de questionnaire.

Pour les entreprises simples, les unités suivent de très près la stratification de première phase sauf en ce qui concerne celle utilisant la branche d'activité économique. En effet, on fait un regroupement des SCIAN à six chiffres afin de réduire le nombre de strates avec de petites populations sujettes à de grandes fractions de sondage. En retour, cela permet de réduire globalement le fardeau des répondants. De plus, pour les petites entreprises simples, on produira également des estimations avec des données fiscales pour les unités sous les seuils. Pour certaines industries le seuil d'exclusion est de \$50,000 et pour d'autres de \$150,000. On utilise également le terme de strates à tirage nul pour ces unités.

#### 2.2.3 Répartition de l'échantillon pour les entreprises simples

La taille de l'échantillon de chaque strate secondaire a été déterminée de façon à produire des estimations avec un CV désiré de 15% pour la variable revenu

brut. Cette détermination a été effectuée avec une répartition proportionnelle à la racine carrée du revenu brut entre les strates de taille. Étant donné que les entreprises simples sélectionnées dans les strates à tirage complet de la première phase reçoivent déjà un questionnaire, il ne reste qu'à sélectionner les entreprises simples de petite et moyenne tailles dans les strates à tirage partiel.

### 2.2.4 Sélection de l'échantillon

Pour la sélection de l'échantillon, une méthode d'échantillonnage aléatoire simple sans remise est utilisée qui ressemble à la première étape de la première phase. Indépendamment dans chaque strate tertiaire, on sélectionne un échantillon par tirage aléatoire séquentiel. Pour cette étape, on réassigne le même nombre aléatoire sur l'intervalle [0,1[ à chaque unité, on ordonne et l'on sélectionne les premières unités de sorte que la taille désirée soit atteinte. De cette façon, la deuxième phase est un sous-échantillon du premier.

## 3. PONDÉRATION ET ESTIMATION

### 3.1 Approche générale : Famille d'estimateur par calage

Dans le contexte de l'EUE, tous les estimateurs proposés proviennent de la famille des estimateurs par calage (Särndal, Deville, 1992). L'approche retenue va permettre de calculer un facteur de calage pour chacune des phases d'échantillonnage. On obtient avec le produit du ou des facteurs de calages et du ou des poids de sondages, les poids finaux pour chacune des unités ayant été sélectionnées pour recevoir un questionnaire. Ce poids final sera utilisé pour l'ensemble des variables à estimer, et ce pour tous les domaines. À noter que dans tous les cas, le poids de sondage est l'inverse de la probabilité de sélection.

### 3.2 Forme générale des estimateurs

Note : La notation utilisée est la notation standard telle que définie par Särndal et Deville.

Telle que décrite, l'acquisition des données est quelque peu complexe. Elle dépend essentiellement de la structure des entreprises; soit complexes ou simples, et de la probabilité de sélection de l'unité. Conséquemment, la forme de l'estimateur d'un total peut être décomposée de la façon suivante:

$$\hat{Y}_{total} = \hat{Y}_{cplex\_tc} + \hat{Y}_{simpl\_tc} + \hat{Y}_{cplex\_tp} + \hat{Y}_{simpl\_tp} + \tilde{Y}_{cplex\_tn} + \tilde{Y}_{simpl\_tn} \quad eq. 1$$

où *cplex* indique les unités provenant des entreprises complexes et *simpl* celles provenant des entreprises simples.

où *\_tc* indique les unités sélectionnées dans les strates à tirage complet; *\_tp* celles sélectionnées dans les strates à tirage partiel et *\_tn* celles sélectionnées dans les strates à tirage nul.

Pour l'estimation d'un domaine donné, il est possible que l'on retrouve ces six types d'unités participant à l'estimation. L'estimation de chacun de ces six types d'unités peut théoriquement être faite de façon différente. De façon concrète, pour l'EUE, on retrouve trois méthodes d'estimation distinctes définies pour trois regroupements de type d'unités.

Le premier groupe contient toutes les unités recevant un questionnaire dès leur sélection à la première phase. Ce sont spécifiquement toutes les unités provenant des entreprises complexes ; sélectionnées dans les strates à tirage complet ou dans les strates à tirage partiel et les unités provenant des entreprises simples sélectionnées dans les strates à tirage complet.

Les estimateurs  $\hat{Y}_{simpl\_tc}$ ,  $\hat{Y}_{cplex\_tp}$  et  $\hat{Y}_{cplex\_tc}$  auront la forme suivante:  $\hat{Y} = \sum g_{gm} w_h y_i$  eq. 2

Le deuxième groupe contient les unités ayant été sélectionnées selon la stratégie à deux phases soient; obtention de données fiscales pour les unités sélectionnées à la première phase et envoi d'un questionnaire pour les unités sélectionnées à la deuxième phase. Spécifiquement, ce sont les unités provenant des entreprises simples et sélectionnées dans les strates à tirages partiels.

L'estimateur  $\hat{Y}_{simpl\_tp}$  aura la forme suivante

$$\hat{Y}_{simpl\_tp} = \sum_{i \in sample} g_{1\_gm} w_{h\_1} g_{2\_gm} w_{g\_2} y_i \quad eq. 3$$

Le troisième groupe contient toutes les unités sélectionnées sous les seuils de couverture. Ce sont des unités provenant des entreprises complexes ou simples sélectionnées dans les strates à tirage nul.

Les estimateurs  $\tilde{Y}_{cplex\_tn}$  et  $\tilde{Y}_{simpl\_tn}$  auront la forme  $\tilde{Y} = \sum g_{gm} w_h \tilde{y}_i$  eq. 4

où les  $y$  ne sont pas des variables provenant des questionnaires, mais plutôt des variables provenant des données fiscales avec des concepts similaires aux données des questionnaires.

### 3.3 Méthodes d'estimation

Toutes les unités de l'échantillon sélectionnées à la première phase sont calées sur de nouveaux comptes

de population pour ainsi produire des totaux calculés avec l'estimateur de post-stratification. Cet ajustement est calculé pour les unités des trois groupes. Pour les unités du groupe un et du groupe trois, les estimations de totaux pour les variables d'intérêt sont directement produites avec ce seul ajustement. Pour les unités du groupe deux, dans un premier temps, ce sont les estimations de totaux pour les variables des données fiscales qui sont produites. Puis, dans un deuxième temps, les unités de l'échantillon de deuxième phase sont réparties en groupe modèles où des paramètres de régression seront calculés, basés sur un modèle construit entre les variables enquêtées et les variables des données fiscales. Le deuxième facteur de calage est basé sur les formules de l'estimateur de régression. Les prochaines sections décrivent les deux types d'estimateurs de façon plus détaillée. La dernière section présente les aspects futurs qui seront à développer dans les prochains mois.

### 3.3.1 L'estimateur de post-stratification

L'EUE est une des premières enquête entreprise importante à Statistique Canada à utiliser cette technique dans la méthode d'estimation. La technique permet de caler les estimations sur des comptes à jour de la population. Elle permet ainsi de corriger par un simple ajustement les erreurs de classification, i.e. le code industriel et le code provincial, et les erreurs de taille, i.e. le revenu (pour les sauts de strates) survenues à la sélection. La technique permet une correction pour les unités mortes et les unités manquantes qui n'auraient pu être enlevées ou mises à temps sur la base de sondage au moment de la sélection de l'échantillon. Elle permet également de répartir l'échantillon de première phase selon les groupes mentionnés dans la section précédente. En définissant les post-strates entre autres, par entreprises complexes et simples, ce qui n'était pas fait au niveau de la stratification, on permet la production de totaux pour des regroupements d'unités, autre que pour ceux de la stratification, par différentes méthodes d'estimation, un peu comme le fait la technique de l'estimation par domaine.

En plus de répartir l'univers par entreprises complexes et simples et par types de tirages, les post-strates sont également définies aux niveaux des branches d'activités industrielles et pour chacune des provinces et territoires.

On obtient ainsi directement l'estimateur pour les unités du premier groupe; l'estimateur de post-stratification. Si on prend les unités provenant des entreprises simples sélectionnées dans les strates à tirage complet comme exemple. Supposons  $N_{pst-simple-1c}$  qui représente le nouveau compte d'établissements dans la population provenant des entreprises simples

avec un revenu plus élevé que la borne de strate à tirage complet, par province et par code SCIAN à trois chiffres.

On obtient:

$$\hat{Y}_{simp-1c} = \sum_{i \in sample} g_{gm} w_h y_i = \sum_{i \in sample} \frac{N_{pst-simp-1c}}{\hat{N}_{pst-simp-1c}} \frac{N_h}{n_h} y_i \quad eq. 5,$$

$$\text{où } \hat{N}_{pst-simp-1c} = \sum_{i \in \text{Echant.}} w_i \text{ pour toutes les variables}$$

d'intérêt et les domaines d'estimation.

La méthode est similaire pour les unités du troisième groupe soient celles sélectionnées dans les strates à tirage nulle où l'estimateur de post-stratification est également utilisé pour produire directement les estimations requises.

Ce n'est pas le cas pour les unités du deuxième groupe, où l'ajustement est appliqué sur les variables des données fiscales, qui vont être utilisées subséquemment dans le modèle de régression. L'estimateur de post-stratification est plutôt utilisé pour la production de totaux fiables pour les variables indépendantes du modèle.

On obtient

$$\hat{X}_{simp-1p} = \sum_{i \in sample} g_{gm} w_h x_i = \sum_{i \in sample} \frac{N_{pst-simp-1p}}{\hat{N}_{pst-simp-1p}} \frac{N_h}{n_h} x_i \quad eq. 6$$

où les  $X_i$  sont les variables provenant des données fiscales sélectionnées à la première phase.

### 3.3.2 L'estimateur de régression

Pour les unités du deuxième groupe, un deuxième facteur de calage est calculé pour les unités de deuxième phase, basé sur l'information auxiliaire obtenue des unités de la première phase. Le facteur de calage est défini pour un groupe modèle  $gm$  donné:

$$g_{gm} = 1 +$$

$$\left( \hat{X}_{simp-1p} - \hat{X}_{phase 2} \right) \left( \sum w_g x_{gm} x'_{gm} / c_{gm} \right)^{-1} x_{mg} / c_{gm} \quad eq. 7$$

et où  $c_{gm}$  est le paramètre de la structure de la variance du modèle dans le groupe modèle  $gm$ .

Une fois l'information des unités de première phase post-stratifiée, telle que définie par l'équation 6, cette information est utilisée dans la formulation de l'estimateur final. En reformulant l'équation 3 avec le terme de l'équation 7 on obtient la formule suivante, qui est la formulation standard de l'estimateur de régression:

$$\hat{Y}_{simp-1p} = \hat{Y}_{phase 2} + \hat{\beta}_{phase 2} (\hat{X}_{simp-1p} - \hat{X}_{phase 2}) \quad eq. 8$$

et où les paramètres de régression sont calculés de la façon suivante avec les unités de deuxième phase:

$$\hat{\beta}_{phase\ 2} = \sum_{phase\ 2} \left( \frac{w_g XX'}{c_{mg}} \right)^{-1} \sum_{phase\ 2} \left( \frac{w_g XY}{c_{mg}} \right) \text{ eq. 9}$$

basé sur le modèle standard :  $Y = B X + \varepsilon$ .

L'estimateur de régression a été démontré comme le plus efficace.

#### 4. DÉVELOPPEMENTS FUTURS

Dans les prochains mois, plusieurs autres aspects devront être développés en plus de la production des estimations préliminaires pour les 7 industries pilotes. Du côté de la pondération et de l'estimation; il y a la mise en place des probabilités inégales de sélection dans une même strate dans le système d'estimation, le développement et la mise en application des méthodes de détection de valeurs influentes, i.e. de taille et de régression dans le système d'estimation. Du côté de

l'estimation de la variance; il y a le développement du calcul de l'estimation de la variance pour tenir compte de l'échantillonnage par réseau et le développement de la partie de la variance due à l'imputation.

#### 5. BIBLIOGRAPHIE

Castonguay, Éline (1998). Le Registre des entreprises, SSC 1998, recueil

Deville, J.-C. , et Särndal, C.E. (1992). Calibration estimators in survey sampling, *JASA*, 87, 376-382

Lavallée, P. And Hidioglou, M.A. (1988) On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33-43.