

L'ESTIMATION POUR L'ENQUÊTE UNIFIÉE SUR LES ENTREPRISES (EUE)

C. Cauchon, A.Côté, M.Simard¹

RÉSUMÉ

La nouvelle Enquête Unifiée sur les Entreprises (EUE) intègre différentes enquêtes annuelles de Statistique Canada. Un système d'estimation à applications généralisées et modulaires a été développé. Ce système sert essentiellement à calculer les poids finaux pour chaque enregistrement de l'échantillon. Ceci nécessite, entre autres, la post-stratification, la détection et le traitement des valeurs aberrantes ainsi que la modélisation des données d'enquête en fonction des données administratives. Une étude empirique comparant les estimations et les mesures de précision créées par le traitement des valeurs aberrantes et par les différentes stratégies d'estimations ont également été préparés.

MOTS CLÉS : Estimation; détection et traitement de valeurs aberrantes; post-stratification; estimateur à deux-phases.

ABSTRACT

The new Unified Enterprise Survey (UES) integrates many of Statistics Canada's annual surveys. An estimation system with general applications that is based on modules has been developed. The system's main objective is to obtain final weights for each selected record. Among the possible steps, there are post-stratification, outlier detection and treatment, as well as modelling between survey and administrative data. An empirical study, comparing the estimates and the variance based on the outlier treatment and the different estimation strategies, has been prepared.

KEY WORDS: Estimation; outlier detection and treatment; post-stratification; two-phase estimator.

1. INTRODUCTION

L'EUE a été mise en place dans le cadre du Projet pour l'Amélioration des Statistiques Économiques et Provinciales (PASEP). Un des objectifs de cette enquête est de produire des statistiques provinciales fiables au niveau de différents secteurs industriels. De plus, les estimations doivent être produites pour deux niveaux d'entités statistiques: l'entreprise ainsi que l'établissement. D'autres objectifs importants doivent également être rencontrés, entre autres, la réduction du fardeau de réponse et une utilisation maximale des données administratives. Pour ce, différentes mesures ont été prises à travers toutes les étapes de l'enquête. Pour le plan d'échantillonnage, un échantillonnage à deux-phases a été retenu. L'échantillon de première

phase consiste à obtenir les données administratives des unités sélectionnées. Pour la deuxième phase, un sous-échantillon du premier est sélectionné pour lesquels les unités recevront un questionnaire.

Lors de la première année, i.e. en 1997, l'EUE débute avec 7 enquêtes pilotes. En 1998, 11 nouvelles industries s'intègrent et pour 1999 le nombre d'enquêtes intégrées s'élève à 23. On prévoit encore l'ajout de 4 industries pour l'EUE-2000. Étant donné les différentes caractéristiques des populations des enquêtes intégrées, les méthodologies proposées doivent être flexibles et conserver une approche unifiée. Le défi a été de développer un système d'estimation permettant une approche commune à tous, mais permettant de tenir compte des spécificités des

¹ Caroline Cauchon (caucar@statcan.ca), Alain Côté, (coteala@statcan.ca), Michelle Simard, (simamic@statcan.ca), Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, 3^{ème} étage Édifice R.H.Coats, Ottawa, Ontario, K1A 0T6

différentes enquêtes. Par exemple, dans les enquêtes intégrées, il existe des industries avec une centaine d'unités représentant \$3 millions de revenus et d'autres ayant des tailles de populations de 120 000 unités représentant des milliards de dollars de revenus dans l'économie canadienne.

Les prochaines sections décrivent le système d'estimation, les différentes stratégies d'estimations possibles, telles la post-stratification ou l'estimateur du quotient et les contraintes méthodologiques rencontrées lors de la finalisation de celles-ci.

2. DESCRIPTION DU SYSTÈME D'ESTIMATION

Le système d'estimation développé pour l'EUE est un système modulaire à application généralisée. Étant donné la multitude d'enquêtes qui s'intègrent à l'EUE, il se doit de satisfaire une approche unifiée. Le même système d'estimation permet donc de produire des estimations avec tous les estimateurs de la famille des estimateurs par calage tel que décrit dans Deville et Särndal et ceci indépendamment du plan d'échantillonnage. Il permet donc de choisir différentes stratégies d'estimation, soient; l'estimateur Horvitz-Thompson, l'estimateur par le quotient, l'estimateur de post-stratification avec effectifs ou variable auxiliaire ou l'estimateur de régression généralisé. Trois types d'estimateurs peuvent être produits: l'estimateur du total, de la moyenne ou encore le ratio de deux variables. Il est basé sur un prototype du SGE (Système Généralisé d'Estimation) de Statistique Canada, développé pour les plans à deux-phases.

L'estimation de l'EUE se fait en plusieurs étapes, toutes correspondantes à un des modules du système. Dans le cas d'un échantillon à une phase, les étapes sont : i) la détection des valeurs aberrantes, ii) le calage choisi selon la stratégie d'estimation iii) l'estimation et l'estimateur de la variance selon la stratégie pour tous les domaines spécifiés. Dans le cas de l'estimateur à deux-phases, il y a des modules supplémentaires soient iv) la détection des valeurs aberrantes pour les données administratives, v) la détection des valeurs aberrantes lors de la régression entre les données d'enquêtes et les données administratives, vi) le calage choisi selon la stratégie d'estimation pour la deuxième phase de l'échantillon, vii) le calcul de l'estimateur de la variance à deux-phases. Chacun de ces modules sera expliqué dans les sections 3 et 4.

3. SYSTÈME D'ESTIMATION POUR L'ESTIMATEUR DE POST-STRATIFICATION

Cette section décrit les différents modules du système d'estimation pour la stratégie d'estimation choisie par la plupart des enquêtes de l'EUE soit, l'estimateur de post-stratification. Le système d'estimation n'utilise que les unités qui ont reçu un questionnaire pour produire les estimations. Les données administratives ne seront utilisées que dans le cas de l'estimateur à deux-phases. Les sous-sections décrivent respectivement, la détection et correction des valeurs aberrantes, la technique de calage proposée, le calcul de l'estimateur de variance ainsi que les indicateurs d'erreurs non-échantillonales.

3.1 Détection des valeurs aberrantes univariée

La détection des valeurs aberrantes est une combinaison de deux méthodes, soient celles de, Hidiroglou-Berthelot (1986) et d'une version modifiée d'une méthode interne à Statistique Canada; la méthode connue sous le nom de sigma-gap. Pour l'EUE, la détection s'est effectuée à partir du revenu total ou encore les dépenses totales car ces deux variables sont communes à toutes les industries et donne une bonne indication de la taille de l'unité.

Pour la méthode Hidiroglou-Berthelot, on utilise la formule suivante:

$$M1_j = \frac{wrev_j - Q(0.5)_i}{Q(0.75)_i - Q(0.5)_i}$$

où $wrev_{ij}$ est le revenu total pondéré de la j ième observation du i ème groupe.

$Q(0.5)_i$ est la médiane des revenus totaux pondérés du i ème groupe.

$Q(0.75)_i$ est le troisième quartile des revenus totaux pondérés du i ème groupe.

Pour la méthode du Sigma-Gap, il faut ordonner les observations à l'intérieur de chacun des groupes.

$$M2_j = \frac{(wrev_j - wrev_{j-1}) - D(0.5)_i}{D(0.75)_i - D(0.5)_i}$$

où $D(0.5)_i$ est la médiane des distances des observations $> Q(0.5)_i$

$D(0.75)_i$ est le troisième quartile de la distance des observations $> Q(0.5)_i$.

Tableau 1: Impact de la détection des valeurs aberrantes sur les estimations.

Valeurs Aberrantes: Province	Sans détection		Avec détection	
	Estimation	Variance	Estimation	Variance
	(\$millions)		(\$millions)	
T-N	300	4 ^E 13	300	4 ^E 13
IPÉ	80	8 ^E 12	80	8 ^E 12
N-É	720	2 ^E 15	690	9 ^E 14
N-B	460	3 ^E 14	460	3 ^E 14
QC	5 800	3 ^E 16	5 250	3 ^E 16
ON	9 700	8 ^E 16	9 100	7 ^E 16
MN	1 100	3 ^E 15	900	3 ^E 15
SK	700	5 ^E 14	570	4 ^E 14
AL	3 350	1 ^E 16	2 400	1 ^E 15
C-B	5 200	9 ^E 16	4 650	9 ^E 15
TNO	30	1 ^E 13	30	1 ^E 13
YK	25	2 ^E 12	25	2 ^E 12
CANADA	27 195	2,1^E17	24 455	1,9^E17

La formule de la méthode Hidiroglou-Berthelot compare la valeur de M1 à une certaine constante C1 tandis que la méthode du sigma-gap modifiée compare la valeur de M2 à une certaine constante C2. Les constantes C1 et C2 sont choisies en fonction de la sensibilité désirée. Pour qu'une observation soit identifiée valeur aberrante, il faut qu'elle soit détectée par les deux méthodes.

Le module de détection des valeurs aberrantes corrige aussi les unités jugées influentes. La méthode retenue est celle qui modifie le poids de sondage en guise de traitement pour les valeurs aberrantes en leur assignant un poids de un. Le poids résiduel sera redistribué aux autres unités de la même strate afin de préserver la représentativité de celle-ci. Toutes les données d'enquêtes qui appartiennent aux strates à tirage partiel sont testées afin de s'assurer qu'elles ne sont pas aberrantes. Dans les étapes subséquentes, i.e. le calage, le calcul de la variance, etc., le système fait en sorte que cette unité garde un poids de 1 comme poids final, de façon à ne représenter que lui-même. Le tableau 1 présente les estimations de totaux et de variance pour une industrie par province. Noter que pour préserver la confidentialité, ce ne sont pas les vraies estimations. Dans certaines provinces, telles l'Alberta, le Québec et l'Ontario, le niveau des estimations diminue de façon considérable. On remarque également que les estimations sont plus stables après la détection et la correction.

3.2 Calage

L'étape suivante consiste à calculer un facteur de calage ou ajustement qui, combiné avec le poids de sondage, produira le poids final pour chaque unité de l'échantillon selon la formule suivante:

$$w_i = a_{li} * g_{li}$$

où a_i est le poids de sondage et;
 g_i l'ajustements de poids ou facteurs de calage tel que

$$\hat{Y}_{pst} = \sum_{i \in S} g_{gm} w_h y_i$$

Dans le cadre de l'EUE, la technique de calage choisie est la post-stratification avec de nouveaux effectifs de la population. Il est absolument nécessaire de post-stratifier car un des objectifs de l'EUE est de produire des estimations au niveau de l'établissement.

Comme les poids de sondage (Parent, Simard, 2000) sont en fait des poids d'unités d'échantillonnage, post-stratifier avec des effectifs d'établissements représente le seul moyen d'obtenir des estimations de la population totale en nombre d'établissements. Conséquemment, toutes les unités de l'échantillon sont calées sur de nouveaux effectifs de population. Pour ce faire, un nouveau fichier de population, ou une image améliorée de la base de sondage, est créé représentant toujours la population cible. Ceci permet, entre autres, de corriger par un simple ajustement macro les erreurs de classification survenues à la sélection, i.e. le code industriel, le code provincial et les erreurs de taille selon le revenu (pour les sauts de strates). 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. Il est à noter que pour assurer des estimations sans biais, le nouveau fichier ne contient que des corrections indépendantes de l'enquête, qui ont été effectuées après le tirage de l'échantillon.

Tableau 2: Impact de la définition des groupes de calage sur les estimations.

Niveau de calage: Province	Province/NAICS3		Province/NAICS3/Cp	
	Estimation (\$millions)	CV (%)	Estimation (\$millions)	CV (%)
T-N	290	4,0	300	2,1
IPÉ	80	3,7	80	3,5
N-É	600	15,7	690	4,3
N-B	450	11,2	460	3,9
QC	4 640	14,5	5 250	3,1
ON	8 820	12,2	9 100	2,8
MN	700	23,7	900	5,9
SK	570	10,4	570	3,7
AL	2 570	16,4	2 400	4,5
C-B	3 560	28,1	4 650	6,3
TNO	30	13,6	30	11,5
YK	25	7,0	25	5,7
CANADA	22 335	7,5	24 455	1,8

La construction des groupes de calage, i.e. les post-strates, doivent obéir à certains critères établis pour l'EUE. Un nombre minimal de 30 unités sélectionnées et de 50 unités dans la population par groupe de calage est requis pour obtenir des estimations fiables et robustes. Le choix des niveaux de post-stratification n'est pas totalement libre pour l'EUE. Étant donné que les estimations sont sans biais au niveau de la post-strate, il est important que la province soit un des niveaux de post-stratification. On se rappelle qu'un des objectifs importants du PASEP est l'amélioration des statistiques provinciales. Il reste ensuite à choisir le niveau de détail du secteur industriel ainsi que de choisir si les unités à tirage complet (Simard, Hidiroglou, 1999) sont ajustées. Le tableau 2 présente l'impact sur les estimateurs de totaux et de variances provinciaux, de la définition des groupes de calage. Deux scénarios ont été présentés. Le premier consiste à définir le groupe de calage par province et par code industriel. Quant au deuxième, il ajoute un autre niveau dans la définition des groupes, celui de la structure des unités, i.e. complexe ou simple. On s'aperçoit que l'ajout de la structure amène une stabilité, i.e. une variance plus petite à l'estimation. Ici, l'hypothèse que la structure d'unité est un facteur d'homogénéité est vérifiée.

3.3 Variance de l'estimateur de post-stratification

Étant donné le réseautage (Simard, Hidiroglou, 1999) ainsi que la difficulté d'adaptation de ce module avec le plan de l'EUE 1997, plusieurs estimateurs de variance ont été testés. La variance de l'estimateur de post-stratification choisie en 1997 est celle calculée par Lehtonen-Pahkinen(1994) :

$$V(\hat{Y}) = \sum_{i=1}^l N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{S_i^2}{n_i}$$

Où n_i est le nombre d'unités de l'échantillon et N_i le nombre d'unité dans la population qui appartiennent à la post-strate l .

Celle-ci a été retenue pour sa simplicité. Pour les années subséquentes, des travaux de recherche ont été entrepris pour produire un estimateur de variance plus approprié.

3.4 Autres indicateurs de qualité

En plus de la variance, trois indicateurs de qualité sont produits pour chacune des estimations: le taux de réponse, le taux d'imputation ainsi que la fraction de réponse. Il faut mentionner que pour l'EUE, il y a deux types d'imputation : l'imputation partielle et l'imputation massive d'enregistrements. Le taux de réponse est une mesure de qualité spécifique à chaque variable ainsi que chaque domaine alors que le taux d'imputation est une mesure reliée au domaine seulement, ce qui implique qu'il est le même peu importe la variable estimée. La fraction de réponse représente la portion pondérée de l'estimateur qui provient des répondants seulement.

Taux de réponse =
$$\frac{\text{nombre de répondants pour une variable donnée dans un domaine donné}}{\text{nombre d'établissements échantillonnés dans le domaine}}$$

$$\text{Taux d'imputation} = 1 - \left(\frac{\text{nombre de répondants pour une variable donnée dans un domaine donné}}{\text{nombre d'établissements échantillonnés dans le domaine}} \right)$$

$$\text{Fraction de réponse} = \frac{\text{somme des poids des répondants pour une variable donnée dans un domaine donné}}{\text{nombre d'établissements estimés dans le domaine}}$$

Ces étapes terminent le processus d'estimation des enquêtes qui choisissent la stratégie avec un ajustement seulement. Les prochaines sections décrivent les autres étapes à compléter lorsqu'une enquête choisit la stratégie d'estimation avec deux ajustements, i.e. utilisant une approche à deux-phases.

4. SYSTÈME D'ESTIMATION POUR L'ESTIMATEUR À DEUX-PHASES

Cette section décrit les modules impliqués dans le cas de l'estimateur à deux-phases. Pour l'EUE 1997, une seule enquête a choisi cette stratégie pour la production de leurs estimations finales.

4.1 Détection des valeurs aberrantes pour la première et la deuxième phases

La détection des valeurs aberrantes est effectuée pour les deux échantillons, i.e., une fois avec les données pondérées de la première phase et une autre avec les données pondérées de la deuxième phase. Le module utilisé est celui décrit à la section 3.1. Toutes les observations qui proviennent des strates à tirage partiel subissent la détection des valeurs aberrantes. Dans un premier temps, la détection et la correction sont effectuées avec les données administratives. Puis, seules celles qui ne sont pas jugées influentes et qui ont été sélectionnées à la deuxième phase subiront une deuxième détection de valeurs aberrantes.

4.2 Calage de la première phase

Le calage de la première phase est en fait une post-stratification tel que décrit dans la section 3.2, mais le calage est effectué sur les enregistrements des données administratives au lieu d'être effectué sur les unités recevant des questionnaires. Par contre pour ce dernier, il est nécessaire de rajouter une autre contrainte pour la définition des post-strates : le niveau complexe-simple (Parent, Simard, 2000). Il faut mentionner que c'est seulement pour les entreprises simples que les données administratives sont disponibles au même niveau que les questionnaires. Ainsi, les données administratives pondérées de la première phase serviront d'information auxiliaire lors

du calage de la deuxième phase. Le choix du code industriel dépendra du nombre d'unités dans l'échantillon ainsi que dans la population (voir section 3.2). Le module de calage offre également la possibilité de caler les unités faisant partie des strates à tirage complet ou encore de ne post-stratifier que les unités à tirage partiel. Cette décision est laissée à la discrétion des analystes. Lors de ce premier calage, un premier ajustement est obtenu.

4.3 Concept de groupe-modèles

Les unités qui seront utilisées pour le calage de la deuxième phase sont les unités qui font partie des entreprises simples qui ont reçu un questionnaire et dont les données administratives sont également disponibles. Une relation, i.e. soit un ratio ou une régression est établie entre ces deux types de données pour chacune des unités appartenant à un groupe modèle donné (voir section 4.5). Dans la construction de ce dernier, une certaine homogénéité des données est recommandée. La province demeure un niveau obligatoire et ce sont en général les analystes qui regroupent les unités selon les différents codes industriels, car ils connaissent mieux les caractéristiques de leurs industries. Par exemple, un analyste peut choisir de grouper ensemble deux codes industriels à 5 chiffres parce qu'ils sont semblables et qu'il n'y a pas suffisamment d'unités dans l'échantillon ou dans la population (voir section 3.2).

4.4 Détection des valeurs aberrantes pour la régression

Le module de détection de valeurs aberrantes pour la régression a pour but d'assurer que certaines grosses unités n'ont pas d'influence sur le calcul des paramètres de régression. Les valeurs influentes ne sont pas retirées de la régression mais ont un impact minimum en leur assignant un poids de régression presque nul. Le système d'estimation permet d'utiliser jusqu'à 25 variables dépendantes pour la régression. Pour détecter les valeurs influentes, on utilise la distance de Cook qui est une des options de la procédure REG de SAS. Dans l'EUE, une unité est jugée influente si la distance de Cook est supérieure à 5.

4.5 Calage de la deuxième phase

Le calage de la deuxième phase rend l'estimateur plus efficace car il permet de réduire sa variance. L'ajustement de poids est calculé en se servant des totaux qui proviennent de la première phase, soit les données administratives. On utilise les groupes modèles définis précédemment. Le module produit les

ajustements pour chaque groupe-modèle ainsi que le poids final qui sera utilisé à l'estimation selon la formule suivante:

$$w_i = a_{1i} * g_{1i} * a_{2i} * g_{2i}$$

où les a_{1i} et a_{2i} sont les poids de sondage respectifs de phase 1 et phase 2 et les g_i , les ajustements de poids ou facteurs de calage respectifs des deux-phases

$$\hat{Y}_{deux-phases} = \sum_{i \in sample} g_{1-gm} w_{h-1} g_{2-gm} w_{g-2} y_i$$

4.6 Variance de l'estimateur à deux-phases

La variance de l'estimateur à deux-phases est donnée dans Arcaro (1997). Elle a été développée pour un des prototypes du système généralisé d'estimation.

$$\begin{aligned} Var\{\hat{Y}\} = & \sum_{h=1}^H N_h^2 (1-f_h) \frac{s_h^2}{n_h} + \\ & \sum_{h=1}^H \sum_{g=1}^G \frac{N_h^2 (1-f_h) M_g^2 (1-f_g) s_{1hg}^2}{n_h^2 (n_h - 1) m_g^2} + \\ & \sum_{g=1}^G M_g^2 (1-f_g) \frac{s_{2g}^2}{m_g} \end{aligned}$$

5. ESTIMATEUR DE POST-STRATIFICATION VS. ESTIMATEUR À DEUX-PHASES

Le choix entre les deux stratégies est basé sur deux critères essentiels pour pouvoir produire l'estimateur à deux-phases : un nombre minimal d'unités par groupe modèle ainsi qu'une bonne corrélation entre les données de questionnaires et les données administratives. Il faut un nombre suffisant d'unités dans le groupe modèle soit environ 30 unités. Si une de ces deux conditions n'est pas respectée, il est préférable d'utiliser l'estimateur de post-stratification. Il sera plus efficace dans le cas où la corrélation serait inférieure à 0.5 (pour estimateur par le quotient) et il sera plus fiable si le nombre d'unités est insuffisant dans les groupes modèles. Il est à noter que sans cette dernière contrainte, l'algorithme a de la difficulté à converger. Dans les mois qui suivent, plusieurs

travaux de recherche seront entrepris quant à la sélection de la stratégie d'estimation finale ainsi que pour celle du calcul de variance.

REMERCIEMENTS

Les auteurs tiennent à remercier Charlie Arcaro et Victor Estevao pour leur travail sur la variance de l'estimateur à deux-phases et le développement du prototype de GES (deux-phases).

RÉFÉRENCES

- Arcaro, C. (1997). Specification for the two-phase prototype estimation system. *Statistics Canada internal document*.
- Arcaro, C. (1997). Estimation for two-phase sampling with auxiliary information. *Statistics Canada internal document*.
- Deville, J.-C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, no 87, pp 376-382.
- Hidiroglou, M, Berthelot, J.-M. (1986). Statistical Edit and Imputation for Periodic Surveys, *Survey Methodology*, 12, pp 73-83.
- Hladky, M. (1998). Survey outlier detection and treatment for the Unified Enterprise Survey. *Statistics Canada internal document*.
- Lehtonen, R., Pahkinen E.J. (1994). Practical methods for design and analysis of complex surveys. Chichester. John Wiley and Sons, pp 95-99.
- Parent, M-N., Simard M. L'échantillonnage avec une approche unifiée : Le cas de l'Enquête Unifiée sur les Entreprises. SSC 2000, recueil.
- Simard, M., Hidiroglou, M. (1999). Estimation For Annual Business Surveys Based On Two-Phase Network Sampling,. *SSC Proceedings of the Survey methods Section*, pp 11-19.