Techniques d'enquête, juin 1985
Vol. 11, no. 1, pp. 73–85
Statistique Canada

Étude empirique de quelques estimateurs de régression pour petits domaines

M.A. HIDIROGLOU et C.E. SÄRNDAL[1]

RÉSUMÉ

La méthode classique d'estimation des caractéristiques d'un petit domaine d'étude est fondée sur l'utilisation de l'estimateur synthétique (SYN). L'avantage de cet estimateur est que sa variance est faible; par contre, il peut être très biaisé dans certains petits domaines dont la structure est différente de celle d'un ensemble de domaines. Särndal (1981) a proposé l'estimateur de régression (RÉG) pour le calcul d'estimations relatives à des domaines. Cet estimateur ne renferme presque aucun biais, mais il présente deux inconvénients: (i) sa variance peut être considérable dans certains petits domaines et (ii) il peut avoir des valeurs négatives dans certains cas où on ne peut pas en admettre.

Dans cette étude, nous proposons un estimateur intermédiaire qui représente un compromis entre les estimateurs SYN et RÉG. Cet estimateur, qu'on appelle "estimateur de régression corrigé" (ERC), a une variance beaucoup moins élevée que celle de l'estimateur RÉG et une erreur quadratique moyenne plus faible que l'estimateur SYN dans les domaines d'étude où ce dernier est extrêmement biaisé. L'ERC ne pose pas le problème des valeurs négatives mentionné plus haut. Ces propriétés sont vérifiées par une étude de Monte Carlo portant sur 500 échantillons.

MOTS CLÉS: Petits domaines; estimateur de régression; estimateur de régression corrigé; biais; erreur quadratique moyenne.

1. INTRODUCTION

L'avantage de l'estimateur synthétique (SYN) est sa faible variance; cependant, cet estimateur comporte les inconvénients suivants: a) il peut être extrêmement biaisé dans certains domaines d'étude et, en général, on ne peut déterminer les domaines dont il s'agit; b) par conséquent, un coefficient de variation (cv) ou un intervalle de confiance calculé pour ces domaines est inutile.

À l'aide du même modèle qui est à la base de l'estimateur SYN, on peut créer un estimateur analogue presque sans biais, l'estimateur de régression généralisé (RÉG), qui permet en outre de calculer, pour chaque estimation relative à un domaine, un intervalle de confiance selon la méthode classique fondée sur le plan de sondage. Un des inconvénients de l'estimateur RÉG, c'est que sa variance estimée (et donc son cv et l'étendue de son intervalle de confiance) peut être beaucoup trop élevée dans des domaines très petits (ce qui, bien entendu, découle directement du nombre insuffisant d'observations dans ces domaines). Par ailleurs, l'estimateur RÉG peut, bien que la probabilité de cette éventualité soit faible, avoir des valeurs négatives dans des cas où on ne peut pas en admettre.

Il paraît donc souhaitable de trouver le juste milieu entre l'estimateur SYN et l'estimateur RÉG. Nous présentons ici une étude empirique des propriétés d'un estimateur intermédiaire, l'estimateur de régression corrigé (ERC). Il renferme un biais faible (mais notable) dans les domaines où l'estimateur SYN est extrêmement biaisé; dans les autres domaines, l'ERC ne comporte presque aucun biais. Un autre avantage de l'ERC est que sa variance est beaucoup

[1] M.A. Hidiroglou, Division des méthodes d'enquête-entreprises, 5-C8, Immeuble Jean talon, Parc Tuney, Ottawa (Ontario) Canada K1A 0T6, et C.E. Särndal, Département de Mathématiques et Statistique, Université de Montréal, Montréal (Québec), Canada H3C 3J7.

moins élevée que celle de l'estimateur RÉG. En outre, l'erreur quadratique moyenne de l'ERC est inférieure à celle de l'estimateur SYN dans les domaines d'étude où ce dernier est extrêmement biaisé. Il est également facile d'établir des intervalles de confiance raisonnables pour le nouvel ERC.

Cet article se compose de cinq sections. La section 2 décrit quelques-uns des estimateurs qui sont utilisés le plus souvent pour le calcul d'estimations relatives à des petits domaines, notamment les estimateurs directs, les estimateurs pour les domaines stratifiés a posteriori et les estimateurs synthétiques, et quelques-uns des estimateurs de régression proposés par Särndal (1981, 1984). La section 3 présente des estimateurs de régression corrigés et en évalue les avantages et les inconvénients. Dans la section 4, on examine les propriétés des estimateurs de régression corrigés et de certains autres estimateurs à l'aide d'une simulation de Monte Carlo faite à partir de données fiscales des entreprises. Enfin, la section 5 présente quelques conclusions générales.

2. ESTIMATEURS

Soit une population $U = \{1, \ldots, k, \ldots, N\}$ divisée en D domaines d'étude distincts $U_{1.}$, $\ldots$, $U_{d.}$, $\ldots$, $U_{D.}$. Soit $N_{d.}$ la taille de $U_{d.}$. (Dans notre étude empirique, les domaines sont définis en fonction de 4 catégories d'activités économiques et des 18 divisions de recensement de la province de la Nouvelle-Écosse. Il y avait $D = 70$ domaines non vides, comme l'ont décrit Hidiroglou, Morry, Dagum, Rao et Särndal (1984).)

La population se divise également selon une deuxième dimension en G groupes distincts, $U_{.1}$, $\ldots$, $U_{.g}$, $\ldots$, $U_{.G}$.

La taille de $U_{.g}$ est représentée par $N_{.g}$. (Dans notre étude, les groupes correspondent à des tranches de revenu d'entreprise brut.) Si on classe les unités de la population selon le domaine et le groupe, on obtient DG catégories U_{dg}; $d = 1, \ldots, D$; $g = 1, \ldots, G$. Soit N_{dg} la taille de U_{dg}.

La taille de la population, N, peut alors être exprimée de la manière suivante:

$$N = \sum_{d=1}^{D} N_{d.} = \sum_{g=1}^{G} N_{.g} = \sum_{d=1}^{D} \sum_{g=1}^{G} N_{dg} \tag{2.1}$$

Soit s un échantillon de taille n prélevé dans U par échantillonnage aléatoire simple (éas). Définissons $s_{d.}$, $s_{.g}$ et s_{dg} comme les parties de s qui appartiennent, respectivement, à $U_{d.}$, $U_{.g}$ et U_{dg}.

Les tailles correspondantes, qui constituent des variables aléatoires, sont représentées par $n_{d.}$, $n_{.g}$ et n_{dg}. À noter que l'équation s'applique également aux n minuscules. La variable qu'on veut étudier, y (= salaires et traitements), a la valeur y_k pour la k^{ieme} unité (= déclaration fiscale d'une entreprise non constituée en société). La variable auxiliaire x (= revenu d'entreprise brut) vaut x_k pour la k^{ieme} unité et la valeur x_k est connue pour tous les $k = 1, \ldots, N$.

On compare ici les estimateurs suivants du total pour l'ensemble d'un domaine, $t_d = \sum_{U_{d.}} y_k$ où $\sum_{U_{d.}}$ représente la sommation pour toutes les unités de $U_{d.}$.

L'estimateur direct par extension (EXT) est:

$$\hat{t}_{d\text{EXT}} = \frac{N}{n} \sum_{s_{d.}} y_k \tag{2.2}$$

L'estimateur pour domaines stratifiés a posteriori (POS) est:

$$\hat{t}_{d\text{POS}} = N_{d.} \bar{y}_{s_{d.}} \tag{2.3}$$

où

$$\bar{y}_{s_{d.}} = \sum_{s_{d.}} \frac{y_k}{n_{d.}}$$

est la moyenne des $n_{d.}$ valeurs de y dans le domaine d. Si $n_{d.} = 0$, nous définissons l'estimateur POS comme nul (ce qui est assez arbitraire puisque, à proprement parler, cet estimateur est alors indéterminé). Ni l'estimateur EXT ni l'estimateur POS ne sont particulièrement avantageux. Ils servent surtout de repères pour comparer les propriétés d'autres estimateurs plus efficaces qu'on présente plus bas.

Deux formes des estimateurs SYN et RÉG ont été examinées, l'une pour le calcul de totaux, l'autre pour le calcul de ratios. L'estimateur SYN repose sur l'hypothèse selon laquelle un modèle donné s'applique à chaque groupe g. Pour le calcul de totaux, il est supposé que la moyenne de chaque groupe est identique dans tous les domaines d. Pour le calcul de ratios, il est supposé que le ratio entre une variable donnée et une variable auxiliaire est constant à l'intérieur d'un groupe particulier dans tous les domaines. Si cette hypothèse d'homogénéité des caractéristiques relatives aux domaines n'est pas exacte pour chaque groupe, les estimateurs SYN peuvent être extrêmement biaisés. L'estimateur RÉG proposé par Särndal (1984) permet a) de produire des estimations qui ne renferment presque aucun biais par rapport au plan de sondage et dont la variance est simple à estimer et les intervalles de confiance faciles à calculer (et raisonnables); b) de renforcer les estimations par l'inclusion de données de l'échantillon de tous les domaines.

Les formules pour le calcul de totaux à base de comptes sont:

Estimateur synthétique à base de comptes (SYN/C):

$$\hat{t}_{d\text{SYN/C}} = \sum_{g=1}^{G} N_{dg} \bar{y}_{s_{.g}} \tag{2.4}$$

où $\bar{y}_{s_{.g}}$ est la moyenne de y dans $s_{.g}$.

Estimateur de régression à base de comptes (RÉG/C):

$$\hat{t}_{d\text{RÉG/C}} = \sum_{g=1}^{G} \{N_{dg} \bar{y}_{s_{.g}} + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\} \tag{2.5}$$

où $\bar{y}_{s_{dg}}$ est la moyenne de y dans s_{dg}, et $\hat{N}_{dg} = Nn_{dg}/n$. Dans cette formule, $\sum_{g=1}^{G} \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})$ est un terme qui corrige le biais de l'estimation et entraîne en général un accroissement considérable de la variance.

Pour le calcul à base de rapports, les formules des estimateurs SYN et RÉG sont:

Estimateur synthétique à base de rapports (SYN/R):

$$\hat{t}_{d\text{SYN/R}} = \sum_{g=1}^{G} X_{dg} \hat{R}_g \tag{2.6}$$

où $X_{dg} = \sum_{U_{dg}} x_k$ et

$$\hat{R}_g = \frac{\sum_{s_{.g}} y_k}{\sum_{s_{.g}} x_k}$$

Estimateur de régression à base de rapports (RÉG/R):

$$\hat{t}_{d\text{RÉG/R}} = \sum_{g=1}^{G} \{X_{dg} \hat{R}_g + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g \bar{x}_{s_{dg}})\} \tag{2.7}$$

3. ESTIMATEURS DE RÉGRESSION CORRIGÉS

Pour obtenir les estimateurs de régression présentés par Särndal (1984), on ajuste un modèle de régression à des variables auxiliaires et on calcule les valeurs prévues par le modèle pour les unités de chaque domaine de la population. Pour un plan de sondage arbitraire P (qui n'est pas nécessairement fondé sur l'éas) comportant des probabilités de sélection π_k (ordre premier) et $\pi_{k\ell}$ (ordre deuxième), soit le modèle de régression

$$E_\xi(y_k) = \underset{\sim}{x}_k'\beta;\ V_\xi(y_k) = v_k$$

où les y_k sont des variables aléatoires indépendantes. Un estimateur de β est

$$\hat{\underset{\sim}{\beta}} = \left(\sum_s \frac{\underset{\sim}{x}_k'\underset{\sim}{x}_k}{v_k\pi_k}\right)^{-1} \sum_s \frac{\underset{\sim}{x}_k'y_k}{v_k\pi_k}$$

où on suppose que les valeurs des v_k sont connues à un coefficient constant près et qu'on peut éliminer ce coefficient quand on calcule $\hat{\beta}$.

Selon la méthode utilisée par Särndal (1984), la formule suivante produit un estimateur presque sans biais pour un total inconnu dans le domaine d:

$$\hat{t}_{d\text{RÉG}} = \sum_{U_d.} \hat{y}_k + \sum_{s_d.} \frac{e_k}{\pi_k} \tag{3.1}$$

où $\hat{y}_k = \underset{\sim}{x}_k'\hat{\beta}$ est la k^{ieme} valeur prévue et $e_k = y_k - \hat{y}_k$ représente la k^{ieme} erreur résiduelle.

Nous appellerons $\sum_{U_d.} \hat{y}_k$ le *terme synthétique* de l'estimateur $\hat{t}_{d\text{RÉG}}$ et le deuxième terme, $\sum_{s_d.} e_k/\pi_k$, le *terme correctif*.

Si $s_d.$ n'est pas un ensemble vide, on peut substituer à l'estimateur RÉG (équation 3.1) un estimateur approximativement non biaisé:

$$\hat{t}_{d\text{SUB}} = \sum_{U_d.} \hat{y}_k + N_d. \frac{\sum_{s_d.} \frac{e_k}{\pi_k}}{\hat{N}_d.} \tag{3.2}$$

où

$$\hat{N}_d. = \sum_{s_d.} \frac{1}{\pi_k}$$

est l'estimation de la taille du domaine d.

Le terme correctif peut maintenant être exprimé sous forme d'estimation par le quotient,

$$\frac{\sum_{s_d.} \frac{e_k}{\pi_k}}{\sum_{s_d.} \frac{1}{\pi_k}},$$

que l'on multiplie par la valeur connue de la taille du domaine d, $N_d.$ (bien entendu, $N_d.$ est connu puisque les totaux N_{dg} sont connus).

Comme la valeur de la taille $n_d.$ est aléatoire, l'estimation par le quotient permet de réduire la variance du terme correctif. Cette modification aura un effet particulièrement notable dans les domaines où la moyenne des résidus est nettement non nulle (c'est-à-dire dans les domaines où l'ajustement du modèle est médiocre).

Si le rendement prévu de l'échantillon dans le domaine d, $E_d = E_p(n_{d.}) = \sum_{U_{d.}} \pi_k$, est élevé (par exemple si $E_d \geq 50$), il est alors pratiquement certain que le rendement réalisé, $n_{d.}$, ne sera pas excessivement faible. Par exemple, dans l'éas, il est très rare d'obtenir des valeurs de $n_{d.} \leq 30$. Dans ces cas, on peut recommander l'estimateur presque non biaisé (équation 3.2) sans aucune modification. Il devrait être beaucoup plus efficace que l'estimateur (3.1), notamment dans les domaines où l'ajustement du modèle n'est pas très bon. Toutefois, en pratique, il arrive souvent que certains domaines sont si petits que le rendement prévu de l'échantillon, E_d, ne dépasse pas 5. C'est ce qui s'est produit dans quelques-uns des domaines définis pour notre étude empirique. Dans ces cas, il est très probable que le rendement réalisé, $n_{d.}$, oscille entre 0 et 5. Notre travail empirique a confirmé, ce qui est intuitivement évident, le fait que la correction des résidus pour ces petits domaines accroît beaucoup la variance, que cette correction soit exprimée sous forme directe, $\sum_{s_{d.}} e_k/\pi_k$, comme dans l'équation (3.1), ou sous forme de quotient, $N_{d.}(\sum_{s_{d.}} e_k/\pi_k)/(\sum_{s_{d.}} 1/\pi_k)$, comme dans l'équation (3.2).

Pour réduire cet accroissement de la variance, nous modifions le terme correctif de l'équation (3.2) d'une manière qui équivaut à accepter un biais faible (dans les domaines où l'ajustement du modèle n'est pas très bon) pour baisser l'accroissement de la variance quand le rendement de l'échantillon, $n_{d.}$, est inférieur au rendement prévu (et on suppose également que le rendement prévu de l'échantillon est déjà faible à l'origine).

La forme du nouveau terme correctif est liée au rapport entre le rendement de l'échantillon, $n_{d.}$, et le rendement prévu, E_d. Le terme correctif $\sum_{s_{d.}} e_k/\pi_k$ sera multiplié par $(\hat{N}_d/N_d)$ quand $n_{d.} < E_d$ et par $(N_d/\hat{N}_d)$ autrement. Le terme correctif qu'on obtient avec cet "amortisseur" adaptable aura pour effet de ne pas "surcorriger" le terme synthétique si quelques-unes des erreurs résiduelles, e_k, ont des valeurs extrêmes quand $n_{d.}$ est faible. À cause de cette "surcorrection", on peut parfois sous-estimer de beaucoup le total pour un domaine d, obtenir des valeurs négatives lorsque seules des valeurs positives sont acceptables ou, à l'inverse, surestimer de beaucoup le total pour un domaine.

On obtient donc l'estimateur de régression corrigé (ERC), dont la définition varie selon que $n_{d.}$ est inférieur ou non à E_d:

$$\hat{t}_{d\text{ERC}} = \sum_{U_{d.}} \hat{y}_k + F_d \sum_{s_{d.}} \frac{e_k}{\pi_k} \tag{3.3}$$

où

$$F_d = \begin{cases} \dfrac{N_{d.}}{\hat{N}_{d.}} \text{ quand } n_{d.} \geq E_d \\ \dfrac{\hat{N}_{d.}}{N_{d.}} \text{ quand } n_{d.} < E_d \end{cases}$$

On peut démontrer que l'estimateur (3.3) est conditionnellement presque sans biais par rapport à $n_{d.}$, pourvu que $n_{d.} \geq E_d$. Si $n_{d.} \leq E_d$, l'ERC renferme un biais conditionnel qui tend à augmenter plus $n_{d.}$ est au-dessous de sa valeur prévue. En même temps, l'ERC est poussé vers la valeur du terme synthétique, ce qui offre une certaine stabilité (variance faible). Globalement, l'ERC défini par l'équation (3.3) renferme un léger biais, mais a une variance beaucoup moins élevée que celle de l'estimateur RÉG.

Nous soulignerons un dernier avantage de l'ERC. À cause de sa variance considérable dans les domaines très petits, l'estimateur RÉG aura, avec une probabilité faible mais positive, des valeurs qui se situent extrêmement loin de la vraie valeur $t_{d.}$. La valeur de l'estimateur RÉG peut même être négative, ce qui, naturellement, est inacceptable pour une variable (telle que les salaires et traitements) qui est par définition non négative. L'estimateur RÉG peut être négatif quand il existe d'importants résidus négatifs, e_k, dans le terme correctif de

l'équation (3.1), et la probabilité d'une valeur négative est particulièrement élevée quand $n_{d.} < E_d$. Avec le nouvel ERC, la possibilité d'obtenir des valeurs négatives est presque nulle. En pratique, si par un hasard peu probable une valeur négative est calculée pour l'ERC, nous recommandons que l'ERC soit redéfini comme étant égal à l'estimateur SYN qui est toujours positif.

Une formule naturelle pour estimer la variance de l'estimateur (3.2) est

$$\hat{V}_p(\hat{t}_{d\text{SUB}}) = \left(\frac{N_{d.}}{\hat{N}_{d.}}\right)^2 \sum_{\substack{k \neq \ell \\ \epsilon s_{d.}}} \sum \Delta_{k\ell} \frac{(e_k - \bar{e}_{s_{d.}})(e_\ell - \bar{e}_{s_{d.}})}{\pi_k \pi_\ell} \tag{3.4}$$

où

$$\bar{e}_{s_{d.}} = \frac{\sum_{s_{d.}} \frac{e_k}{\pi_k}}{\sum_{s_{d.}} \frac{1}{\pi_k}}$$

et

$$\Delta_{k\ell} = \begin{cases} 1 - \pi_k & \text{si } \ell = k \\ 1 - \dfrac{\pi_k \pi_\ell}{\pi_{k\ell}} & \text{si } \ell \neq k. \end{cases}$$

Nous soutenons que cette même formule peut bien servir à estimer la variance de l'ERC (3.3). Il est vrai que l'estimateur (3.3) est différent de l'estimateur (3.2) quand le rendement de l'échantillon est inférieur au rendement prévu; toutefois, il est peu probable que la différence entre ces deux estimateurs puisse être assez grande pour miner gravement la validité d'un intervalle de confiance de t_d centré sur $\hat{t}_{d\text{ERC}}$ à partir de la variance estimée par l'équation (3.4).

Pour un échantillon aléatoire simple dans lequel, pour $g = 1, \ldots, G$,

$$E_\xi(y_k) = \beta_g;\ V_\xi(y_k) = \sigma_g^2;\ k \in U_{.g}, \tag{3.5}$$

on obtient la formule

$$\hat{\beta}_g = \frac{\sum_{s_{.g}} y_k}{n_{.g}} = \bar{y}_{s_{.g}},$$

qui permet de définir l'estimateur corrigé à base de comptes (ERC/C)

$$\hat{t}_{d\text{ERC/C}} = \sum_{g=1}^{G} \{N_{dg} \bar{y}_{s_{g.}} + F_d \hat{N}_{dg} (\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\} \tag{3.6}$$

où E_d dans la formule pour F_d est maintenant

$$E_d = E_{eas}(n_{d.}) = \frac{nN_{d.}}{N}$$

et

$$\hat{N}_{dg} = n_{dg}\left(\frac{N}{n}\right)$$

où

$$\bar{y}_{s_{dg}} = \begin{cases} \dfrac{\sum_{s_{dg}} y_k}{n_{dg}} & \text{pour } n_{dg} \geq 1 \\ 0 & \text{autrement.} \end{cases}$$

L'ERC/C comportera un certain biais qui, cependant, est généralement beaucoup moins élevé que celui de l'estimateur SYN/C.

Les hypothèses sur lesquelles l'estimateur corrigé pour le calcul de ratios (ERC/R) est fondé sont, pour $g = 1, \ldots, G$,

$$E_\xi(y_k) = \beta_g x_k;\ V_\xi(y_k) = \sigma_g^2 x_k,\ k \in U_{.g}.$$

L'ERC/R est donc, dans le cas de l'échantillonnage aléatoire simple,

$$\hat{t}_{d\text{ERC/R}} = \sum_{g=1}^{G} \{X_{dg}\hat{R}_g + F_d\hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g\bar{x}_{s_{dg}})\} \tag{3.7}$$

où

$$\hat{R}_g = \frac{\sum_{d=1}^{D} \hat{N}_{dg}\bar{y}_{s_{dg}}}{\sum_{d=1}^{D} \hat{N}_{dg}\bar{x}_{s_{dg}}},$$

et

$$X_{dg} = \sum_{U_{dg}} x_k.$$

Drew, Singh et Choudhry (1982) ont élaboré des estimateurs pour petits domaines qui ressemblent aux estimateurs décrits dans cette étude. Leur estimateur à base de comptes est

$$\hat{t}_{d\text{KNO/C}} = \sum_{g} N_{dg}\{W'_{dg}\bar{y}_{s_{dg}} + (1 - W'_{dg})\bar{y}_{s.g}\} \tag{3.8}$$

et leur estimateur à base de rapports est

$$\hat{t}_{d\text{KNO/R}} = \sum_{g} X_{dg}\left\{W'_{dg}\frac{\bar{y}_{s_{dg}}}{\bar{x}_{s_{dg}}} + (1 - W'_{dg})\frac{\bar{y}_{s.g}}{\bar{x}_{s.g}}\right\} \tag{3.9}$$

où

$$W'_{dg} = \begin{cases} \dfrac{n_{dg}}{E_{dg}} \text{ si } n_{dg} \leq E_{dg} \\ 1 \quad \text{autrement} \end{cases}$$

et $E_{dg} = n(N_{dg}/N)$. Dans le contexte de la présente étude, si on remplace W'_{dg} dans les équations (3.8) et (3.9) par

$$W''_{dg} = \begin{cases} \left(\dfrac{n_{d.}}{E_d}\right)\left(\dfrac{n_{dg}}{E_{dg}}\right) & \text{si } n_{d.} < E_d \\ \left(\dfrac{E_d}{n_{d.}}\right)\left(\dfrac{n_{dg}}{E_{dg}}\right) & \text{si } n_{d.} \geq E_d \end{cases}$$

on obtient $\hat{t}_{d\text{ERC/C}}$.

4. RÉSULTATS DE L'ÉTUDE EMPIRIQUE

Pour étudier les propriétés des estimateurs décrits dans les sections précédentes, nous avons entrepris une simulation. La province de la Nouvelle-Écosse a été choisie comme univers et la population comprend $N = 1{,}678$ unités (déclarations fiscales d'entreprises non constituées en société). La variable qu'on veut analyser est y, les salaires et traitements. Nous utilisons une seule variable auxiliaire, x, le revenu d'entreprise brut. On suppose que les valeurs de $x_1, \ldots, x_N$ sont connues.

La population a été classée en domaines par rapport à 4 secteurs d'activité économique et à 18 régions. Les secteurs d'activité économique sont: commerce de détail (515 unités), construction (496 unités), hébergement (114 unités) et autres activités économiques (553 unités). Les valeurs des coefficients de corrélation globaux entre les salaires et traitements et le revenu d'entreprise brut étaient 0.42 pour le commerce de détail, 0.64 pour la construction, 0.78 pour l'hébergement et 0.61 pour les autres activités économiques. Les régions sont les 18 divisions de recensement de la Nouvelle-Écosse. On a obtenu 70 domaines non vides (il y avait 4 fois 18 domaines possibles, mais 2 ne comprenaient aucune unité). Ainsi, il faut estimer un total, t_d, pour chacun des 70 domaines chaque fois qu'un échantillon est prélevé.

Pour la simulation de Monte Carlo, 500 échantillons aléatoires simples, s, de $n = 419$ unités chacun ont été tirés de la population de $N = 1{,}678$ unités. Les unités choisies pour l'échantillon ont été classées selon le secteur d'activité économique et la division de recensement. La population aurait également pu être divisée en fonction d'une deuxième dimension, par exemple selon la tranche de revenu. Mais pour les besoins de cette étude, on a supposé que tous les déclarants fiscaux sont compris dans une seule tranche de revenu ($G = 1$).

Les résultats sont résumés pour chaque petite région dans les secteurs d'activité économique COMMERCE DE DÉTAIL et HÉBERGEMENT sous forme de tableaux et de graphiques. Les tableaux 1 à 4 présentent le biais conditionnel relatif et l'erreur quadratique moyenne. La figure 1 est composée de 8 graphiques, dont un pour chacun des 8 estimateurs étudiés. Chaque graphique contient une "bande de distribution" verticale pour chacune des 18 divisions de recensement dans le secteur d'activité COMMERCE DE DÉTAIL. Les points maximur et minimum de chaque bande de distribution correspondent respectivement au 90^{ieme} et au 10^{ieme} centile de la distribution des 500 valeurs de $(\hat{t}_{d.} - t_{d.})/t_{d.}$. Par conséquent, quand une bande de distribution est à peu près centrée sur le niveau horizontal 0, il s'ensuit que l'estimateur correspondant est presque sans biais pour le domaine en question, autrement l'estimateur est biaisé pour ce domaine. Plus une bande est courte, plus la variance de l'estimateur est faible dans un domaine.

Les tableaux et les graphiques permettent de tirer les conclusions suivantes (la conclusion C résume les résultats de cette étude et les conclusions A et B reposent sur les travaux de Särndal et Råbäck (1983) et Hidiroglou et coll. (1984)):

A. Les estimateurs SYN/C et SYN/R sont extrêmement biaisés dans certains domaines, notamment ceux où le modèle implicite ne s'ajuste pas bien. Toutefois, un de leurs avantages est le fait que leur variance est toujours faible par rapport à celle des autres estimateurs. L'erreur quadratique moyenne des deux estimateurs SYN est donc très élevée dans les domaines où le biais est important (où la qualité de l'ajustement du modèle est mauvaise); par contre, leur erreur quadratique moyenne est faible dans les domaines où il y a peu de biais (où le modèle s'ajuste bien aux données).

B. Les estimateurs RÉG/C et RÉG/R sont essentiellement non biaisés. Leur variance est généralement moins élevée que celle des estimateurs EXT et POS, mais elle est toujours beaucoup plus élevée que celle des estimateurs SYN/C et SYN/R.

Tableau 1

Rendement moyen de l'échantillon et biais relatif de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population Secteur d'activité: COMMERCE DE DÉTAIL. Régions: 18 divisions de recensement en Nouvelle-Écosse.

Région	Rendement moyen	Estimateur EXT	POS	SYN/C	ERC/C	RÉG/C	SYN/R	ERC/R	RÉG/R
1	1.76	−0.02	−0.13	0.12	0.02	−0.03	0.30	0.09	−0.02
2	5.45	0.00	−0.04	−0.36	−0.10	−0.02	−0.27	−0.08	−0.02
3	3.90	−0.02	0.01	−0.08	−0.02	0.00	−0.01	−0.01	0.00
4	3.02	0.01	−0.05	0.15	0.05	0.01	0.13	0.04	0.04
5	5.93	0.00	0.01	0.21	0.05	0.00	0.13	0.03	0.00
6	7.63	−0.02	−0.01	0.28	0.07	0.01	0.10	0.02	0.00
7	8.61	0.02	0.01	−0.16	−0.03	0.01	−0.18	−0.03	0.01
8	5.64	−0.02	−0.01	0.34	0.10	0.03	0.24	0.06	0.01
9	24.64	0.00	0.00	−0.02	0.00	0.00	−0.01	0.00	0.01
10	8.92	−0.02	−0.02	0.15	0.02	−0.01	0.09	0.00	−0.01
11	8.35	−0.03	−0.02	0.08	0.01	0.00	0.10	0.02	0.00
12	10.58	0.01	0.00	−0.27	−0.05	0.00	−0.18	−0.03	0.00
13	0.48	−0.04	−0.58	0.61	0.36	0.04	1.00	0.58	0.04
14	2.80	0.03	−0.03	0.33	0.11	0.00	0.24	0.10	0.02
15	4.21	0.06	−0.01	0.28	0.06	0.00	0.30	0.07	−0.01
16	2.24	0.03	−0.05	0.74	0.26	0.03	0.94	0.32	0.02
17	23.95	−0.01	−0.01	−0.02	0.00	0.00	−0.05	−0.01	0.00
18	0.54	0.07	−0.54	0.63	0.34	−0.06	0.67	0.35	−0.06

Tableau 2

Erreur quadratique moyenne de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population Secteur d'activité: COMMERCE DE DÉTAIL. Régions: 18 divisions de recensement en Nouvelle-Écosse.

Région	Estimateur EXT	POS	SYN/C	ERC/C	RÉG/C	SYN/R	ERC/R	RÉG/R
1	3,209	2,206	96	697	1,397	462	769	1,484
2	42,598	24,623	21,782	12,725	17,358	13,110	10,256	14,380
3	10,469	6,853	357	2,592	4,212	146	2,333	3,782
4	5,626	3,657	324	746	1,186	257	1,206	1,853
5	14,554	9,681	2,999	5,090	7,360	1,294	3,993	5,974
6	12,308	5,686	6,713	3,423	4,289	1,255	1,747	2,515
7	34,865	17,988	6,912	9,387	13,451	8,161	12,019	17,239
8	12,066	8,630	5,772	3,694	5,045	2,981	3,528	4,986
9	72,974	40,440	5,776	24,025	29,250	5,068	21,292	25,832
10	22,091	9,433	4,559	5,832	7,927	2,009	5,365	7,272
11	23,519	12,505	1,778	6,738	9,578	2,348	7,890	11,063
12	46,588	21,874	35,310	13,558	17,084	17,454	12,222	16,514
13	635	244	161	95	228	422	287	783
14	3,871	2,849	692	1,254	2,141	378	1,373	2,346
15	8,088	3,511	2,249	1,892	2,806	2,651	1,985	2,937
16	3,245	2,127	3,316	1,563	2,516	5,333	1,741	2,654
17	81,211	47,753	5,503	28,957	35,232	7,681	27,457	33,136
18	1,003	306	169	187	654	186	184	637

Tableau 3

Rendement moyen de l'échantillon et biais relatif de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population Secteur d'activité: HÉBERGEMENT. Régions: 16 divisions de recensement en Nouvelle-Écosse.

Région	Rendement moyen	Estimateur							
		EXT	POS	SYN/C	ERC/C	RÉG/C	SYN/R	ERC/R	RÉG/R
1	0.25	0.01	−0.75	−0.08	−0.06	−0.01	0.36	0.28	0.01
2	1.37	−0.06	−0.21	0.25	0.10	0.02	0.25	0.11	0.02
3	1.02	0.06	−0.26	0.19	0.09	0.04	0.12	0.06	0.03
4	0.23	−0.10	−0.77	−0.33	−0.26	−0.07	−0.15	−0.13	−0.05
5	2.04	0.03	−0.13	0.21	0.08	0.03	0.18	0.06	0.01
6	1.49	0.04	−0.13	0.17	0.10	0.03	0.03	0.02	0.01
7	1.53	0.01	−0.18	−0.29	−0.11	−0.01	−0.30	−0.12	−0.02
8	1.54	0.03	−0.19	−0.42	−0.17	−0.01	−0.26	−0.11	−0.02
9	6.83	0.01	−0.02	0.13	0.02	0.00	0.12	0.02	0.00
10	1.26	−0.01	−0.26	0.40	0.17	0.03	0.30	0.13	0.02
11	3.06	0.04	−0.02	0.51	0.21	0.08	0.40	0.16	0.06
12	1.80	0.02	−0.16	−0.08	−0.05	−0.03	−0.23	−0.10	−0.03
14	1.04	0.02	−0.33	−0.52	−0.23	−0.07	−0.32	−0.15	−0.06
15	1.54	−0.03	−0.23	−0.21	−0.13	−0.08	−0.15	−0.11	−0.08
17	3.08	−0.07	−0.05	−0.03	−0.01	0.00	−0.14	−0.07	−0.03
18	0.52	0.04	−0.54	3.26	3.20	0.60	2.97	2.92	0.50

Tableau 4

Erreur quadratique moyenne de chacun des huit estimateurs pour 500 échantillons aléatoires simples prélevés dans l'ensemble de la population Secteur d'activité: HÉBERGEMENT. Régions: 16 divisions de recensement en Nouvelle-Écosse.

Région	Estimateur							
	EXT	POS	SYN/C	ERC/C	RÉG/C	SYN/R	ERC/R	RÉG/R
1	1,142	283	9	7	25	58	44	164
2	7,467	5,082	877	631	1,077	747	455	726
3	878	442	48	163	242	24	116	163
4	155	43	7	6	17	3	3	6
5	15,200	8,392	2,091	2,270	3,230	1,271	1,208	1,785
6	5,239	3,906	253	1,038	2,193	54	396	792
7	21,197	8,781	3,569	1,831	3,016	3,709	1,812	2,948
8	14,071	6,738	3,608	2,122	4,018	1,492	947	1,766
9	50,606	27,867	9,980	11,413	14,344	6,575	7,779	9,991
10	2,219	993	590	362	665	317	151	280
11	10,535	5,774	6,366	5,126	7,154	3,867	2,752	3,673
12	16,787	10,485	543	1,148	1,944	1,245	1,130	1,836
14	51,471	25,644	9,669	8,221	14,155	3,972	3,189	5,077
15	59,207	41,381	4,861	10,548	18,119	2,759	4,262	6,636
17	29,632	25,211	1,501	3,023	4,754	1,765	2,123	3,214
18	286	99	2,062	2,112	5,623	1,607	1,646	4,561

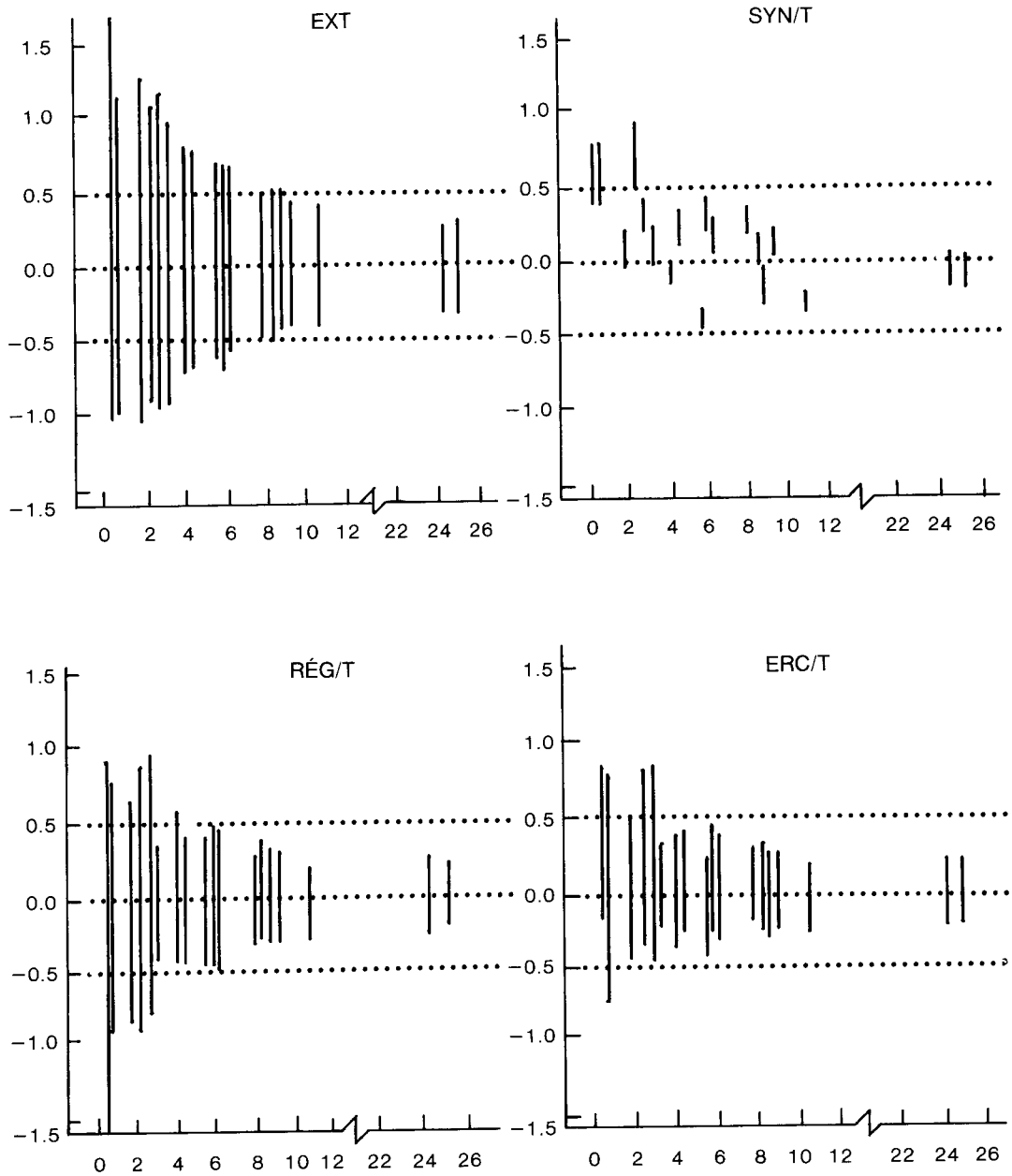
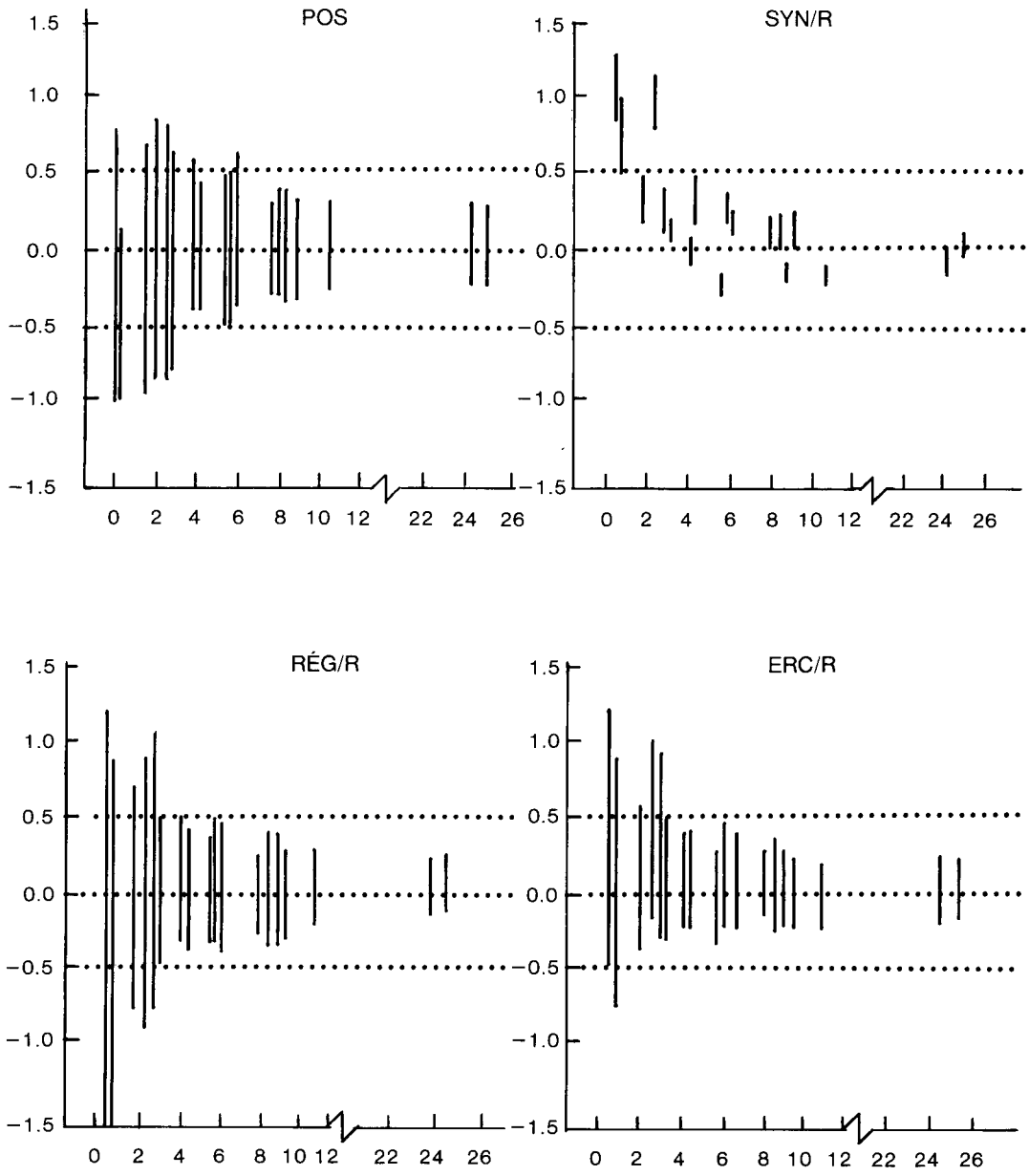


Figure 1: Secteur d'activité: COMMERCE DE DÉTAIL. Régions: 18 divisions de recensement en Nouvelle-Écosse. Bande de distribution de l'erreur relative des estimateurs - l'abscisse représente le rendement moyen de l'échantillon

Figure 1 (continue)



Dans les domaines les plus petits, aucun des estimateurs non biaisés (EXT, POS, RÉG/T, RÉG/R) n'est avantageux du point de vue de la variance, surtout les estimateurs RÉG. Ce problème est résolu par les deux formes corrigées des estimateurs RÉG, les ERC.

C. Les deux formes de l'ERC, l'ERC/T et l'ERC/R, renferment un biais négligeable dans les cas où les estimateurs SYN sont presque sans biais (par exemple COMMERCE DE DÉTAIL, région 17); autrement, les ERC contiennent un certain biais qui, cependant, est généralement moins élevé que celui des estimateurs SYN (par exemple COMMERCE DE DÉTAIL, région 2). La variance et l'erreur quadratique moyenne des ERC sont beaucoup plus faibles, dans tous les domaines, que celles des estimateurs RÉG. Cette constatation est particulièrement fréquente dans les domaines les plus petits. En comparaison des autres estimateurs, on voit que les ERC (comme prévu) ont néanmoins une variance élevée dans pratiquement tous les domaines. Par contre, l'erreur quadratique moyenne des ERC est moins élevée que celle des estimateurs SYN dans les domaines où ces derniers sont extrêmement biaisés. Le tableau 6, par exemple, montre que l'erreur quadratique moyenne de l'ERC/R est plus faible que celle de l'estimateur SYN/R dans 9 petits domaines sur 16. Ce résultat découle évidemment du fait que, dans les domaines où l'estimateur SYN renferme un biais considérable, le carré du biais accroît énormément l'erreur quadratique moyenne de l'estimateur SYN, alors que le carré du biais n'est pas très important pour les ERC. Comme nous ignorons lesquels des domaines créent les biais élevés, il est généralement mieux d'utiliser les ERC pour obtenir des estimations fiables dans tous les domaines.

5. CONCLUSIONS

En somme, nous constatons que les ERC produisent des résultats qui permettent de les considérer comme des outils prometteurs pour les travaux futurs sur les estimations relatives aux petites régions. La méthode recommandée de calcul d'un intervalle de confiance à partir des ERC est décrite à la section 3.

Nous pensons que la méthode de calcul des ERC décrite plus haut représente un moyen simple de ramener les estimations un peu vers la valeur des estimateurs SYN, qui sont stables, quand le rendement de l'échantillon est moins élevé que prévu. Cet objectif (bien qu'atteint par des approches très différentes) est également la base de certains autres travaux récents, comme ceux fondés sur l'estimation empirique bayésienne (Fay et Herriot, 1979) et l'utilisation d'un estimateur dépendant de l'échantillon (Drew, Singh et Choudhry, 1982).

BIBLIOGRAPHIE

DREW, J.D., SINGH, M.P., et CHOUDHRY, G.H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada. *Techniques d'enquête*, 8, 19-54.

FAY, R.E, et HERRIOT, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

HIDIROGLOU, M.A., MORRY, M., DAGUM, E.B., RAO, J.N.K., et SÄRNDAL, C.E. (1984). Evaluation of alternative small area estimators using administrative data. Communication présentée aux réunions de l'ASA, Philadelphie, août 1984.

SÄRNDAL, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustments for nonresponse. *Bulletin of the International Statistical Institute*, 49:1, 494-513. (Actes de la 43ieme séance, Buenos Aires).

SÄRNDAL, C.E., et RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review*, 1983:5 (Essais en hommage à T.E. Dalenius), 33-40.

SÄRNDAL, C.E. (1984). Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.