Estimation pour les petits domaines : théorie et pratique à Statistique Canada

Michel Arsène HIDIROGLOU
Statistique Canada

1. INTRODUCTION

Au cours des dernières années, la demande de données sur les petits domaines s'est fortement accrue. Cette augmentation s'explique par l'utilité de ces données dans l'élaboration de politiques et de programmes gouvernementaux, dans l'allocation de divers fonds, et dans la planification régionale. Afin de satisfaire à ces nouvelles demandes, plusieurs agences de statistique nationales et régionales, dont Statistique Canada, ont introduit des programmes visant à produire des estimations pour les petits domaines. Lorsque ces données sont produites à partir d'un recensement complet de la population, il n'y a évidemment aucun problème d'estimation. Très souvent cependant, les données disponibles ont été recueillies à partir d'un sondage inapte à produire ce genre d'estimations. Néanmoins, s'il existe des données administratives incorporant des informations sur les petits domaines et qui sont fortement corrélées avec les caractéristiques faisant l'objet de l'estimation, alors il existe plusieurs méthodes pour produire des estimations convenables pour ces domaines. L'exposé qui suit décrira quelques unes de ces méthodes.

Il est bien sûr possible de produire, des données pour les petits domaines en se servant de l'estimateur direct, c'est-à-dire en utilisant les poids de l'échantillon. Cependant, cet estimateur n'est pas sans inconvénients. Premièrement, s'il s'agit d'estimer des totaux, il est conditionnellement biaisé, c'est-à-dire, son espérance pour les échantillons de même taille que celle réalisée ne donne pas la valeur attendue. Deuxièmement, lorsque le petit domaine ne renferme que peu d'unités, la variance sera très élevée. Afin de remédier à ces problèmes, il est nécessaire d'élaborer des estimateurs qui tiennent compte des autres petits domaines dans la population. Ces estimateurs sont construits en se servant de modèles qui lient les données d'enquête aux données auxiliaires provenant de banques de données administratives ou de recensements. La gamme de méthodes étudiées à Statistique Canada inclut les suivantes: l'estimation synthétique, l'estimation par le quotient (effectué par post-strates), les méthodes d'estimation qui dépendent de la taille réalisée de l'échantillon dans le petit domaine, et les méthodes de régression. Une évaluation de ces méthodes en fonction de leurs biais, leurs

variances et leurs erreurs quadratiques moyennes a été effectuée à partir d'une simulation utilisant des données tirées d'une enquête auprès des entreprises.

Cette simulation a été effectuée avec des données provenant d'une enquête ponctuelle. Cependant, des organisations telles que Statistique Canada mènent des enquêtes annuelles, voire mensuelles, dont l'Enquête sur la population active. Il est alors possible d'élaborer des estimateurs qui utilisent l'information supplémentaire que l'on peut retrouver dans une série chronologique, et ainsi améliorer davantage les estimations pour les petits domaines.

2. ESTIMATEURS SIMPLES

Soit une population $U = \{1,..., k,...,N\}$, divisée en D petits domaines $U_1, U_2,...,U_D$. Ces domaines sont exhaustifs et disjoints. C'est-à-dire, $U = \bigcup_{d=1}^{D} U_d \quad et \quad U_{d_1} \cap U_{d_2} = \emptyset \quad si \quad d_1 \neq d_2$. Nous (1)

désirons obtenir une estimation pour le total (ou la moyenne) d'une variable cible dans chacun de ces domaines. Nous supposons que la taille, (N_d), de chacun des domaines U_d est connue. Nous supposons aussi qu'on peut subdiviser cette population selon un autre critère de classification, en G groupes $U_{.1}, U_{.2},...,U_{.G}$, exhaustifs et disjoints. Normalement, ces groupes sont des post-strates. Ces deux divisions répartissent la population en DG cellules U_{dg} $(d = 1,..., D: g = 1,...,G)$, dont nous supposons la taille (N_{dg}), connue.

La taille de la population (N) peut alors s'exprimer par

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

Soit maintenant l'échantillon "s", de taille n, tiré de la population U avec probabilité p(s), où $P(k \in s) = \pi_k > 0$ et $P(k \in s \text{ et } l \in s) = \pi_{kl} > 0$ pour tous les $k \neq l$. L'échantillon s se subdivise en $s_d = s \cap U_d \text{ et } s_{dg} = s \cap U_{dg}$. Les tailles respectives de l'échantillon dans s_d et s_{dg} sont n_d et n_{dg}. Nous voulons estimer le total (ou la moyenne)

$$Y_d = \sum_{k \in U_d} y_k \qquad (\bar{Y}_d = Y_d / N_d) \tag{2.2}$$

Par la suite, nous indiquerons les sommes du genre $\sum_{k \in U_d} y_k$ par $\sum_{U_d} y_k$.

2.1 Le cas où des données auxiliaires n'existent pas

L'estimateur le plus simple de Y_d est l'estimateur direct,

$$\hat{Y}_{dDIR} = \Sigma_{s_d} y_k/\pi_k \quad . \tag{2.3}$$

Cet estimateur est inconditionnellement sans biais pour n'importe quel plan d'échantillonage. Dans le cas d'un échantillon aléatoire simple, on peut démontrer que le biais et la variance conditionnels sont

$$biais_c(\hat{Y}_{dDIR}) = [(N/n)n_d - N_d]\bar{y}_{U_d}$$

et

$$V_c(\hat{Y}_{dDIR}) = (Nn_d/n)^2(1/n_d - 1/N_d)S^2_{U_d} \tag{2.4}$$

où n_d est la taille de s_d et $\bar{y}_{U_d}$ et $S^2_{U_d}$ sont, respectivement, la moyenne et la variance de y pour le domaine U_d. Nous notons que le biais conditionnel est près de 0 si n_d est près de son espérance nN_d/N. Cependant, on peut difficilement justifier un intervalle de confiance conditionnel pour un tel estimateur.

L'estimateur synthétique simple de Y_d est

$$\hat{Y}_{dSYN/C}(1) = N_d \bar{y}_s \tag{2.5}$$

où $\bar{y}_s = (\Sigma_s y_k/\pi_k)/(\Sigma_s 1/\pi_k)$. "C" indique que l'on se sert des tailles connues N_d de la population. Cet estimateur est biaisé sauf si:

Hypothèse 1. $\bar{y}_{U_d} = \bar{y}_U$, $d=1,\ldots,D$

où $\bar{y}_U$ est la moyenne pour la population U. Cette hypothèse est décrite par le modèle suivant

$$y_k = \mu + e_k \ , \ k=1,\ldots,N \tag{2.6}$$

où $E(e_k) = 0$, $E(e_k^2) = \sigma^2$ et $E(e_k e_j) = 0$, $k \neq j$. L'estimateur des moindres carrés de μ pour ce modèle est $\hat{\mu} = \bar{y}_s$. Etant donné que

$$Y_d = \sum_{s_d} y_k + \sum_{U_d - s_d} y_k ,$$

et que le meilleur prédicteur sans biais (MPSB) de y_k est $\hat{\mu} = \bar{y}_s$ (Holt et coll. 1979), nous avons:

$$\begin{aligned} \hat{Y}_{dSYN/C}(2) &= \sum_{s_d} y_k + \sum_{U_d - s_d} \hat{y}_k \\ &= n_d \bar{y}_{s_d} + (N_d - n_d) \bar{y}_s \\ &= N_d \bar{y}_s + n_d (\bar{y}_{s_d} - \bar{y}_s) . \end{aligned} \qquad (2.7)$$

Le biais conditionnel de $\hat{Y}_{dSYN/C}(2)$ est:

$$biais(\hat{Y}_{dSYN/C}(2)) = (N_d - n_d)(\bar{Y}_U - \bar{Y}_{U_d}) .$$

Il est à noter que $\hat{Y}_{dSYN/C}(2)$ est préférable à $\hat{Y}_{dSYN/C}(1)$: $\hat{Y}_{dSYN/C}(2)$ est égal à Y_d lorsque $n_d = N_d$, alors que $\hat{Y}_{dSYN/C}(1)$ ne l'est pas.

Dans le cas où l'estimation synthétique s'effectue indépendamment pour chacun des groupes g, l'estimation est sans biais si:

Hypothèse 2. $\bar{y}_{U_{dg}} = \bar{y}_{U_{.g}}$, $g = 1, \ldots, G$ (2.8)

où $\bar{y}_{U_{dg}}$ est la moyenne des N_{dg} unités de la population qui appartiennent au petit domaine d et au groupe g, $\bar{y}_{U_{.g}} = \sum_{U_{.g}} y_k / N_{.g}$; $N_{.g} = \sum_{d=1}^{D} N_{dg}$. Un estimateur synthétique est obtenu en

remplacant la moyenne $\bar{y}_{s_{dg}} = (\Sigma_{s_{dg}} y_k/\pi_k) / (\Sigma_{s_{dg}} 1/\pi_k)$ qui figure dans l'estimateur

$$\hat{Y}_{dPOSG/C} = \sum_{g=1}^{G} N_{dg} \bar{y}_{s_{dg}} \quad (n_{dg} > 0 \text{ pour tous les g}) \tag{2.9}$$

par la moyenne $\bar{y}_{s_{.g}} = (\Sigma_{s_{.g}} y_k/\pi_k) / (\Sigma_{s_{.g}} 1/\pi_k)$:

$$\hat{Y}_{dSYNG/C}(1) = \sum_{g=1}^{G} N_{dg} \bar{y}_{s_{.g}} \ . \tag{2.10}$$

Le modèle sous-jacent est

$$y_k = \mu_g + e_k \quad si \quad k \in U_{dg} \tag{2.11}$$

où $E(e_k) = 0$, $E(e_k^2) = \sigma_g^2$ et $E(e_k e_j) = 0$, $k \neq j$. Comme auparavant, nous avons que

$$Y_d = \Sigma_g \{\Sigma_{s_{dg}} y_k + \Sigma_{U_{dg}-s_{dg}} \hat{y}_k\}$$

et que le MPSB de y_k est $\hat{\mu}_g = \bar{y}_{s_{.g}}$, ce qui mène à l'estimateur

$$\begin{aligned} \hat{Y}_{dSYNG/C}(2) &= \sum_{g=1}^{G} \{\Sigma_{s_{dg}} y_k + \Sigma_{U_{dg}-s_{dg}} \hat{y}_k\} \\ &= \sum_{g=1}^{G} \{N_{dg} \bar{y}_{s_{.g}} + n_{dg} (\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\}. \end{aligned} \tag{2.12}$$

2.2 Le cas où il existe une variable auxiliaire

S'il existe une variable auxiliaire x pour chacun des éléments de la population U et qu'elle ait une corrélation non triviale avec la variable y, nous pouvons en profiter pour améliorer nos

estimations. Nous supposons alors que le rapport $R_{U_d} = \bar{y}_{U_d}/\bar{x}_{U_d}$ ne dépend pas du domaine, c'est-à-dire,

Hypothèse 3. $R_{U_d} \doteq R = \bar{y}_U/\bar{x}_U$, d = 1,...,D (2.13)

Un estimateur synthétique pour Y_d est

$$\hat{Y}_{dSYN/R}(1) = X_d(\bar{y}_s/\bar{x}_s) \tag{2.14}$$

où "R" est utilisé pour montrer que l'on sert de X_d, le total (connu) de la variable x pour le petit domaine d. Le modèle sous-jacent est

$$y_k = \beta x_k + e_k \tag{2.15}$$

où $E(e_k) = 0$. $E(e_k^2) = \sigma^2 x_k$ et $E(e_k e_j) = 0$ pour $k \neq j$, et le meilleur "prédicteur" sans biais (MPSB)

$$\begin{aligned} {}_0\hat{Y}_{dSYN/R}(2) &= n_d \bar{y}_{s_d} + (N_d \bar{x}_{U_d} - n_d \bar{x}_{s_d})\bar{y}_s/\bar{x}_s \\ &= X_d (\bar{y}_s/\bar{x}_s) + n_d(\bar{y}_{s_d} - \bar{x}_{s_d}\bar{y}_s/\bar{x}_s) \end{aligned} \tag{2.16}$$

L'estimateur $\hat{Y}_{dSYN/R}(2)$ est égal à Y_d lorsque $n_d = N_d$.

On peut aisément généraliser cet estimateur pour le cas où l'hypothèse 3 s'applique aux groupes g, c'est-à-dire, $R_{U_{dg}} \doteq R_g = \bar{y}_{U_g}/\bar{x}_{U_g}$, d = 1,...,D, où $R_{U_{dg}} = \bar{y}_{U_{dg}}/\bar{x}_{U_{dg}}$.

3. ESTIMATEURS COMPLEXES

3.1 Un estimateur de régression qui corrige le biais

Plusieurs méthodes d'estimation pour les petits domaines reposent sur des méthodes de régression. Pour ces cas, nous avons un vecteur $\mathbf{x}$ de variables auxiliaires observées pour toutes les unités de la population U. Il s'agit donc d'élaborer des modèles linéaires reliant la variable cible y, observée pour l'échantillon s, à $\mathbf{x}$, un vecteur de dimension p. Un modèle de ce genre, qu'on représente ici par ξ, suppose que $y_1, y_2, \ldots, y_N$ sont indépendants et que

$$E_\xi(y_k) = \boldsymbol{x}'_k\boldsymbol{\beta} \; ; \; V_\xi(y_k) = \sigma_k^2 \quad . \tag{3.1}$$

L'estimateur pondéré de $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ est

$$\boldsymbol{B}_U = [\Sigma_U(\boldsymbol{x}_k\boldsymbol{x}'_k)/\sigma_k^2]^{-1} \; \Sigma_U(\boldsymbol{x}_k y_k/\sigma_k^2) \quad . \tag{3.2}$$

Pour le plan de sondage décrit au début de la section 2, un estimateur de $\boldsymbol{B}_U$ est

$$\boldsymbol{B}_s = [\Sigma_s(\boldsymbol{x}_k\boldsymbol{x}'_k)/\sigma_k^2\pi_k]^{-1} \; \Sigma_s(\boldsymbol{x}_k y_k/\sigma_k^2\pi_k) \quad . \tag{3.3}$$

Hypothèse 4. $\quad \boldsymbol{B}_{U_d} \doteq \boldsymbol{B}_U \;$, d = 1,...,D $\qquad$ (3.4)

Si l'on admet le même modèle (3.1) pour les petits domaines s_d, un estimateur synthétique de U_d est

$$\hat{Y}_{dSYN} = \Sigma_{U_d}\hat{y}_k = \Sigma_{U_d}(\boldsymbol{x}'_k\,\boldsymbol{B}_s) \tag{3.5}$$

Le biais inconditionnel de cet estimateur est

$$biais_I(\hat{Y}_{dSYN}) \doteq \Sigma_{U_d}(\mathbf{x}'_k \mathbf{B}_U - y_k) \quad .$$

et son biais conditionnel (c'est-à-dire, pour les échantillons de taille n_d) est

$$biais_c(\hat{Y}_{dSYN}) \doteq \Sigma_{U_d}(\mathbf{x}'_k \boldsymbol{\beta}_c - y_k)$$

où

$$\boldsymbol{\beta}_c = \left\{E_c\left(\Sigma_s \frac{\mathbf{x}_k \mathbf{x}'_k}{v_k}\right)\right\}^{-1} \left\{E_c\left(\Sigma_s \frac{\mathbf{x}_k y_k}{v_k}\right)\right\} \quad . \tag{3.6}$$

Cette expression du biais conditionnel est donnée dans Särndal et Hidiroglou (1989). L'opération espérance conditionnelle $E_c(.)$ s'effectue sur le sous-ensemble de tous les échantillons s pour lesquels la taille n_d de s_d est la même que la taille de l'échantillon réalisée. Ce principe ne s'applique que dans le cas d'un échantillon aléatoire simple.

Afin de corriger le biais inconditionnel, Särndal (1984) a proposé l'estimateur suivant

$$\hat{Y}_{dRE} = \Sigma_{U_d} \hat{y}_k + \Sigma_{s_d} e_k/\pi_k \tag{3.7}$$

où $e_k = y_k - \hat{y}_k$ est le résidu. Quoique cet estimateur corrige le biais de l'estimateur synthétique $\hat{Y}_{dSYN}$, il est conditionnellement biaisé. Hidiroglou et Särndal (1985) ont proposé une version conditionnellement sans biais de cet estimateur, donnée par

$$\hat{Y}_{dMRE} = \Sigma_{U_d} \hat{y}_k + \frac{N_d}{\hat{N}_d} \Sigma_{s_d} e_k/\pi_k \tag{3.8}$$

où $\hat{N}_d = \Sigma_{s_d} 1/\pi_k$. Cet estimateur possède plusieurs avantages par rapport au précédent. Premièrement, $\hat{Y}_{dMRE}$ a une variance plus petite que $\hat{Y}_{dRE}$, car il tient compte du rapport entre la taille de la population N_d et de son estimateur $\hat{N}_d$. Deuxièmement, $\hat{Y}_{dMRE}$ est

conditionnellement sans biais, alors que $\hat{Y}_{dRE}$ ne l'est pas. Finalement, lorsque $n_d = N_d$ et que l'échantillon est un échantillon aléatoire simple, $\hat{Y}_{dMRE}$ est égal à Y_d. Dans le cas où n_d est très petit (5 unités ou moins), la variance du terme de correction pour $(N_d/\hat{N}_d) \sum_{s_d} e_k/\pi_k$, pourrait être élevée. Ceci peut nous mener à des situations inacceptables, à savoir, des estimations négatives pour un petit domaine alors que celles-ci devraient être positives. Afin d'éviter une telle éventualité, Särndal et Hidiroglou (1989) ont proposé d'atténuer le terme de correction lorsque $\hat{N}_d < N_d$. Ceci nous donne l'estimateur atténué (ARE):

$$\hat{Y}_{dARE} = \sum_{U_d} \hat{y}_k + \left(\frac{\hat{N}_d}{N_d}\right)^{\lambda-1} \sum_{s_d} e_k/\pi_k \quad , \tag{3.9}$$

avec

$$\begin{aligned} \lambda &= 0 \quad si \quad \hat{N}_d \geq N_d \\ &= h \quad si \quad \hat{N}_d < N_d \end{aligned} ,$$

où h est une constante positive. Une valeur de h=2 semblerait être adéquate. Cet estimateur est conditionnellement sans biais si $\hat{N}_d \geq N_d$, sinon il est conditionnellement biaisé. Son biais est alors

$$biais_c(\hat{Y}_{dARE}) \doteq [1-(\hat{N}_d/N_d)^{\lambda}] \sum_{U_d} (y_k - \mathbf{x}'_k B_U) \tag{3.10}$$

$\hat{Y}_{dARE}$ ressemble beaucoup à $\hat{Y}_{dMRE}$, ce qui nous mène à nous servir de la variance estimée de ce dernier pour construire des intervalles de confiance pour $\hat{Y}_{dARE}$. Cette variance conditionnelle estimée est donnée par:

$$v(\hat{Y}_{dMRE}) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{k \neq l} \sum_{s_d} \Delta_{kl} \frac{(e_k - \bar{e}_{s_d})(e_l - \bar{e}_{s_d})}{\pi_k \pi_l}$$

$$où \qquad \bar{e}_{s_d} = (\Sigma_{s_d} e_k/\pi_k) \; / \; (\Sigma_{s_d} 1/\pi_k) \tag{3.11}$$

$$et \qquad \Delta_{kl} = \begin{cases} 1-\pi_k \; si \; l=k \\ 1-\pi_k\pi_l/\pi_{kl} \; si \; l \neq k. \end{cases}$$

3.2 Un estimateur de régression avec erreur emboîtée

Nous avons supposé dans le modèle (3.1) que les erreurs $e_k = y_k - \boldsymbol{x}'_k\beta$ étaient indépendantes. Cependant, il est souvent raisonnable de supposer que les erreurs sont indépendantes si elles proviennent de domaines différents, mais dépendantes à l'intérieur de chaque domaine. Ce modèle a été proposé par Battese, Harter et Fuller (1988) pour estimer le rendement moyen de certains produits agricoles dans des comtés (petits domaines) de l'Iowa, en reliant des données du satellite Landsat à des données de sondage. Leur modèle, représenté ici par η est

$$y_{dk} = \boldsymbol{x}'_{dk}\beta + u_{dk} \tag{3.12}$$

où $\boldsymbol{x}'_{dk} = (x_{dk1}, \ldots, x_{dkp})$, $u_{dk} = v_d + e_{dk}$ $pour$ $k \varepsilon s$. Les erreurs $v_d (d=1,2,\ldots,D)$ sont indépendantes et à distribution normale avec moyenne 0 et variance σ_v^2 . En plus, elles sont indépendantes des erreurs e_{dk} (k = 1, 2, ..., n_d), qui sont aussi indépendantes et à distribution normale, avec moyenne 0 et variance σ_e^2 . Ceci entraîne la structure de covariance suivante pour les erreurs u_{dk}

$$\begin{aligned} E_\eta(u_{dk}u_{rq}) &= \sigma_v^2+\sigma_e^2 \quad d=r \ , \ k=q \\ &= \sigma_v^2 \quad\quad d=r \ , \ k\neq q \\ &= 0 \quad\quad d\neq r \end{aligned} \tag{3.13}$$

Si nous posons $x_{dk1} = 1$ et $\beta = \alpha$, ce modèle donnera des ordonnées à l'origine aléatoires, à savoir, les variables $\alpha_d = \alpha + v_d$. La moyenne estimée pour le petit domaine d pour le modèle (3.12) est

$$\begin{aligned} \bar{y}_{U_d} &= \bar{\boldsymbol{x}}'_{U_d}\beta + v_d + \bar{e}_{U_d} \\ &\doteq \bar{\boldsymbol{x}}'_{U_d}\beta + v_d \end{aligned} \tag{3.14}$$

où $\bar{\mathbf{x}}_{U_d}$ est le vecteur des moyennes (connues) de $\mathbf{x}_{dk}$ et où on suppose N_d assez grand pour que $\bar{e}_{U_d} = N_d^{-1} \sum_{U_d} e_{dk} \doteq 0$. Nous constatons que le premier terme de (3.14) est l'estimateur synthétique $\bar{Y}_{dSYN} = \hat{Y}_{dSYN} / N_d$, et que v_d apporte une correction du biais à celui-ci.

Si les variances σ_v^2 *et* σ_e^2 sont connues,

$$\hat{\boldsymbol{\beta}}_N = (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{y}) \qquad (3.15)$$

où $\boldsymbol{X} = col_{1 \le d \le D}\, col_{1 \le k \le n_d} (\boldsymbol{x}'_{dk})$; $\boldsymbol{y} = col_{1 \le d \le D}\, col_{1 \le k \le n_d} (y_{dk})$;

$\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_D)$ avec $\boldsymbol{V}_d = \sigma_e^2 \, \boldsymbol{I}_{n_d} + \sigma_v^2 \, \mathbf{1}_{n_d} \mathbf{1}'_{n_d}$ est l'estimateur généralisé des moindres carrés de $\boldsymbol{\beta}$. $\boldsymbol{I}_{n_d}$ est une matrice identité d'ordre n_d et $\mathbf{1}_{n_d}$ est un vecteur de longueur n_d dont les éléments sont des "1".

Si toutes les erreurs u_{dk} ($k = 1, 2, \ldots, n_d$) étaient connues, alors le meilleur "prédicteur" de v_d serait l'espérance conditionnelle de v_d, étant donné la moyenne $\bar{u}_{s_d}$ de l'échantillon, où $\bar{u}_{s_d} = n_d^{-1} \sum_{s_d} u_{dk}$ *et* $u_{dk} = y_{dk} - \boldsymbol{x}'_{dk} \boldsymbol{\beta}$. Puisque v_d et $\bar{u}_{s_d}$ suivent une distribution bivariée normale avec moyenne (0,0) et matrice de covariance

$$\begin{pmatrix} \sigma_v^2 & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + n_d^{-1} \sigma_e^2 \end{pmatrix} ,$$

l'espérance de v_d étant donné $\bar{u}_{s_d}$, est

$$E(v_d|\bar{u}_{s_d}) = \bar{u}_{s_d}\,\gamma_d \tag{3.16}$$

où $\gamma_d = \sigma_v^2(\sigma_v^2+n_d^{-1}\sigma_e^2)^{-1}$. La variance du meilleur "prédicteur" est $\sigma_v^2(1-\gamma_d)$ (Battese, Harter et Fuller, 1988). Donc un "prédicteur" possible pour v_d est

$$\tilde{v}_d = \tilde{u}_{s_d}\,\gamma_d \tag{3.17}$$

où $\tilde{u}_{s_d} = n_d^{-1}\sum_{s_d}(y_{dk}-\mathbf{x}'_{dk}\tilde{\boldsymbol{\beta}}_N)$. L'estimateur de la moyenne $\bar{y}_{u_d}$ est

$$\begin{aligned}\bar{y}_{dN} &= \bar{\mathbf{x}}'_{U_d}\tilde{\boldsymbol{\beta}}_N+\tilde{v}_d \\ &= \bar{\mathbf{x}}'_{U_d}\tilde{\boldsymbol{\beta}}_N+\gamma_d(\bar{y}_{s_d}-\bar{\mathbf{x}}'_{s_d}\tilde{\boldsymbol{\beta}}_N)\,.\end{aligned} \tag{3.18}$$

Ceci implique que l'estimateur du total Y_{U_d} est

$$\begin{aligned}\hat{Y}_{dN} &= \textstyle\sum_{U_d}\tilde{y}_{dk}+\gamma_d N_d\,(\bar{y}_{s_d}-\bar{\mathbf{x}}'_{s_d}\tilde{\boldsymbol{\beta}}_N) \\ &= \textstyle\sum_{U_d}\tilde{y}_{dk}+\gamma_d N_d \sum_{s_d}(y_{dk}-\mathbf{x}'_{s_d}\tilde{\boldsymbol{\beta}}_N)/n_d \\ &= \textstyle\sum_{U_d}\tilde{y}_{dk}+\gamma_d \dfrac{N_d}{\hat{N}_d}\sum_{s_d}(y_{dk}-\mathbf{x}'_{s_d}\tilde{\boldsymbol{\beta}}_N)/(n/N)\end{aligned} \tag{3.19}$$

où $\tilde{y}_{dk} = \mathbf{x}'_{dk}\tilde{\boldsymbol{\beta}}_N$, $\bar{\mathbf{x}}_{s_d}$ $(\mathbf{x}_{s_d})$ est le vecteur de moyennes (totaux) de $\mathbf{x}_{dk}$ basées sur l'échantillon s_d et $\hat{N}_d = \frac{N}{n}n_d$. On constate que cet estimateur est exactement l'estimateur $\hat{Y}_{dMRE}$ (3.8) si $\gamma_d=1$ et si l'échantillon est un échantillon aléatoire simple. Nous notons que $\gamma_d=1$ implique $\sigma_v^2\to\infty$, et donc la variabilité des ordonnées à l'origine aléatoires α_d est grande. Par

contre, $\gamma_d=0$ implique $\sigma_v^2=0$, et on obtient l'estimateur synthétique $\hat{Y}_{dSYN}$: l'hypothèse 4, $\boldsymbol{\beta}_{U_d} \doteq \boldsymbol{\beta}_U$ est donc satisfaite dans ce cas. Remarquons que $\hat{Y}_{dN}$ est le meilleur "prédicteur" sans biais (MPSB) pour Y_{U_d} et qu'il est équivalent à $\gamma_d \hat{Y}_{dPOST/C} + (1-\gamma_d)\hat{Y}_{dSYN}$ où $0 \leq \gamma_d \leq 1$. Quoique $\hat{Y}_{dN}$ soit conditionnellement biaisé, son erreur quadratique moyenne est inférieure à celle de $\hat{Y}_{dARE}$.

Puisqu'on ne connaît pas les vraies valeurs de σ_v^2 et σ_e^2 , on doit les estimer. Des estimateurs sans biais de σ_v^2 et σ_e^2 sont

$$\tilde{\sigma}_e^2 = (n-D-p)^{-1} \sum_d \sum_k \hat{e}_{dk}^2 \qquad (3.20)$$

et

$$\tilde{\sigma}_v^2 = n_*^{-1} [\sum_d \sum_k \hat{u}_{dk}^2 - (n-p)\,\tilde{\sigma}_e^2] \qquad (3.21)$$

où $n_* = n - tr[(\boldsymbol{X}'\boldsymbol{X})^{-1} \sum_{d=1}^{D} n_d^2 \bar{\boldsymbol{x}}_{s_d} \bar{\boldsymbol{x}}'_{s_d}]$. Les $\{\hat{e}_{dk}\}$ sont les résidus de la régression ordinaire de y_{dk} sur $\mathbf{x}_{dk}$. Ces valeurs estimées sont substituées dans l'expression (3.19) et l'estimateur qui en résulte est dénoté par $\tilde{Y}_{dN}$.

3.3 Un estimateur de régression empirique de Bayes

Nous pouvons étendre l'hypothèse 1 au modèle de régression en supposant que

Hypothèse 5. $\bar{y}_{U_d} \doteq \bar{\boldsymbol{x}}'_{U_d}\boldsymbol{\beta}$, $d=1,\ldots,D$ (3.22)

$\bar{y}_{s_d}$ est modelé en fonction de $\bar{\boldsymbol{x}}'_{U_d}$ en supposant le modèle simple $\bar{y}_{s_d} = \bar{\boldsymbol{x}}'_{U_d}\boldsymbol{\beta} + e_d$ où

$E(e_d) = 0$ et $E(e_d^2) = \sigma^2$. L'estimateur synthétique de régression est

$$\hat{Y}_{dREG} = N_d \bar{\boldsymbol{x}}'_{U_d} \boldsymbol{\beta}$$
$$= \Sigma_{U_d}(\boldsymbol{x}'_k \boldsymbol{\beta})$$

où

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'y$$

avec $\boldsymbol{X} = col_{1 \le d \le D}(\bar{\boldsymbol{x}}'_{U_d})$ et $\boldsymbol{y} = col_{1 \le d \le D}(\bar{y}_{s_d})$.

Fay et Herriot (1979) introduisent un terme aléatoire dans le modèle (3.22) en supposant que $\bar{y}_{U_d} = \bar{\boldsymbol{x}}'_{U_d}\boldsymbol{\beta} + v_d$ et que les v_d sont indépendants avec moyenne zéro et variance "A". Ce modèle ressemble beaucoup à celui de Battese, Harter et Fuller (1988). Nous n'avons qu'à revoir l'équation (3.14) pour noter que ces deux modèles, l'un au niveau des unités et l'autre au niveau de moyennes connues, considèrent les moyennes $\bar{y}_{U_d}$ des petits domaines comme des valeurs aléatoires. Nous combinons ce modèle avec celui où on tient compte de l'erreur de l'échantillon, c'est-à-dire, $\bar{y}_{s_d} = \bar{y}_{U_d} + \bar{e}_{s_d}$. La moyenne de $\bar{e}_{s_d}$ est zéro et sa variance est $F_d = S_{U_d}^2 / n_d$, où $S_{U_d}^2$ est la variance des y au niveau de la population U_d. Si on ne connaît pas cette variance, on l'estime à partir de l'échantillon. Pour résumer, nous avons la situation suivante

$$\bar{y}_{s_d} = \bar{y}_{U_d} + \bar{e}_{s_d} \quad et \quad \bar{y}_{U_d} = \bar{\boldsymbol{x}}'_{U_d}\boldsymbol{\beta} + v_d \tag{3.23}$$

où $\bar{\boldsymbol{e}} = (\bar{e}_{s_1}, \ldots, \bar{e}_{s_D})'$ et $\boldsymbol{v} = (v_1, \ldots, v_{s_D})'$ sont indépendants et avec distributions N(0 , F) et

N(0 , AI), respectivement, où F = diag(F_1,...,F_D). Nous devons donc estimer $\boldsymbol{\beta}$ et "A" à partir

des données. Avec les hypothèses ci-dessus, l'estimateur général de moindres carrés pour $\bar{y}_{U_d}$ est:

$$y^{*}_{U_d} = \bar{\boldsymbol{x}}'_{U_d}\,\boldsymbol{\beta}_{EB}$$

où $\boldsymbol{\beta}_{EB} = (\boldsymbol{x}'\boldsymbol{V}^{-1}\boldsymbol{x})^{-1}\,\boldsymbol{x}'\boldsymbol{V}^{-1}\boldsymbol{y}$ avec V = diag ($A+F_1,\ldots,A+F_D$). Pour "A" connu, nous avons

$$E\left(\sum_d \frac{(\bar{y}_{U_d}-y^{*}_{U_d})^2}{A+F_d}\right) = D-p \quad . \tag{3.24}$$

Fay et Herriot (1979) résolvent (3.24) en faisant disparaître l'espérance. C'est-à-dire

$$\sum_d \frac{(\bar{y}_{U_d}-y^{*}_{U_d})^2}{A^{*}+F_d} = D-p \tag{3.25}$$

est résolu de façon itérative pour A^{*}. L'estimateur empirique de Bayes pour $\bar{y}_{U_d}$ est

$$\begin{aligned} \bar{y}_{dEB} &= \frac{A^{*}}{A^{*}+F_d}\bar{y}_{s_d} + \frac{F_d}{A^{*}+F_d}y^{*}_{U_d} \\ &= \bar{\boldsymbol{x}}'_{U_d}\,\hat{\boldsymbol{\beta}}_{EB} + \frac{A^{*}}{A^{*}+F_d}\,(\bar{y}_{s_d}-\bar{\boldsymbol{x}}'_{U_d}\,\hat{\boldsymbol{\beta}}_{EB}) \end{aligned} \tag{3.26}$$

Notons que cet estimateur ressemble beaucoup à celui de Battese, Harter et Fuller (1988). Si $A^{*}\to 0$ on se rapproche de l'estimateur synthétique $\bar{\boldsymbol{x}}'_{U_d}\,\boldsymbol{\beta}_{EB}$, tandis que si $F_d\to 0$ on se rapproche de l'estimateur direct $\bar{y}_{s_d}$, où $\bar{y}_{s_d}=\hat{Y}_{dDIR}/N_d$

3.4 Résultats d'une étude empirique

Afin d'étudier les propriétés des estimateurs décrits dans les sections précédentes, nous avons entrepris une simulation. La province de la Nouvelle-Ecosse a été choisie comme univers. La population comprenait N = 1,678 unités (déclarations d'impôt sur le revenu pour des entreprises non constituées en société) . La variable analysée étaient le montant des salaires et traitements (y) des salariés. Nous avons utilisé une seule variable auxiliaire, le revenu brut de l'entreprise (x). Les valeurs de $x_1,...,x_N$ étaient connues.

La population a été répartie en domaines selon 4 secteurs d'activité économique et 18 régions. Les secteurs d'activité économiques étaient: le commerce de détail (515 unités), le bâtiment et les travaux publics (496 unités), l'hébergement (114 unités), les autres activités économiques constituant le quatrième secteur (553 unités). Les valeurs des coefficients de corrélation entre les salaires et les traitements d'une part, et le revenu brut d'entreprise, d'autre part, étaient 0,42 pour le commerce de détail, 0,64 pour le bâtiment, 0,78 pour l'hébergement et 0,61 pour les autres activités économiques. On a obtenu 70 domaines sur 72 (2 ne comprenaient aucune unité). Pour chacun des 70 domaines, on doit estimer un total Y_d pour chaque échantillon prélevé. Pour la simulation de Monte Carlo, 500 échantillons aléatoires simples de n = 419 unités chacun ont été tirés de la population de N = 1,678 unités. Ceci correspond à un taux d'échantillonnage de 25%. Les unités sélectionnées ont été classées selon le secteur d'activité économique et la division de recensement (région). La population aurait pu être subdivisée selon une deuxième variable, par exemple la tranche de revenu. Pour cette étude, cependant on a supposé que toutes les entreprises étaient comprises dans une seule tranche de revenu (G=1). Nous avons étudié deux versions principales des estimateurs. La première dépend du nombre d'unités n_{dg} de l'échantillon s qui se situent dans le domaine d et le groupe g. Cette version (ESTG/C) se traduit par le modèle suivant: $E_\xi(y_k) = \beta_g$ et $V_\xi(y_k) = \sigma_g^2$ pour g= 1,2,...,G. La deuxième dépend de l'estimation $\sum_{s_d} x_k$, où la somme porte sur les unités de l'échantillon qui sont dans le domaine d et le groupe g. Cette version (ESTG/R) suit le modèle suivant: $E_\xi(y_k) = \beta_g x_k$ et $V_\xi(y_k) = \sigma_g^2 x_k$ pour g= 1,2,...,G. Le nombre d'entreprises N_{dg} et les totaux des données auxiliaires sont connus pour chaque domaine U_{dg}. Ces modèles ont été appliqués séparément pour chaque secteur industriel. Pour l'estimateur de régression avec erreur emboîtée et l'estimateur de régression empirique de Bayes, il faut diviser les données par $\sqrt{x_k}$ afin d'obtenir un estimateur par quotient (EST/R).

Tableau 1. Biais Relatif Absolu (BRA) Moyen avec un seul groupe (G = 1)

	Secteur d'activité économique			
Estimateur*	Commerce de Détail	BTP	Hébergement	Autre
DIR	.023	.020	.036	.027
POS1/C	.085	.054	.265	.032
POS1/R	.108	.050	.270	.050
SYN1/C	.207	.173	.584	.337
SYN1/R	.324	.157	.414	.264
ARE1/R (h = 2)	.091	.047	.243	.078
N1/R	.203	.084	.263	.147
EB1/R	.176	.116	.380	.212

Tableau 2. Efficacité Relative (ER) Moyenne avec un seul groupe (G = 1)

	Secteur d'activité économique			
Estimateur	Commerce de Détail	BTP	Hébergement	Autre
DIR	1.00	1.00	1.00	1.00
POS1/C	1.34	1.35	1.29	1.18
POS1/R	1.24	1.86	1.86	1.64
SYN1/C	1.88	1.46	2.07	1.71
SYN1/R	2.27	2.92	3.26	2.04
ARE1/R (h = 2)	1.80	2.10	2.38	1.78
N1/R	2.08	2.38	2.56	2.39
EB/R	1.95	1.97	2.55	1.54

* C/R indique que le nombre d'entreprises / le revenu brut des entreprises pour chaque petite région est censé être connu.

Les propriétés inconditionnelles et conditionnelles de ces estimateurs ont été calculées pour chaque secteur d'activité économique.

Examinons en premier les propriétés inconditionnelles de ces estimateurs pour un seul groupe, c'est-à-dire, G = 1: le biais relatif absolu (BRA) et l'efficacité relative (ER) moyenne. Les résultats sont présentés aux tableaux 1 et 2.

Le biais relatif absolu (BRA) est donné par

$$BRA(\hat{Y}_{EST}) = \frac{1}{500D} \sum_{d=1}^{D} \left| \sum_{r=1}^{500} (\hat{Y}_{dEST}^{(r)} / Y_d - 1) \right|$$

où D est le nombre de petites régions. D est égal à 18 pour tous les secteurs, sauf pour l'hébergement, où il est égal à 16. $\hat{Y}_{dEST}^{(r)}$ est la valeur obtenue pour l'estimateur EST lors de la r-ième itération.

L'efficacité relative moyenne est donnée par

$$ER(\hat{Y}_{EST}) = [EQM(\hat{Y}_{DIR}) / EQM(\hat{Y}_{EST})]^{1/2},$$

où

$$EQM(\hat{Y}_{EST}) = \frac{1}{500D} \sum_{d=1}^{D} \sum_{r=1}^{500} (\hat{Y}_{dEST}^{(r)} - Y_d)^2$$

Les estimateurs sont classés par ordre de grandeur du biais relatif absolu, à savoir: DIR, ARE1/R, POS1/C, POS1/R, N1/R, EB1/R, SYN1/R et SYN1/C.

Par contre, tous les estimateurs sont plus efficaces que DIR. Tel que prévu théoriquement, l'estimateur synthétique SYN1/R est supérieur aux autres, car l'estimateur synthétique a la plus petite variance. Lorsque son biais est faible pour chacun des petits domaines, il a la plus petite erreur quadratique moyenne. Les estimateurs qui se servent de la variable auxiliaire x, (POS1/R et SYN1/R) sont en général plus efficaces que ceux qui suppose que la taille N_{dg} de la population est connue (POS1/C et SYN1/C). Ceci est surtout évident pour les secteurs où la corrélation entre y et x est forte (l'hébergement). Si on se restreint aux estimateurs du genre ESTG/R, leur rang (du meilleur ER au pire) est N1/R, ARE1/R, EB1/R et POS1/R. Le N1/R est supérieur aux autres estimateurs car sa variance est plus petite, quoique sa performance n'est que légèrement supérieure à celle de l'estimateur ARE1/R.

Passons maintenant aux caractéristiques conditionnelles. Nous examinerons le biais relatif conditionnel et la racine carrée de l'erreur quadratique moyenne conditionnelle. Ici le terme

"conditionnel" a une signification particulière. Une réalisation de l'échantillon s fournira n_d unités se situant dans un domaine quelconque U_d (petite région et secteur). Pour chacun des domaines U_d, si on se restreint dans nos calculs aux échantillons qui ont la même taille n_d, nos statistiques dépendront de cette taille et seront dites "conditionnelles". Les deux statistiques conditionnelles que nous avons examinées sont le biais relatif conditionnel (BRC)

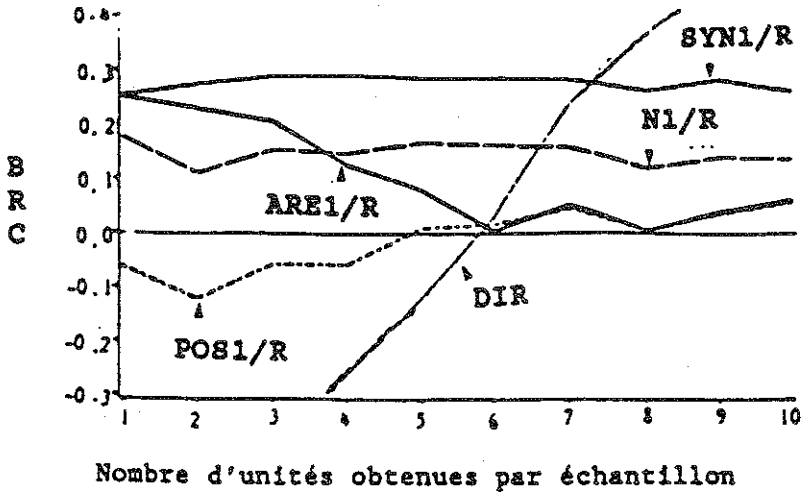
$$BRC\ (\hat{Y}_{dEST}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{Y}_{dEST}^{(r)}/Y_d - 1)$$

et la racine carrée de l'erreur quadratique moyenne conditionnelle (REQMC)

$$REQMC\ (\hat{Y}_{dEST}) = \left\{ \frac{1}{R} \sum_{r=1}^{R} (Y_{dEST}^{(r)} - Y_d)^2 \right\}^{1/2} .$$

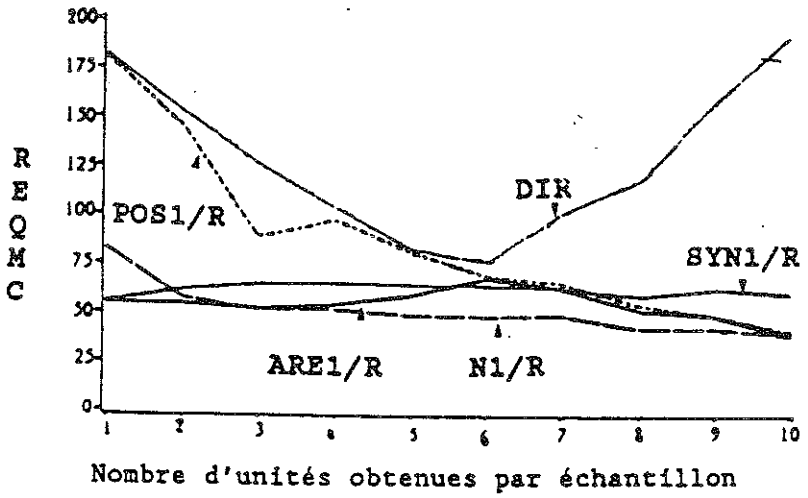
R est le nombre d'échantillons parmi les 500 tirés pour lesquels le nombre d'unités se situant dans le domaine U_d était n_d. On présente les résultats pour une seule région et un seule secteur, le commerce de détail. Pour ce domaine, $N_d = 23$ et l'espérance de la taille de l'échantillon est 23/4 = 5.75. Le graphique 1 présente le comportement du BRC tandis que le graphique 2 présente celui de la REQMC pour quelques estimateurs.

En ce qui concerne le BRC, les résultats empiriques appuient bien la théorie (voir le graphique 1). Le biais de l'estimateur DIR augmente de façon linéaire au fur et à mesure qu'on s'éloigne de la taille espérée $E(n_d)$. Le biais de l'estimateur synthétique (SYN1/R) est presque constant. L'estimateur par le quotient POS1/R est légèrement biaisé, quoique dans l'ensemble, les biais s'annulent. On peut attribuer le biais observé pour les tailles d'échantillon 1,2,9 et 10 au faible nombre d'observations obtenues (7,25,31 et 10). L'estimateur de régression avec erreur emboîtée, N1/R, a un biais constant égal à environ la moitié de celui associé à SYN1/R. L'estimateur de régression qui corrige le biais, ARE1/R, se comporte tel que prévu. Son biais conditionnel est sensiblement égal à celui de l'estimateur synthétique SYN1/R pour des échantillons de taille 1, et se rapproche de zéro au fur et à mesure que la taille de l'échantillon se rapproche de $E(n_d)$. Son biais est essentiellement égal à zéro lorsque cette taille est supérieure à $E(n_d)$.



Graphique 1: Biais Relatif Conditionnel (BRC) pour une région et le secteur commerce de détail

Les résultats pour la REQMC sont présentés dans le graphique 2. L'estimateur ayant la plus faible REQMC est N1/R; celle-ci a une valeur quasi constante, peu importe le nombre n_d d'unités obtenues. Les REQMC de ARE1/R et de SYN1/R se situent légèrement au-dessus de la REQMC de N1/R. La REQMC de DIR est une fonction quadratique de n_d et atteint, son minimum à $E(n_d) = 5.75$. La REQMC de cet estimateur est particulièrement élevée. L'estimateur par le quotient POS1/R n'a une faible REQMC que lorsque n_d s'approche de son maximum.



Graphique 2: Racine carrée de l'erreur quadratique moyenne conditionnelle (REQMC) pour une région et le secteur commerce de détail

4. ESTIMATEURS TEMPORELS

Dans les sections précédentes, nous avons examiné les méthodes les plus courantes utilisées pour obtenir des estimations pour les petits domaines. Ces méthodes se servent de données auxiliaires pour améliorer les estimations provenant d'une enquête ponctuelle. Cependant, de nombreuses enquêtes menées par des organisations telles que Statistique Canada, ont une caractère continu. On peut alors élaboré des méthodes qui peuvent tirer profit de l'information supplémentaire comprise dans une série chronologique.

Choudhry et Hidiroglou (1987) ont examiné le modèle suivant:

$$\bar{y}_{s_d}(t) = \boldsymbol{x}'_{U_d}(t)\,\boldsymbol{\beta} + \boldsymbol{z}'_d\boldsymbol{\nu} + e_d(t) \qquad (4.1)$$

où d = 1,2,...,D ; t = 1,2,...,T;

$\boldsymbol{x}'_{U_d}(t)$ est un vecteur de données auxiliaires de dimension p pour le petit domaine d au temps t. On suppose que le premier élément de $\bar{\boldsymbol{x}}'_{U_d}(t)$ est égal à 1 pour tous les d et t: ceci implique une ordonnée à l'origine dans le modèle (4.1); $\mathbf{z}'_d$ est un vecteur de dimension D-1, où le i-ième élément de $\mathbf{z}'_d$, z_{di}, est défini par

$$z_{di} = \begin{cases} 1 & si\ i \epsilon U_d \\ 0 & autrement \end{cases} .$$

Le vecteur $\boldsymbol{z}_d$ représente les effets fixes du modèle. On suppose aussi que les erreurs $e_d(t)$ suivent un modèle autorégressif,

$$e_d(t) = \rho e_d\,(t-1) + \varepsilon_d(t) \ , \qquad (4.2)$$

où le coefficient d'autocorrélation ρ ne dépend ni du domaine ni du temps. Les erreurs $\varepsilon_d(t)$ sont indépendantes et à distribution normale, avec moyenne zéro et variance σ_d^2. L'ensemble des équations (4.1) et (4.2) nous conduisent au modèle suivant:

$$y = X\beta + Z v + e \tag{4.3}$$

où $y = col_{1\le d\le D} col_{1\le t\le T} (\bar{y}_{s_d}(t))$,

$X = col_{1\le d\le D} col_{1\le t\le T} (x'_{D_d}(t))$,

$Z = col_{1\le d\le D}(z'_d) \otimes 1_T$, $1'_T = (1,1,\ldots,1)_T$,

et $e = col_{1\le d\le D} col_{1\le t\le T}(e_d(t))$

En plus,

$$E(e) = 0 , \quad Cov(e) = (\Sigma_D \otimes \Gamma) = R$$

où $\Sigma_D = diag\ (\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2)$ et Γ est une matrice de dimension TxT avec (i,j) -ème élément $\gamma_{ij} = (1-\rho^2)^{-1} \rho^{|i-j|}$. L'estimateur généralisé de moindres carrés de (β', v') est

$$\begin{pmatrix} \beta \\ \hat{v} \end{pmatrix} = \left(\begin{pmatrix} X' \\ Z' \end{pmatrix} R^{-1} (X, Z) \right)^{-1} \begin{pmatrix} X' \\ Z' \end{pmatrix} R^{-1} y \ . \tag{4.4}$$

Notons que (4.4) dépend de deux inconnues: ρ et σ_d^2 . Ces deux inconnues sont estimées en se servant de la méthode de Gauss-Newton [Hartley (1961)]. Une estimation de σ_d^2 pour chacune des régions s'obtient en se servant de l'estimateur non-pondéré de moindres carrés. On divise chaque résidu par $\hat{\sigma}_d$ et à partir de ces résidus pondérés, on obtient une estimation $\hat{\rho}$ de ρ . Ces estimations sont alors introduites dans l'équation (4.4) et le processus répété jusqu'à ce que les estimations de ρ et de σ_d convergent. Le meilleur "prédicteur" non-biaisé (MPNB) de y, selon Goldberger (1964), est

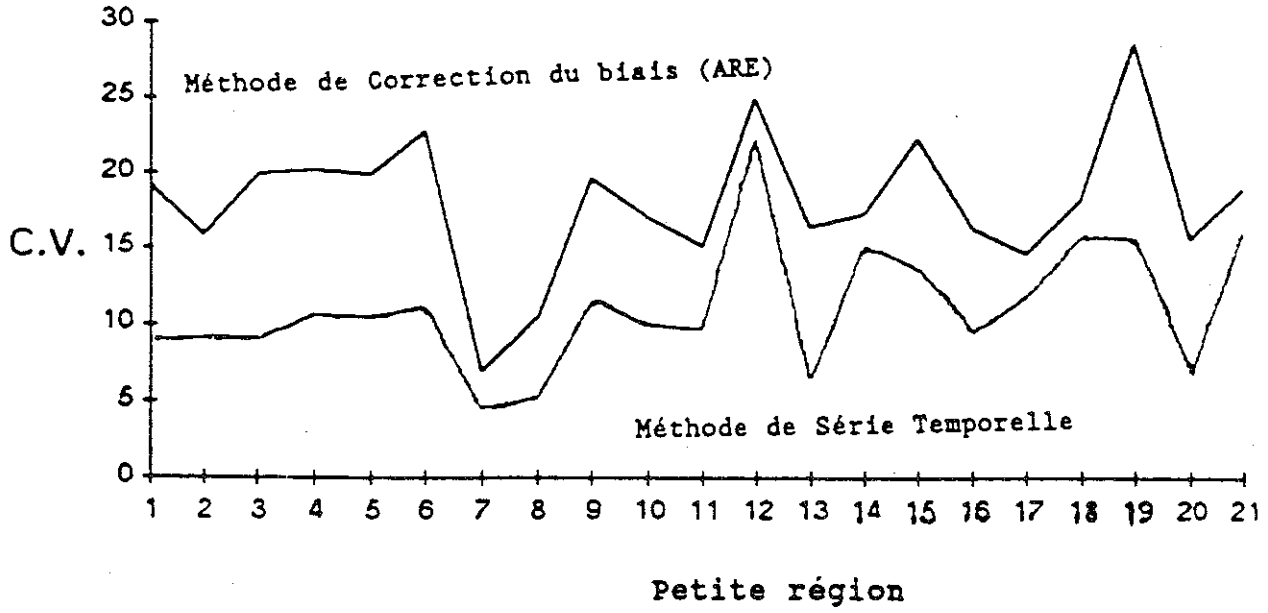
$$y^{*}_{s_d}(t) = \hat{y}_{s_d}(t) + \beta\,(y_{s_d}(t-1) - \hat{y}_{s_d}(t-1))$$
$$\text{où} \quad \hat{y} = (\boldsymbol{X}, \boldsymbol{Z})\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{v}} \end{pmatrix} \qquad (4.5)$$

La matrice de covariance pour y^{*} est

$$Cov(y^{*}) = (\boldsymbol{X}, \boldsymbol{Z})\left(\begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{pmatrix} R^{-1} (\boldsymbol{X}, \boldsymbol{Z})\right)^{-1}\begin{pmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{pmatrix} + \frac{\rho^2}{1-\rho^2}\Sigma_D \otimes I_T \qquad (4.6)$$

où I_T est une matrice identité de dimension T.

Cette méthode a été utilisée pour estimer le taux de chômage dans 21 petites régions de la Colombie-Britannique, pour lesquelles nous avions 36 mois (janvier 1983 à décembre 1985) de données. Les données auxiliaires utilisées dans la régression étaient $x_{U_{d,1}}(t)$ = log (nombre mensuel de bénéficiaires des prestations de l'assurance chômage), et $x_{U_{d,2}}(t)$ = l'estimation du taux d'activité, ainsi que les 20 variables dans $\mathbf{z}_d$ représentant l'effet de chacune des 21 petites régions. La variable que nous cherchons à estimer est le taux de chômage de chaque petite région $\bar{y}_{s_d}(t)$. Dans le modèle, on prend le logarithme naturel de cette variable. $x_{U_{d,2}}(t)$ est une estimation ayant une erreur d'échantillonnage, mais celle-ci est petite par rapport à celle de $\bar{y}_{s_d}(t)$. Pour ces données, l'estimation de l'autocorrélation ρ était 0.53 avec une écart-type de 0.03. Le coefficient de détermination pondéré (R^2) était 0.98. Le graphique ci-dessous montre la moyenne sur 36 mois du coefficient de variation estimé, pour un estimateur du type $\hat{Y}_{dARE}$ et pour l'estimateur temporel y^{*}. Pour l'ensemble des 21 petites régions, la moyenne sur les 36 mois était de 11% pour l'estimateur temporel, et de 18% pour l'estimateur du type $\hat{Y}_{dARE}$.



Graphique 3: Moyenne des coefficients de variation estimés du taux de chômage.

5. CONCLUSIONS

Nous avons passer en revue les principales méthodes d'estimations pour les petits domaines. Nous avons aussi examiné leur comportement, vis-à-vis du biais et de l'erreur quadratique moyenne, à l'aide d'une étude empirique. Parmi les méthodes étudiées la méthode qui corrige le biais possède plusieurs avantages. Elle est simple à utiliser et le biais conditionnel n'existe que lorsque la taille réalisée de l'échantillon (n_d) pour le domaine d est supérieur à la taille espérée ($E(n_d)$). De plus, on peut calculer des intervalles de confiance qui sont valables. Par contre, l'erreur quadratique moyenne pourrait être supérieure à celle de l'estimateur purement synthétique, de l'estimateur de régression avec erreur emboîtée, ou de l'estimateur de régression empirique de Bayes. Il faut souligner cependant que ces estimateurs visent précisement à minimiser cette erreur quadratique moyenne. En outre, ils ont quelques inconvénients: i) les estimations qui en résultent peuvent être sujettes à un biais conditionnel, peu importe la taille réalisée de l'échantillon, ii) il est alors impossible de construire un intervalle de confiance valable et iii) il existe souvent trop peu d'observations pour permettre une estimation fiable des composantes de la variance.

Prasad et Rao (1990) ont récemment englobé le modèle de régression avec erreur emboîtée et l'estimateur de régression empirique en se servant du modèle général linéaire avec erreur mixte dû à Henderson (1975). Ce modèle a la forme

$$\boldsymbol{y} = \boldsymbol{X}\,\beta + \boldsymbol{Z}\boldsymbol{v} + \boldsymbol{e}\,, \qquad (5.1)$$

où y est un vecteur d'observations provenant d'un échantillon, X et Z sont des matrices connues, et v et e sont des erreurs dont les distributions sont indépendantes, avec moyenne θ et matrices de covariance G et R, respectivement. Le meilleur "prédicteur" sans biais (MPSB) de $\mu = l'\beta + m'v$ est

$$t(\theta, y) = l'\hat{\beta} + m'GZ'V^{-1}(y - X\hat{\beta}) \qquad (5.2)$$

où $V = R + ZGZ'$ est la matrice de covariance de y et $\hat{\beta} = (X'V^{-1}X)^{-1}(X'V^{-1}y)$ est l'estimateur des moindres carrés généralisés de β. Prasad et Rao (1990) estiment l'erreur quadratique moyenne de $t(\hat{\theta}, y)$ où $\hat{\theta}$ est l'estimateur de θ obtenue en utilisant les matrices de variance estimées $\hat{G}$ et $\hat{R}$. Cette erreur quadratique moyenne est

$$MSE\,[\,t(\hat{\theta})\,] = MSE\,[\,t(\theta)\,] + E\,[\,t(\hat{\theta}) - t(\theta)\,]^2 . \qquad (5.3)$$

Les méthodes temporelles semblent être particulièrement prometteuses. Choudhry et Rao (1989) ont modifié le modèle (4.1) en supposant que dans $z_d' v$, $v = (v_1, v_2, \ldots, v_D)'$ représente des ordonnées à l'origine aléatoires. C'est-à-dire, $v \sim MVN\,(0, \sigma_v^2 I)$ où I est la matrice identité d'ordre D. Le modèle qui en résulte a la forme de l'équation (5.1) et β peut être estimé en se servant de (5.2). Cette version du modèle est, en moyenne, 2.5 fois plus efficace vis-à-vis de l'erreur quadratique moyenne, que celle dont il était question à l'équation (4.1).

Aucune des méthodes synthétiques décrites dans cet article ne sert à produire des estimations officielles à Statistique Canada. La seule méthode utilisée couramment est celle de Drew, Singh et Choudhry (1982). Cette méthode, qui ressemble beaucoup à celle de Särndal et Hidiroglou (1989), produit des estimations annuelles du taux de chômage pour les petites régions.

REMERCIEMENTS

L'auteur tient à remercier Yves Bélanger, Jean Dumais, Carl-Erik Särndal et Jackie Yiptong pour leurs commentaires utiles. L'auteur remercie particulièrement Georges Lemaître qui a revu et corrigé la version finale du texte.

BIBLIOGRAPHIE

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988), " An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28-36.

Binder, D.A. and Dick, J.P. (1989), "Modelling and Estimation of Repeated Surveys," *Survey Methodology*, 15, 29-45.

Choudhry, G.H. and Hidiroglou, M.A. (1987), "Small Area Estimation: Some Investigations at Statistics Canada," in *Proceedings of the 46th Session of the International Statistical Institute*, 1-19.

Choudhry, G.H. and Rao, J.N.K. (1989), "Small Area Estimation Using Models that Combine Time Series and Cross-Sectional Data," in Proceedings of the Statistics Canada Symposium on Analysis of Data in Time (à paraître).

Drew, J.D., Singh, M.P. and Choudhry, G.H. (1982)," Evaluation of Small Area Techniques for the Canadian Labour Force Survey," *Survey Methodology*, 8, 17-47.

Fay, R.E., and 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.

Goldberger, A.S. (1964), Econometric Theory, John Wiley and Sons: New York.

Hartley, H.D. (1961), "The Modified Gauss Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares," *Technometrics*, 3, 269-280.

Henderson, C.R. (1975)," Best Linear Unbiased Estimation and Prediction Under a Selection Model," *Biometrics*, 31, 423-447.

Hidiroglou, M.A., and Särndal, C.E. (1985), "An Empirical Study of Some Regression Estimators for Small Domains," *Survey Methodology*, 11, 65-77.

Holt, D., and Smith, T.M.F. (1979), "Post-Stratification," *Journal of the Royal Statistical Society*, Sec. A, 142, 33-46.

Prasad, N.G.N., and Rao, J.N.K. (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (1986), "Synthetic Estimators, SPREE and Best Model-Based Predictors of Small Area Means," Document interne à Carleton University, Ottawa, Canada.

Särndal, C.E. (1984), "Design-Consistent Versus Model-Dependent Estimators for Small Domains," *Journal of the American Statistical Association*, 79, 624-631.

Särndal, C.E., and Hidiroglou, M.A. (1989), "Small Domain Estimation: A Conditional Analysis.", *Journal of the American Statistical Association*, 84, 266-275.