Assemblée annuelle de la SSC, Mai 2014
Recueil de la Section des méthodes d'enquête

INTÉGRER LES CARACTÉRISTIQUES DES ÉTABLISSEMENTS LORS DU CALAGE D'UNE ENQUÊTE AUPRÈS D'ENTREPRISES

Naïma Gouzi[1]

RÉSUMÉ

Dans le contexte des enquêtes économiques, une entreprise peut avoir plusieurs entités de production appelées établissements. Chaque entité exploite un secteur d'activité dans une région géographique. L'objectif d'une enquête est d'estimer le total des variables d'intérêt associées à un ensemble spécifique de secteurs d'activité. Les enquêtes utilisent la même population d'entreprises; une entreprise peut être dans le champ de plusieurs enquêtes. Un grand échantillon d'entreprises est d'abord sélectionné pour toutes les enquêtes afin de corriger l'information auxiliaire comme la mauvaise classification. Ensuite, chaque enquête sélectionne son propre sous-échantillon. Les sous-échantillons peuvent se chevaucher. Ce document présente les enjeux de l'utilisation de l'information auxiliaire des établissements pour améliorer les estimations de chaque enquête.

MOTS CLÉS : Base unique; échantillon à deux phases; enquêtes chevauchantes; enquêtes-entreprises.

ABSTRACT

In the context of business surveys, an enterprise can have several production entities, called establishments. Each establishment works in an area of activity in a geographic region. The purpose of a survey is to estimate the total of the variables of interest associated with specific areas of activity. Surveys use the same population of enterprises; an enterprise can therefore be in the scope of several surveys. First, a large sample of enterprises is selected for all surveys to correct auxiliary information, such as misclassification. Each survey then selects its own subsample; subsamples can overlap. This document presents the issues surrounding the use of auxiliary information on establishments to improve each survey estimate.

KEY WORD: Single frame; Two phases sample; Overlapping surveys; Business survey.

1. INTRODUCTION

Statistique Canada a lancé le Programme intégré de la statistique des entreprises (PISE) dont l'objectif principal est l'élaboration d'un système général qui comprend la majorité des enquêtes économiques à Statistique Canada. Ce système vise à uniformiser la méthodologie et l'infrastructure de ces enquêtes. Par ce programme, Statistique Canada veut à la fois remanier l'Enquête unifiée auprès des entreprises (EUE), qui englobe plus de soixante enquêtes économiques annuelles, et l'élargir pour qu'elle puisse compter éventuellement environ 120 enquêtes-entreprises annuelles et infra-annuelles. Chaque enquête vise un ensemble de sous-populations où une sous-population est définie par les dimensions industrielles et géographiques. Le PISE n'utilise qu'une seule base de sondage stratifiée pour ces 60 enquêtes où une entreprise peut être dans le champ de plusieurs enquêtes économiques. Une entreprise est classifiée dans une seule strate selon l'information disponible dans la base de sondage. Un grand échantillon de première phase est sélectionné. Une fois que l'information disponible sur la base de sondage est validée auprès des unités échantillonnées, chaque enquête reclassifie les unités échantillonnées et procède à la sélection de son propre sous-échantillon de deuxième phase d'une façon indépendante des autres enquêtes. Pour les échantillons à deux phases, l'unité d'échantillonnage est l'entreprise.

L'objectif principal des enquêtes auprès des entreprises est d'estimer le total de variables d'intérêt au niveau d'établissements définis par une entité de production exploitée dans un secteur d'activité et une région géographique. L'objectif du présent projet est d'examiner l'intégration des variables auxiliaires dans l'estimation à l'aide du calage. Deux types de variables auxiliaires sont disponibles : les variables disponibles dans la base de sondage pour toutes les unités de la population et les variables observées pour l'échantillon de première phase. Le poids d'échantillonnage est alors multiplié par un facteur d'ajustement en fonction des caractéristiques des établissements qui forment l'entreprise. Puisque chaque variable auxiliaire peut se multiplier en une panoplie de variables de calage pour une entreprise, l'objectif est de déterminer un sous-ensemble de variables de calage à utiliser. Cette réduction est nécessaire étant donné la grande

[1] Naïma Gouzi, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Canada, K1A 0T6, Naima.Gouzi@statcan.gc.ca

dimension du vecteur des variables auxiliaires causée par les nombreux domaines représentant le secteur des activités et les régions géographiques. De plus, elle concerne l'incidence du choix des variables de calage sur les propriétés des estimateurs.

La section 2 donne un bref aperçu du plan de sondage du PISE. La section 3 fait état du calage pour un plan à deux phases où plusieurs enquêtes se partagent un échantillon de première phase. Finalement, dans la dernière section, les résultats d'une étude par simulation sont présentés. Ils comparent différentes options pour le calage où l'efficacité d'un estimateur calé est évaluée en termes d'erreur quadratique moyenne.

2. PLAN D'ENQUÊTE DU PISE

2.1 Population, variables et paramètres d'intérêt

Les enquêtes-entreprises ont en commun une base de sondage *F*, appelée Registre des entreprises, qui contient une liste de toutes les entreprises existantes au canada. Pour cette étude, il sera supposé que cette base est complète. Les paramètres d'intérêt de base pour le PISE sont principalement les totaux d'une caractéristique pour plusieurs sous-populations. Une sous-population est définie par la dimension industrielle et la dimension géographique. Supposons que nous avons N_a industries et N_g régions géographiques. Pour une caractéristique d'intérêt y, nous avons $N_a N_g$ paramètres d'intérêt à estimer, le vecteur des paramètres est donc $\boldsymbol{Y} = (Y_1,...,Y_I)^T$ où $I = N_a N_g$ est le nombre total de sous-populations. L'indice $i \equiv (ag)$, $i = 1,...,I$ indique une activité industrielle a dans une région g pour $a \in \mathbf{A}$ et $g \in \mathbf{G}$, où $\mathbf{A} = \{1,...,a,...N_a\}$ est l'ensemble des activités et $\mathbf{G} = \{1,...,g,...N_g\}$ est l'ensemble des régions géographiques. Si y représente le revenu de l'entreprise, alors la valeur $y_{i;k} = y_{ag;k}$ peut être vue comme le revenu d'un établissement ou d'un groupe d'établissements appartenant à l'entreprise k exploitant l'activité a dans la région g. Si l'entreprise ne pratique pas l'activité a dans la région g, alors $y_{i;k} = y_{ag;k} = 0$.

De la même façon que les variables d'intérêt, une variable auxiliaire disponible dans la base de sondage peut également être partitionnée en un vecteur de composantes associées aux activités et régions géographiques : $\boldsymbol{x}^{(F)} = (x_1^{(F)},...,x_i^{(F)},...,x_I^{(F)})^T$. Ainsi, la valeur de la variable auxiliaire $x_k^{(F)}$ pour l'entreprise k peut être décomposée comme suit : $x_k^{(F)} = \sum_i^I x_{i;k}^{(F)}$ où $x_{i,k}^{(F)}$ est la valeur de $x^{(F)}$ si l'entreprise k pratique l'activité a dans la région géographique g et $x_{i,k}^{(F)} = 0$ si ce n'est pas le cas.
Le total Y_i d'une sous-population i pour une variable d'intérêt y à être estimée par l'enquête e peut s'écrire sous la forme :

$$Y_{ei} = \sum_k y_{ei;k}, \; e = 1,...,N_e, \; i = 1,...,I_e \tag{2.1}$$

où $y_{ei;k}$ est la valeur de la variable d'intérêt pour l'entreprise k dans la sous-population i de l'enquête e, I_e est le nombre total de sous-populations de l'enquête e pour $e \in \mathbf{E}$, où $\mathbf{E} = \{1,...,e,...,N_e\}$ désigne l'ensemble des enquêtes et $\sum_k$ représente la sommation de toutes les entreprises de la population.

2.2 Plan d'échantillonnage

Pour simplifier le processus d'échantillonnage, le PISE utilise une approche à base unique stratifiée, où les strates sont définies par la dimension industrielle et la dimension géographique. Une entreprise, même complexe (c.-à-d. qui opère dans plusieurs secteurs d'activités ou dans plusieurs zones géographiques), apparaît dans une seule strate, soit dans une seule dimension industrielle et géographique. Chaque sous-population est traitée comme un domaine d'étude. Pour faciliter l'illustration du choix de la strate pour une entreprise complexe, il sera supposé que le nombre de strates H est égal au nombre de domaines I et qu'une entreprise k a I proportions $(p_{1k},..., p_{Ik})$, où p_{ik} est la proportion de l'entreprise k dans le domaine i. L'entreprise k peut être placée dans la strate h si sa proportion dans la strate h est la plus grande de toutes les proportions, soit $p_{hk} = \max(p_{1k},..., p_{Ik})$.

Le plan d'échantillonnage du PISE est un plan particulier à deux phases. La première phase sert principalement à mettre à jour l'information disponible sur la base de sondage. Les strates sont définies à partir des variables de classification disponibles sur la base de sondage. Un échantillon de Bernoulli stratifié $s^{(1)}$, dont l'unité d'échantillonnage est

l'entreprise, est prélevé. Une fois l'information auxiliaire mise à jour pour les unités de l'échantillon de première phase, le vecteur $x_k^{(F)} = (x_{1;k}^{(F)},...,x_{i;k}^{(F)},...,x_{I;k}^{(F)})^T$ est corrigé pour obtenir $x_k = (\boldsymbol{x}_{1;k},...,\boldsymbol{x}_{i;k},...,\boldsymbol{x}_{I;k})^T$ pour chaque $k \in s^{(1)}$. Ensuite, plusieurs bases de sondage sont formées à partir de l'échantillon de première phase, de telle sorte qu'une entreprise ayant plusieurs activités et donc appartenant à plusieurs enquêtes se voit placée dans plusieurs bases de sondage. Chaque enquête est indépendante des autres enquêtes. Elle place chaque entreprise dans une seule strate utilisant l'information corrigée et sélectionne un échantillon de Bernoulli stratifié.

2.3 Estimateurs

L'estimateur Horvitz-Thompson (HT) du total Y_{ei} donné par (2.1) pour l'enquête e est

$$\hat{Y}_{ei} = \Sigma_k \, d_{e;k} y_{ei;k} \, . \tag{2.2}$$

Le poids d'échantillonnage pour l'entreprise k appartenant à l'enquête e est donné par

$$d_{e;k} = d_k^{(1)} d_{e;k}^{(2|1)}$$

où $d_k^{(1)}$ est le poids de la première phase avec $d_k^{(1)} = 0$ si l'entreprise k n'est pas sélectionnée à la première phase, et où $d_{e;k}^{(2|1)}$ est le poids conditionnel de la deuxième phase de l'enquête e étant donné l'échantillon de la première phase. Toutefois, l'estimateur HT n'utilise pas l'information auxiliaire disponible.

3. CALAGE

L'estimation par calage est une méthode utilisée fréquemment dans les enquêtes par sondage pour intégrer de l'information auxiliaire sur la population ou sur une partie de la population. La méthode permet en particulier de construire des poids de calage pour n'importe quelle enquête complexe et fait coïncider l'estimateur calé du total d'une variable à partir de l'échantillon avec celui provenant d'une autre source considérée comme certaine. Par conséquent, cette méthode confère à l'estimateur la propriété d'assurer la cohérence entre les résultats et les totaux disponibles (Deville et Särndal, 1992). Aussi, la méthode améliore souvent l'efficacité de l'estimation (Särndal, 2007).

L'ensemble des poids de calage w_k pour un plan à une phase est obtenu en minimisant la fonction

$$\Sigma_{k \in s} \, d_k G(w_k / d_k) \tag{3.1}$$

tout en respectant des contraintes appelées équations de calage : les estimations de chaque variable auxiliaire doivent être calées aux totaux connus de cette variable auxiliaire, où d_k est le poids de sondage et $G(.)$ est une fonction de distance choisie au préalable. Le choix d'une fonction repose sur l'objectif à atteindre, par exemple soit minimiser la variabilité des poids calés, soit minimiser l'étendue de leur distribution, soit obtenir une certaine forme de distribution. Dans le cas de la forme linéaire, où $G(u) = (1/2)(u-1)^2$, le problème d'optimisation possède une solution analytique, le poids de calage devient $w_k = d_k g_k$, et le facteur d'ajustement g_k est donné par $g_k = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})^T \{\Sigma_k \, d_k \boldsymbol{x}_k \boldsymbol{x}_k^T\}^{-1} \boldsymbol{x}_k$, c'est-à-dire que l'estimateur du total de la population est l'estimateur par régression (GREG). Ici $\hat{\boldsymbol{X}}$ désigne l'estimateur HT du total connu $\boldsymbol{X}$.

Cependant, du point de vue pratique, le choix d'une fonction n'est pas sans conséquence. La forme linéaire conduit, dans certains cas, à des poids négatifs ou à des poids extrêmement grands, ce qui n'entre pas dans le cadre acceptable d'une enquête. Afin d'éviter l'obtention de poids aberrants, le poids de calage s'obtient en minimisant la fonction (3.1) sous la contrainte des équations de calage ainsi qu'en ajoutant d'autres contraintes sur les bornes ayant la forme $l_k \le g_k \le u_k$ pour tout $k \in s$ avec $l_k \le 1 \le u_k$. Le nombre de contraintes de calage à respecter a une influence directe sur la procédure du calage. En effet, dans certains cas, les calculs peuvent être fastidieux, et, dans d'autres, la solution n'existe tout simplement pas. Le cas échéant, afin d'obtenir une solution, les contraintes sur les bornes doivent être relâchées.

L'estimation par calage sous un plan à deux phases a aussi été examinée par plusieurs auteurs dans le contexte où deux sources d'information auxiliaire sont connues. Dans le cas de la population, le total $\boldsymbol{X}^{(F)} = \Sigma_k \, \boldsymbol{x}_k^{(F)}$ est connu où $\boldsymbol{x}_k^{(F)}$ est un vecteur connu pour chaque entreprise appartenant à l'échantillon de la première phase. Par conséquent, il est également connu pour chaque échantillon de deuxième phase de chaque enquête e, $e = 1,...,N_e$. Pour l'échantillon $s^{(1)}$ de première phase, le vecteur $\boldsymbol{x}_k$ est observé pour chaque entreprise de première phase et est par conséquent connu pour toutes les

entreprises de deuxième phase de chaque enquête. Le total $\boldsymbol{X} = \sum_k \boldsymbol{x}_k$ est inconnu mais peut être estimé à partir de l'échantillon de première phase à l'aide de l'estimateur HT $\hat{\boldsymbol{X}}^{(1)} = \sum_k d_k^{(1)} \boldsymbol{x}_k$.

En 1987, une adaptation de l'estimation par régression a été proposée par Särndal et Swensson, avec ses variantes dans le cadre du sondage à deux phases. L'approche du calage décrite dans Deville et Särndal (1992) a été appliquée par Dupont (1995) à l'échantillonnage à deux phases. Dupont (1995) a étudié différentes stratégies d'estimation selon deux configurations alternatives pour l'information auxiliaire, en reliant les deux approches possibles pour aborder le problème de calage : le modèle de régression et l'estimation par calage. L'utilisation conjointe de deux informations auxiliaires conduit à différentes stratégies d'utilisation dans le cadre du sondage dont il est question dans ce texte ou de tout autre sondage à deux phases. Une stratégie communément utilisée est constituée de deux étapes :

Étape 1 : Obtenir les poids de calage $w_k^{(1)}$ de première phase en minimisant la fonction

$$\sum_{k \in s^{(1)}} d_k^{(1)} G(w_k^{(1)} / d_k^{(1)})$$

sous la contrainte que $\sum_k w_k^{(1)} \boldsymbol{x}_k^{(F)} = \boldsymbol{X}^{(F)}$.

Étape 2 : Obtenir les poids de calage $w_{e;k}$ de deuxième phase pour chaque enquête e en minimisant la fonction

$$\sum_{k \in s_e} d_{e;k} G(w_{e;k} / d_{e;k})$$

sous la contrainte que $\sum_k w_{e;k} \boldsymbol{x}_{e;k} = \tilde{\boldsymbol{X}}_e^{(1)}$ où $\tilde{\boldsymbol{X}}_e^{(1)} = \sum_k w_k^{(1)} \boldsymbol{x}_{e;k}$.

L'estimateur de calage du total (2.1) est alors $\tilde{\tilde{Y}}_{ei} = \sum_k w_{e;k} y_{ei;k}$.

Dans le cas du PISE, plusieurs raisons font que plusieurs variables auxiliaires ne peuvent pas toutes être utilisées : a) un grand nombre de domaines de calage ont un petit effectif; b) le biais des estimateurs des totaux de population calculés en utilisant les poids de calage est faible si ces variables sont reliées à la variable d'intérêt, mais leur variance peut être élevée si l'on utilise un trop grand nombre de variables auxiliaires (Clark et Chambers, 2008); c) le problème d'optimisation devient complexe à résoudre et peut nécessiter beaucoup de temps de calcul informatique; et d) la solution peut tout simplement ne pas exister. Bankier (1990) a proposé une méthode pour calculer les estimateurs calés en vertu de laquelle on élimine certaines colonnes de la matrice de données auxiliaires afin de réduire le nombre de conditions de la matrice de produits croisés et d'éviter ainsi les situations indésirables comme des poids négatifs, des poids aberrants ou une dépendance linéaire exacte entre les colonnes. Au lieu de nous fonder sur la sélection d'un sous-ensemble de variables auxiliaires, nous nous intéressons plutôt à un choix d'agrégation de domaines de calage.

4. SIMULATION

Une étude par simulation est présentée ici pour à la fois illustrer la complexité du calage à deux phases du PISE et évaluer la performance des diverses combinaisons de variables de calage envisagées. Considérons le cas le plus simple de deux enquêtes $N_e = 2$. La population de la première enquête est de taille $N_1 = 1000$ entreprises et est générée à l'aide du modèle $y \sim N(\mu_y, \sigma_y^2)$, où $\mu_y = 100$ et $\sigma_y^2 = 50$. Les valeurs de la variable auxiliaire de la première enquête sont générées conditionnellement étant donné y à l'aide du modèle $x_1 \mid y \sim N(\mu_{x_1|y}, \sigma_{x_1|y}^2)$, où $\mu_{x_1|y} = \mu_{x_1} + (\sigma_{x_1} \rho_{x_1;y} / \sigma_y)(y - \mu_y)$, $\mu_{x_1} = 100$, $\sigma_{x_1/y}^2 = \sigma_{x_1}^2 (1 - \rho_{x_1;y}^2)$ et $\sigma_{x_1}^2 = 50$.

Le paramètre d'intérêt de cette simulation est le total de l'enquête 1, soit $Y = \sum_k y_k$. Afin de mesurer l'effet des autres enquêtes sur l'estimation du paramètre d'intérêt de la première enquête, la variable auxiliaire de la deuxième enquête de taille $N_2 = 1000$ est générée selon le modèle $x_2 \sim N(\mu_{x_2}, \sigma_{x_2}^2)$ pour les unités ne chevauchant pas la première enquête et selon le modèle $x_2 \mid x_1 \sim N(\mu_{x_2|x_1}, \sigma_{x_2|x_1}^2)$ pour les unités chevauchantes, où $\mu_{x_2} = 100$, $\sigma_{x_2}^2 = 50$, $\mu_{x_2|x_1} = \mu_{x_2} + (\sigma_{x_2} \rho_{x_2,x_1} / \sigma_{x_1})(x_1 - \mu_{x_1})$, et $\sigma_{x_2|x_1}^2 = \sigma_{x_2}^2 (1 - \rho_{x_2;x_1}^2)$. Au total, 54 populations ont été générées en variant le coefficient de variation $\rho_{y;x_1}$ entre y et x_1; le coefficient de variation $\rho_{x_2;x_1}$ entre x_1 et x_2; l'indice de chevauchement φ entre les deux enquêtes et le taux d'erreur de classification τ dans la base de sondage : $\{0{,}3;0{,}6;0{,}9\} \times \{0{,}3;0{,}6;0{,}9\} \times \{0;0{,}5;1\} \times \{0;0{,}10\}$. En cas d'erreur de classification, les

deux variables auxiliaires sont interchangées puisqu'une unité incorrectement classée pour une enquête se trouve forcément dans l'autre enquête. Deux strates correspondant aux deux enquêtes sont définies dans la base de sondage.

À partir de chaque population, $B=1000$ échantillons stratifiés de Bernoulli de première phase avec fraction d'échantillonnage égale à $f_1^{(1)} = f_2^{(1)} = 0{,}8$ ont été sélectionnés. Afin de corriger l'information pour les unités échantillonnées, les valeurs de x_1 et x_2 sont interchangées en cas d'erreur. Les unités appartenant à la première enquête sont identifiées. Ensuite, un sous-échantillon de Bernoulli est sélectionné à partir de chaque échantillon de première phase de la première enquête en utilisant une fraction d'échantillonnage égale à $f_1^{(2|1)} = 0{,}6$. Pour chaque échantillon, le GREG est utilisé pour obtenir les poids calés pour chaque phase. Le tableau 1 dresse toutes les combinaisons des totaux de variables auxiliaires utilisées pour le calage dans cette simulation.

Tableau 1 : Combinaisons des totaux de variables auxiliaires pour le calage

	Phase 1	**Phase 2**
Pas de calage	ϕ	
Tailles seulement	$\boldsymbol{N}^{(F)} = (N_1^{(F)}, N_2^{(F)})$	$\widetilde{N}_1^{(1)}$
Totaux seulement	$\boldsymbol{X}^{(F)} = (X_1^{(F)}, X_2^{(F)})$	$\widetilde{X}_1^{(1)}$
Tailles et totaux	$(\boldsymbol{N}^{(F)}, \boldsymbol{X}^{(F)})$	$\widetilde{N}_1^{(1)}, \widetilde{X}_1^{(1)}$
Tailles et agrégation des totaux	$(\boldsymbol{N}^{(F)}, (X_1^{(F)} + X_2^{(F)}))$	

Il y a 15 combinaisons de calage pour ce cas simple. Le nombre de combinaisons pour le PISE, qui englobe 11 000 domaines à la première phase et 2 300 domaines à la deuxième phase pour la plus grande enquête, est donc très important. Si l'on suppose que $\hat{\theta}$ désigne un estimateur quelconque du total de la population, $\hat{\theta}$, a été calculé à partir de chaque échantillon b ($b = 1,...,B$) de chaque population et de sa moyenne $\overline{\hat{\theta}} = \sum_b \hat{\theta}_b / B$, où $\hat{\theta}_b$ est la valeur de $\hat{\theta}$ de l'échantillon b. Le biais relatif et l'erreur quadratique moyenne (EQM) de l'estimateur $\hat{\theta}$ sont calculés respectivement comme $BR(\hat{\theta}) = (\overline{\hat{\theta}} - Y)/Y$ et $EQM(\hat{\theta}) = B^{-1}\sum_{b=1}^{B}(\hat{\theta}_b - Y)^2$.

Le graphique 1 présente le rapport de l'EQM, $EQM(\widetilde{Y})/EQM(\hat{Y})$ entre l'estimateur calé et l'estimateur HT dans le cas d'un chevauchement complet entre les deux enquêtes avec un taux d'erreur de 10 % dans la base de sondage. Chaque courbe représente une combinaison des deux coefficients de corrélation $\rho_{y;x_1}$ et $\rho_{x_2;x_1}$ (en tout neuf courbes) pour les 15 combinaisons de calage (l'axe des x). L'axe des y est le rapport des EQM. Le graphique révèle trois ensembles de courbes engendrées par le coefficient de corrélation $\rho_{y;x_1}$ qui diffèrent de façon frappante : 1) le premier ensemble de courbes, en rouge, est celui d'une corrélation faible $\rho_{y;x_1} = 0{,}3$; 2) le deuxième ensemble de courbes, en bleu, est celui d'une corrélation moyenne $\rho_{y;x_1} = 0{,}6$; et 3) le dernier ensemble de courbes, en vert, est celui d'une forte corrélation $\rho_{y;x_1} = 0{,}9$.

On peut facilement voir que lorsque l'on ne fait pas de calage à la première phase, le gain par rapport à l'estimateur de HT est minime (30 %) et cela peu importe la façon de caler à la deuxième phase. Utiliser seulement les tailles à la deuxième phase permet de gagner autour de 50 % en termes d'EQM par rapport à l'estimateur HT. Lorsque l'on ajoute les totaux, soit à la première phase ou à la deuxième phase, le gain se fait en fonction de la corrélation $\rho_{y;x_1}$.

Avec une corrélation faible, le gain est de 50 % si l'on ajoute les totaux, par contre le gain est de 45 % si l'on ajoute les totaux et les tailles. Avec une corrélation moyenne, le gain est de 41 % avec les totaux et de 37 % avec les totaux et les tailles. En présence d'une forte corrélation, on gagne presque 88 % dans les deux cas. Les mêmes conclusions sont observées lorsque les totaux sont agrégés.

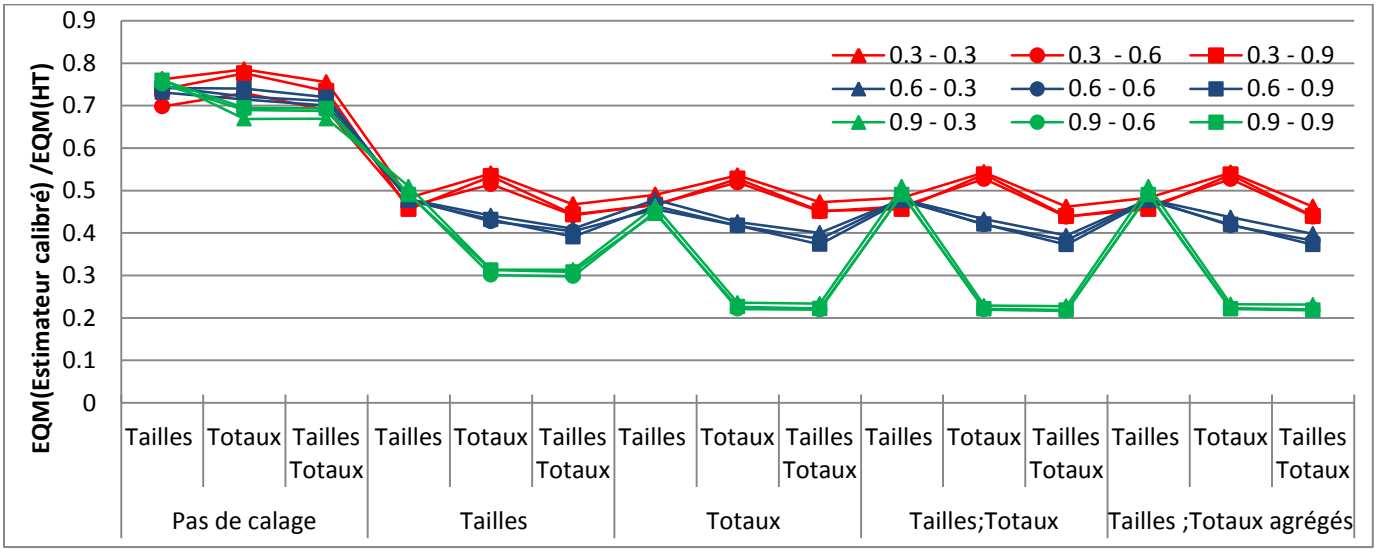
Des résultats similaires ont été obtenus pour les autres taux de chevauchement et les erreurs de classification dans la base de sondage.

DISCUSSION

Des études similaires ont été effectuées avec des données historiques. Les résultats des études empiriques confirment l'étude par simulation. Il est donc plus avantageux d'utiliser l'information auxiliaire lors d'un calage à deux phases. Agréger les totaux donne des conclusions similaires comparativement aux totaux sans agrégation. Cependant, les avantages de l'agrégation sont nombreux lorsque les variables de calage sont nombreuses.

Ces conclusions à l'appui, les estimations de la première phase du PISE ont été calées avec le revenu provenant des données des taxes distribuées au secteur d'activité et aux régions géographiques. Le même niveau d'agrégation établi à l'échantillonnage a été utilisé pour le secteur d'activité et les régions géographiques ont été regroupées en cinq régions. Le nombre de variables de calage à la première phase a été réduit de 11 000 à 1 000 variables. À la deuxième phase, les totaux de revenu agrégés au même niveau qu'à la première phase ont été utilisés et les tailles des strates ont été ajoutées. Pour la plus grande enquête, les variables de calage ont été réduites de 2 300 à 800 variables.

Graphique 1 : Rapport entre l'EQM de l'estimateur calé et l'estimateur HT
L'indice de chevauchement est à 100 % et le taux d'erreur dans la base est de 10 %



RÉFÉRENCES

Bankier, M.D. (1990). "Two Steps Generalized Least Squares Estimation". Ottawa: Statistique Canada. Division des méthodes d'enquêtes sociales, rapport interne.

Clark, R. G., et Chambers, R. L. (2008). « Calage adaptif pour la prédiction de totaux de population finie ». *Techniques d'enquête*. **34**, 181-192.

Deville, J.-C., et Särndal, C.-E. (1992). "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, **87**, 376-382.

Dupont, F. (1995). « Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire » *Techniques d'enquête*, **21**, 141-150.

Särndal, C.-E. (2007). « La méthode de calage dans la théorie et la pratique ». *Techniques d'enquête*, **33**, 113-135.

Särndal, C.-E. et Swensson, B. (1987). "A General View of Estimation for Two Phases of Selection with Applications to Two-Phases Sampling and Nonresponse". *International Statistical Review*, **55**, 279-294.