# *PONDÉRATION ET ESTIMATION DANS LES ENQUÊTES-ENTREPRISES*

*M. A. Hidiroglou, E. Sarndal, et D. A. Binder, Statistique Canada*

## 1. Introduction

La fréquence des enquêtes-établissements peut être infra-annuelle (mensuelle, trimestrielle) ou annuelle; ces enquêtes visent à produire des estimations de totaux, de moyennes et de rapports de même que des estimations de variations d'une période à l'autre. De façon générale, l'objectif des enquêtes-établissements annuelles à Statistique Canada est de produire des données structurelles sur des variables comme la finance, la production, l'emploi et la propriété. Ces estimations sont publiées pour le plus bas niveau de détail pour lequel il existe une demande, pourvu qu'elles soient appuyées par des sources de données et que l'organisme statistique dispose des ressources nécessaires pour produire de telles estimations. Par ailleurs, les enquêtes infra-annuelles ont pour but de mesurer à point nommé les tendances économiques. Ces estimations sont publiées au niveau national et pour divers niveaux d'agrégation géographique ou industrielle. Cet article expose brièvement les plans de sondage le plus souvent utilisés pour les enquêtes annuelles et infra-annuelles. Nous étudions aussi des méthodes d'estimation ponctuelle et d'estimation de la variance pertinentes.

Au Canada, les enquêtes annuelles étaient auparavant réalisées à la manière d'un recensement; on effectuait une multitude d'envois postaux et de suivis dans le but d'obtenir le plus haut taux de réponse possible. Aujourd'hui, ces enquêtes prennent de plus en plus la forme d'enquêtes par sondage à cause des coûts prohibitifs et du lourd fardeau de réponse rattachés aux recensements. Dans les enquêtes annuelles, on procède généralement à un échantillonnage aléatoire simple stratifié de grappes (groupes d'unités) ou d'unités. Les strates sont définies en fonction d'un niveau d'agrégation industrielle et géographique approprié (strates primaires), puis en fonction de la taille des unités (strates secondaires). Les strates secondaires consistent en une strate à tirage complet et en plusieurs strates à tirage partiel, où se fait un échantillonnage. La strate à tirage complet est nécessaire parce que les distributions des variables économiques sont fortement asymétriques. Dans le cas des unités de la strate à tirage complet, les données sont recueillies au moyen d'une enquête faite directement auprès des établissements. Pour les unités des strates à tirage partiel, on recueille les données par enquête directe ou on consulte les fichiers administratifs pertinents.

Les plans d'échantillonnage utilisés pour les enquêtes infra-annuelles ressemblent à ceux des enquêtes annuelles. Cependant, à cause de la nécessité de produire des estimations justes de la variation entre deux périodes et d'alléger le

fardeau de réponse, il faut envisager une forme quelconque de renouvellement de l'échantillon. Les plans d'échantillonnage qui tiennent compte du caractère dynamique des bases de sondage et de la nécessité de réduire le fardeau de réponse ont été analysés dans Sunter (1977), Brewer, Early et Hanif (1984), Schiopu-Kratina et Srinath (1991), et Hidiroglou, Choudhry et Lavallée (1991). Le lecteur peut aussi consulter le chapitre rédigé par Nash et Monsour.

La plupart des enquêtes-entreprises ont pour objectif principal de produire des estimations ponctuelles non biaisées ou quasi non biaisées pour des variables telles que des totaux, des moyennes ou des rapports, ainsi que pour les mesures de précision correspondantes. Les estimations ponctuelles sont normalement nécessaires pour des domaines particuliers que l'on veut étudier. Un domaine peut être la population tout entière ou une sous-population particulière. Les domaines étudiés peuvent se confondre avec les strates d'échantillonnage ou peuvent les chevaucher partiellement. On mesure habituellement la précision des estimations pour domaines au moyen du *coefficient de variation* basé sur le plan, que l'on calcule en divisant l'écart-type estimé de l'estimation ponctuelle par l'estimation ponctuelle proprement dite et en exprimant le résultat en pourcentage. Les méthodes de pondération et d'estimation correspondantes reflètent le plan d'échantillonnage. Si cela convient, on peut se servir d'information supplémentaire pour accroître l'efficience des estimations. L'information supplémentaire peut provenir de fichiers administratifs mis à jour régulièrement ou peut être constituée de totaux annuels établis au moyen d'une enquête indépendante. Si les données auxiliaires tirées des fichiers administratifs sont corrélées raisonnablement avec la ou les variables étudiées, on peut intégrer cette information supplémentaire au processus d'estimation, celui-ci pouvant prendre plusieurs formes : méthode du quotient, méthode de stratification a posteriori, méthode de régression ou méthode itérative du quotient. Le calcul de ces estimateurs consiste, en clair, à déterminer, suivant une métrique particulière ou une fonction de distance, de "nouveaux" poids qui se rapprochent le plus possible des poids initiaux. Cela se fait de manière que si on applique ces nouveaux poids aux variables auxiliaires, on obtiendra des totaux d'échantillon qui concordent parfaitement avec les totaux auxiliaires pour la population. À ce propos, nous décrirons sommairement la théorie qu'ont élaborée Särndal, Swensson et Wretman (1992) et Deville et Särndal (1992). Nous illustrerons cette théorie par plusieurs techniques de pondération d'usage courant, en nous servant comme exemple d'enquêtes-entreprises annuelles et infra-annuelles de Statistique Canada.

L'estimation du niveau ou de la variation dans les enquêtes infra-annuelles soulève plusieurs questions méthodologiques, comme i) la non-réponse et ii) l'exactitude du calcul de la variance du rapport de deux estimations infra-annuelles (tendance), compte tenu de ce que la base de sondage et l'échantillon évoluent constamment. Ces questions seront aussi traitées dans cet article.

Pour calculer la variance et le coefficient de variation du rapport de deux estimations de niveau mensuelles (tendances), il faut connaître la covariance des deux estimations. Dans cet article, nous allons décrire une méthode pour calculer cette covariance suivant trois hypothèses : i) la population a changé dans l'intervalle à cause des créations et des disparitions; ii) la composition de l'échantillon a aussi changé à cause des créations, des disparitions et du renouvellement; et iii) aucune information supplémentaire n'est intégrée au processus d'estimation. Notons que cette covariance est aussi utile pour calculer le coefficient de variation d'agrégats comme les totaux annuels et pour faire concorder les totaux infra-annuels estimés avec les totaux annuels observés. L'opération qui consiste à harmoniser des séries infra-annuelles avec des valeurs annuelles s'appelle "étalonnage".

Le plan de l'article est le suivant : la section 2 expose la notation et les définitions nécessaires à l'étude de la pondération et de l'estimation dans les enquêtes-établissements. Dans la section 3, il est question de pondération au moyen de totaux auxiliaires; la théorie générale exposée par Särndal, Swensson et Wretman (1992) y est illustrée à l'aide de cas particuliers bien connus, comme l'estimateur de régression, l'estimateur de stratification a posteriori et l'estimateur de la méthode itérative du quotient. Nous examinons aussi des techniques de pondération appliquées dans le but de compenser la non-réponse totale. L'importante question de l'estimation par domaines est traitée dans la section 4. Quant à la section 5, elle traite la pondération et l'estimation dans les enquêtes à passages répétés; nous donnons l'exemple de l'estimation de la variance de l'écart entre deux totaux estimés, étant donné une population et un échantillon variables. Cette théorie peut s'appliquer à l'estimation de la variance dans le cas de l'étalonnage ainsi qu'à l'estimation composite. Enfin, la section 6 sert de conclusion.

## 2. Population, échantillon et groupes de modèle

Posons $U = \{1, ..., k, ..., N\}$ comme l'ensemble d'indices pour les N unités d'une population finie d'établissements. Une enquête est réalisée; nous désignons par $s$ un échantillon probabiliste d'unités prélevé dans $U$ au moyen d'un plan d'échantillonnage donné. Les probabilités de sélection qui découlent de ce plan sont désignées par $\pi_k = P(k \in s)$ et $\pi_{k\ell} = P(k \& \ell \in s)$. Nous supposons que ces probabilités sont connues et positives. Posons $a_k = 1/\pi_k$, c'est-à-dire le poids d'échantillonnage de l'unité k. En règle générale, on a recours à l'échantillonnage aléatoire simple stratifié dans les enquêtes-entreprises qui ont pour base de sondage une liste. Dans ce cas, $a_k = 1/f_h$ pour tous $k$ de la strate h, où $f_h = n_h/N_h$ est la fraction de sondage de la strate. Posons $y_k$ comme la valeur, pour l'unité de population k, de la variable étudiée y. Le total de $y$ pour la population est désigné par $Y = \Sigma\, y_k$. (Si $A$ est un ensemble d'unités, on écrit $\Sigma_A$ pour représenter $\Sigma_{k\in A}$; par exemple, $Y = \Sigma_{k\in U}\, y_k = \Sigma_U\, y_k$.) Un des objectifs de l'enquête est d'estimer Y. On a aussi besoin, en règle générale, d'estimations de totaux pour divers domaines à l'étude; les questions spéciales que soulève l'estimation par domaine sont traitées dans la section 4.

$$\hat{Y}_\pi = \Sigma_s\, a_k\, y_k$$

L'estimateur de Horvitz-Thompson (estimateur HT), peut être souvent amélioré grâce à l'utilisation d'information supplémentaire. L'information supplémentaire dont il est question dans cet article est constituée de totaux connus pour une ou plusieurs variables auxiliaires. Ces totaux peuvent se rapporter à la population entière ou à des sous-populations particulières. L'effectif d'une sous-population est un exemple simple de total connu. Notre objectif est d'utiliser le plus efficacement possible cette information dans le processus d'estimation.

Nous employons le terme groupe de modèle pour désigner une sous-population au sujet de laquelle on connaît un ou plusieurs totaux de variable auxiliaire. Nous désignons généralement le groupe de modèle par le symbole $U_p$, où $U_p = U$. Soit $x_{pk}$ la valeur pour l'unité k d'un vecteur auxiliaire $x_p$ rattaché à $U_p$. Plus précisément, nous appelons $U_p$ un groupe de modèle si :

a) la valeur auxiliaire $x_{pk}$ peut être observée pour chaque unité $k \in s_p = s \cap U_p$, et

b) le total auxiliaire pour le groupe, c.-à-dire. $X_p = \sum_{U_p} x_{pk} = N_p$ est connu.

Nous supposons que pour chaque unité $k \in s$, il est possible de déterminer le groupe de modèle auquel appartient $k$ et de calculer la paire de valeurs $(y_k, x_{pk})$, de sorte que l'on puisse effectuer une régression linéaire de $y$ par rapport à $x_p$ dans chaque groupe. Le vecteur $x_p$, pour lequel le total $X_p$ pour le groupe de modèle est connu, peut être composé de différentes variables dans les différents groupes, d'où la présence de l'indice p dans $x_p$. Pour une population d'établissements commerciaux par exemple, nous pouvons avoir $x_{1k}$ = revenu brut de l'entreprise $k$ dans le groupe de modèle $U_1$, $x_{2k}$ = nombre d'employés de l'entreprise $k$ dans le groupe de modèle $U_2$, et ainsi de suite. La condition est que l'on connaisse le revenu brut global des entreprises qui constituent $U_1$, que l'on connaisse le nombre total des employés des établissements qui constituent $U_2$, t ainsi de suite. Idéalement, $x_p$ est un bon prédicteur de la variable $y$ dans le groupe de modèle. Dans le cas élémentaire où $x_{pk} = 1$ pour tous $k \in U_p$, nous avons $X_p = \sum_{U_p} x_{pk}$, où $N_p$ est l'effectif du groupe de modèle. Les groupes de modèle correspondent donc à des strates formées a posteriori (voir section 3.1). La connaissance de l'effectif des groupes, $N_p$, peut accroître considérablement la précision des estimations.

## 3. Pondération d'observations à l'aide de totaux auxiliaires connus

L'utilisation d'information supplémentaire dans les enquêtes-établissements peut contribuer à accroître la précision des estimations ou bien à réduire la taille effective de l'échantillon. Cela est possible lorsque les variables auxiliaires sont bien corrélées avec les variables étudiées. Prenons, par exemple, l'Enquête sur l'emploi, la rémunération et les heures de travail (EERH), qu'effectue mensuellement Statistique Canada. Cette enquête repose sur un échantillonnage aléatoire simple stratifié avec renouvellement, la stratification étant faite selon la province (PROV), la branche d'activité (CTI) et la tranche d'effectif (TAILLE). Les strates qui servent à l'échantillonnage sont définies selon le niveau à 3 chiffres de la Classification type des industries (CTI3), par province et par tranche d'effectif (4 tranches). Schiopu-Kratina et Srinath (1991) décrivent en détail la méthode d'échantillonnage de l'EERH. Les échantillons de l'EERH sont tirés dans une population fortement asymétrique. Par conséquent, un des groupes de taille correspond à une strate à tirage complet qui renferme de grands établissements. Les autres groupes de taille correspondent à des strates à tirage partiel qui renferment de plus petits établissements et dans lesquelles se fait un échantillonnage. On s'est servi de l'estimateur HT pendant plusieurs années pour cette enquête. Maintenant que l'on dispose de variables auxiliaires bien corrélées provenant des fichiers administratifs, de nouvelles techniques seront utilisées après le remaniement de l'enquête.

Nous allons maintenant décrire une méthode générale de pondération pour des enquêtes où l'on dispose d'information supplémentaire. Nous supposons qu'il existe P groupes de modèle $U_p$, $p = 1, ..., P$, formant une partition de $U$, c'est-à-dire un ensemble de sous-populations disjointes et entières. Notons que l'on peut obtenir des totaux auxiliaires pour des niveaux d'agrégation plus détaillés que la partition ci-dessus. Cependant, cette partition représente un compromis de telle sorte qu'aucun groupe de modèle ne contient un trop petit nombre d'unités échantillonnées. Si

P = 1, la population entière constitue le seul groupe de modèle. Il suffit alors de connaître les totaux auxiliaires pour l'ensemble de la population.

L'échantillon $s$, tiré de la population $U$ suivant le plan d'échantillonnage donné, peut être divisé en groupes de modèle de la façon suivante : $s = s_p$, où $s_p = s \cap U_p$ est la portion de l'échantillon qui correspond au groupe de modèle p. Le poids d'échantillonnage $a_k = 1/\pi_k$ n'est pas le seul poids rattaché à l'unité k; il y a aussi le poids g. Celui-ci reflète le total auxiliaire connu $X_{U_p} = \sum_{U_p} x_{pk}$ relatif a groupe de modèle auquel appartient l'unité $k$. Le poids g pour l'unité k est défini par l'expression

$$g_k = 1 + (X_p - \hat{X}_{p\pi})(\sum_{s_p} a_k x_{pk} x'_{pk} / c_k)^{-1} x_{pk} / c_k \qquad (3.1)$$

si $k \in s_p$, $\hat{X}_{p\pi} = \sum_{s_p} a_k x_{pk}$ où est l'estimateur HT du total auxiliaire connu $X_p$ pour le groupe. (Dans cet article, les estimateurs identifiés par un "chapeau" et l'indice inférieur $\pi$ sont des estimateurs HT.) Les constantes connues $c_k$ sont déterminées par la structure de variance du modèle de régression hypothétique défini en (3.3) ci-dessous. Le poids total attribué à l'unité k est le produit des deux poids, $a_k$ (selon le plan) et $g_k$ (selon les données auxiliaires). En faisant tout d'abord la sommation à l'intérieur des groupes, puis pour l'ensemble des groupes, nous obtenons l'estimateur du total pour la population entière, $Y = \Sigma_U y_k$, c'est-à-dire

$$\hat{Y}_{GREG} = \sum_{p=1}^{P} \sum_{s_p} a_k g_k y_k \qquad (3.2)$$

La série de poids g calculés selon la formule (3.1) pour $p = 1, ..., P$ renferme l'information supplémentaire qui se rattache à l'ensemble particulier de groupes de modèle utilisés dans l'estimation.

Nous allons maintenant présenter deux méthodes de calcul qui aboutissent à l'estimateur (3.2).

### 3.1 Régression

Supposons que la population a été divisée en P groupes de modèle $U_p$, p = 1,...,P. Dans le cas de l'EERH par exemple, ces groupes pourraient être formés des strates d'échantillonnage originales ou pourraient correspondre à des sous-groupes de la population formés arbitrairement et pour lesquels il existe des totaux auxiliaires. Pour le groupe p, considérons le modèle de régression selon lequel :

$$y_k = x_{pk} \beta_p + \epsilon_k \text{ pour } k \in U_p \qquad (3.3)$$

où $E_\xi(\epsilon_k) = 0$, $Var_\xi(\epsilon_k) = c_k\,\sigma^2$, et $Cov_\xi(\epsilon_k, \epsilon_l) = 0$ pour tous $k \neq \ell$, l'indice inférieur $\xi$ désignant les moments par rapport au modèle. Dans le modèle ci-dessus, $\boldsymbol{\beta}_p$ est estimé au moyen de l'échantillon s par $\hat{\boldsymbol{B}}_p$, qui est défini comme la solution de

$$(\Sigma_{s_p} a_k\, x_{pk}\, x'_{pk} / c_k)\ \hat{\boldsymbol{B}}_p = \Sigma_{s_p} a_k\, x_{pk}\, y_k / c_k$$

$Y_p = \Sigma_{U_p} y_k$ C'est ce qui représente le système d'équations normales lorsque les données $\{(y_k, \boldsymbol{x}_{pk}) : k \in s_p\}$ servent à l'ajustement du modèle (3.3). Le but des poids $a_k$ dans ce système est de faire de $\hat{\boldsymbol{B}}_p$ un estimateur convergent selon le plan du vecteur des coefficients de régression, $\mathbf{B}_p$, pour la population -- en ajustement optimal (au sens des moindres carrés généralisés) -- lorsque toutes les unités de $U_p$ sont observées. L'ajustement par régression produit aussi les résiduels $e_k = y_k - \boldsymbol{x}'_{pk}\,\hat{\boldsymbol{B}}_p$ pour $k \in s_p = s \cap U_p$, $p = 1, \ldots, P$. Le total pour le groupe de modèle, est estimé par $\hat{Y}_{p\pi} + (\boldsymbol{X}_p - \hat{\boldsymbol{X}}_{p\pi})'\hat{\boldsymbol{B}}_p$ la somme de l'estimateur HT, $\hat{Y}_{p\pi} = \Sigma_{s_p} a_k\, y_k$ et d'un facteur de correction de régression, $(\boldsymbol{X}_p - \hat{\boldsymbol{X}}_{p\pi})'\hat{\boldsymbol{B}}_p$. Si nous voulons connaître l'estimateur du total pour la population entière, nous faisons une sommation par rapport aux groupes, c'est-à-dire

$$\hat{Y}_{GREG} = \Sigma_{p=1}^{P} \{\hat{Y}_{p\pi} + (\boldsymbol{X}_p - \hat{\boldsymbol{X}}_{p\pi})'\hat{\boldsymbol{B}}_p\} \tag{3.4}$$

Si cet estimateur est exprimé comme une somme linéaire pondérée appliquée à l'échantillon, $\Sigma_s w_k y_k$, on peut vérifier facilement que le poids $w_k$ est précisément $w_k = a_k g_k$, où $g_k$ est défini en (3.1).

Les résidus de la régression, $e_k$, sont nécessaires pour calculer l'estimation de la variance de $\hat{V}(\hat{Y}_{GREG})$ ou de $\hat{V}$ tout court. Cet estimateur de la variance est défini par l'expression

$$\hat{V} = \Sigma\Sigma_s (\Delta_{kl} / \pi_{kl})(g_k e_k / \pi_k)(g_l e_l / \pi_l) \tag{3.5}$$

où $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, $\pi_{kk} = \pi_k$ et $\Sigma\Sigma_s$ est une forme abrégée de la double sommation $\Sigma_{k \in s} \Sigma_{l \in s}$.

La justification théorique de la pondération des résidus par des poids g dans la formule (3.5) est donnée dans Särndal, Swensson et Wretman (1989). Bien que $\hat{V}$ soit défini comme une double sommation dans (3.5), il n'est pas calculé comme tel ordinairement dans la pratique. On ramène plutôt, pour chaque plan d'échantillonnage, le membre de droite de l'équation à une forme qui se prête au calcul. Prenons, par exemple, le cas de l'échantillonnage aléatoire

simple stratifié sans remise (ÉASSR). Alors, $s = \bigcup_{h=1}^{H} s_h$, où $s_h$ est un ÉASSR tiré de la strate h, h = 1,...,H.

Dans ce cas, (3.5) devient

$$\hat{V} = \sum_{h=1}^{H} N_h^2 \{(1-f_h)/n_h\} \sum_{s_h} (g_k e_k - \overline{ge}_h)^2/(n_h - 1) \tag{3.6}$$

où $\overline{ge}_h = \sum_{s_h} g_k e_k / n_h$ et $f_h = n_h/N_h$ est la fraction de sondage pour la strate h. Comme autre exemple ayant rapport à l'ÉASSR, voici

$$\hat{V} = N^2 \{(1-f)/n\} \sum_{s} (g_k e_k)^2/(n-1) \tag{3.7}$$

où f=n/N et $\overline{ge} = \sum_{s} g_k e_k / n$ est supposé égal à zéro. Notons que $\overline{ge} = 0$ lorsque $c_k$, dans la structure de variance du modèle, satisfait l'équation $c_k = \lambda' x_k$ pour tous k et pour un vecteur constant $\lambda$. Par exemple, pour la structure de variance homoscédastique, $Var_\xi(\varepsilon_k) = \sigma^2$ pour tous k, nous avons $\Sigma_s g_k e_k = 0$ si le modèle de régression contient un terme d'ordonnée à l'origine.

Une mesure de précision qu'utilisent couramment les organismes d'enquête est le coefficient de variation (de plan) estimé, que l'on désigne en abrégé par c.v.. On calcule le c.v. de l'estimateur GREG défini en (3.2) à l'aide de l'estimateur $V$ défini en (3.5), c'est-à-dire

$$cv = \{\hat{V}\}^{1/2} / \hat{Y}_{GREG} \tag{3.8}$$

**Exemple 3.1 :** Estimation par régression pour l'EERH

Tous les employeurs ont l'obligation de remettre les retenues sur la paye à Revenu Canada. Les données mensuelles correspondantes sont offertes à l'usage de Statistique Canada. Il existe une forte corrélation entre les versements mensuels (x) et les principales variables étudiées dans l'EERH, par exemple la rémunération et l'emploi (y). Cependant, cette corrélation est quelque peu amoindrie par l'irrégularité avec laquelle sont rapportées les données relatives aux versements. Lee et Croal (1989) concluent dans leur étude que, si les versements mensuels des déductions salariales servent de variable auxiliaire, l'estimateur par régression donnera de bien meilleurs résultats que l'estimateur HT pour les petites strates de l'EERH. Ils constatent cependant que l'estimation par régression convient pour un groupe PROV x CTI2 donné (qui est une agrégation de niveaux à quatre chiffres de la Classification type des industries) uniquement si deux conditions sont satisfaites : i) la taille de l'échantillon est au moins de 10, et ii) le degré de corrélation entre le nombre d'employés et les versements mensuels dépasse un certain seuil.

Les groupes de modèle sont définis au niveau PROV x CTI2 pour les strates de petite taille, c'est-à-dire que pour chaque groupe PROV x CTI2, un modèle de régression simple $y_k = \alpha + \beta x_k + \epsilon_k$ est ajusté. On peut décrire l'estimateur du total y au niveau PROV x CTI2 comme un estimateur par régression composé parce que les strates sont combinées pour l'ajustement du modèle. Pour une province (PROV) donnée, posons *h* comme l'indice du groupe de taille (TAILLE), *i* comme l'indice du groupe de la CTI (CTI2) et $s_{hi}$ comme l'échantillon du groupe *hi* (TAILLE x CTI2).

Les poids d'échantillonnage sont $a_k = N_{hi} / n_{hi}$ pour tous k dans la strate hi, où $N_{hi}$ et $n_{hi}$ désignent, respectivement, la taille de la population et la taille de l'échantillon pour la strate hi. Le coefficient de régression **B** pour la population est estimé par

$$\hat{B} = \sum_h \sum_i \sum_{k \in s_{hi}} a_k (y_k - \tilde{y}_s)(x_k - \tilde{x}_s) / \sum_h \sum_i \sum_{k \in s_{hi}} a_k (x_k - \tilde{x}_s)^2$$

où $\tilde{y}_s = \hat{Y}_{pi} / \hat{N}_{pi}$ , avec $\hat{N}_\pi = \sum_h \sum_i \sum_{k \in s_{hi}} a_k$ l'expression pour $\tilde{x}_s$ est analogue.

L'estimateur par régression du total Y pour un groupe PROV x CTI2 donné est

$$\hat{Y}_{REG} = \hat{Y}_\pi + \hat{B}(X - \hat{X}_\pi)$$

où $\hat{Y}_\pi = \sum_h \sum_i \sum_{k \in s_{hi}} a_k y_k$ et $\hat{X}_\pi = \sum_h \sum_i \sum_{k \in s_{hi}} a_k x_k$ sont les estimateurs HT des totaux *Y* et *X* respectivement.

Les poids g sont définis par l'expression

$$g_k = 1 + (X - \hat{X}_\pi)(x_k - \tilde{x}_s) / \sum_h \sum_i \sum_{k \in s_{hi}} a_k (x_k - \tilde{x}_s)^2$$

L'expérimentation de cet estimateur révèle des gains d'efficience appréciables par rapport à l'estimateur HT,

### 3.2 Stratification a posteriori

La stratification a posteriori est un cas particulier de la méthode de régression. Elle est souvent utilisée dans les enquêtes de grande envergure, principalement dans le but d'accroître l'efficience des estimateurs (voir Holt et Smith, 1979; Rao, 1985; Särndal et Hidiroglou, 1989; et Valliant, 1993). La stratification a posteriori peut contribuer à réduire sensiblement la variance, par comparaison à l'estimation HT ordinaire. L'exemple classique que l'on donne dans les manuels est celui de l'é.a.s. (échantillonnage aléatoire simple). Dans ce cas, l'estimateur de stratification a posteriori engendre des gains d'efficacité appréciables lorsque les moyennes des strates formées a posteriori sont très dispersées. De plus, il suscite l'intérêt des statisticiens par ses propriétés avantageuses dans le contexte de l'inférence conditionnelle (à ce sujet, voir Holt et Smith, 1979).

L'estimateur de stratification a posteriori est un cas particulier de (3.2). Il vient d'un modèle qui est un cas particulier de (3.3) en ce sens que $x_{pk} = 1$ pour tous $k \in U_p$. Autrement dit, ce modèle est

$$y_k = \beta_p + \epsilon_k \qquad \text{pour } k \in U_p \tag{3.9}$$

où $E_\xi(\epsilon_k) = 0, Var_\xi(\epsilon_k) = \sigma_p^2$ ,et $Cov_\xi(\epsilon_k, \epsilon_\ell) = 0$ pour $k \neq \ell$. Les groupes de modèle sont alors appelés «strates formées a posteriori». L'information supplémentaire dont il faut disposer dans ce cas est l'effectif des strates formées a posteriori, $N_p = \sum_{U_p} x_{pk}$ pour $p = 1, ..., P$. La formule générale (3.2) donne donc l'estimateur de stratification a posteriori

$$\hat{Y}_{POST1} = \sum_{p=1}^{P} N_p \tilde{y}_{s_p} \tag{3.10}$$

ù $\tilde{y}_{s_p} = \sum_{s_p} a_k y_k / \hat{N}_p$ , avec $\hat{N}_p = \sum_{s_p} a_k$ et $s_p = s \cap U_p$. On déduit l'estimateur de la variance de $\hat{Y}_{POST1}$ de l'expression (3.5) en posant, pour p=1,...,P, $e_k = y_k - \tilde{y}_{s_p}$ for $k \in s_p$ pour $k \notin s_p$. Les poids g sont définis $g_k = N_p / \hat{N}_p$ for all $k \in s_p$ , pour tous $k \notin s_p$. La stratification a posteriori est souvent utilisée dans les enquêtes-entreprises dès qu'on dispose d'une classification industrielle ou d'une classification selon la taille plus détaillées. Considérons une enquête où des strates sont constituées en fonction d'une "ancienne" classification et où d'autres strates, formées a posteriori, sont constituées en fonction d'une classification récente. Supposons que le plan est appliqué avec une fraction de sondage $n_h/N_h$ dans la strate h, $h = 1, ..., H$. Alors, $a_k = N_h/n_h$ pour tous k dans la strate h. En règle générale, les strates formées a posteriori recoupent toutes les strates d'échantillonnage. Posons $N_{hp}$ comme la portion de l'effectif de la strate d'échantillonnage $h$ contenue dans la strate "a posteriori" $p$ (case $hp$), de sorte que $N_h = \sum_{p=1}^{P} N_{hp}$, et posons $N_p = \sum_{h=1}^{H} N_{hp}$. Par conséquent, si nous supposons qu'il existe de l'information supplémentaire dans chaque strate formée a posteriori pour toutes les strates d'échantillonnage, de sorte que nous connaissons avec exactitude les effectifs $N_p$, mais que les effectifs $N_{hp}$ sont inconnus, l'estimateur (3.10) devient

$$\hat{Y}_{POST2} = \sum_{p=1}^{P} N_{.p} \frac{\sum_{h=1}^{H} N_{h.}\, n_{hp}\, \bar{y}_{s_{hp}} / n_{h.}}{\sum_{h=1}^{H} N_{h.}\, n_{hp} / n_{h.}} \qquad (3.11)$$

L'avantage que présente (3.11) par rapport à l'estimateur HT, $\sum_{h=1}^{H} N_{h.}\, \bar{y}_{s_h}$, est que si les strates sont considérablement périmées parce que de nombreux établissements ont changé de catégorie, l'estimateur (3.11) aura une variance beaucoup plus faible.

La situation est tout autre lorsqu'il existe de l'information supplémentaire pour chaque case, de sorte que les $N_{hp}$ sont connus. Dans ces circonstances, il faut connaître les effectifs qui découlent de la répartition des établissements de l'"ancienne" et de la "nouvelle" classifications selon la tranche de taille.
Le modèle pertinent dans ce cas est

$$y_k = \beta_{hp} + \epsilon_k$$

pour les unités k de la case *hp*. En supposant de nouveau un ÉASSR à l'intérieur des strates, nous déduisons de (3.2) un autre estimateur de stratification a posteriori

$$\hat{Y}_{POST3} = \sum_{h=1}^{H} \sum_{p=1}^{P} N_{hp}\, \bar{y}_{s_{hp}} \qquad (3.12)$$

**Exemple 3.2** : Stratification a posteriori pour l'EERH

Avant octobre 1990, l'échantillon de l'EERH provenait d'une base de sondage qui s'appuyait sur l'ancien registre des entreprises (RE) de Statistique Canada. En octobre 1990, la nouvelle «Base de données du registre central» (BDRC) était substituée au RE. L'ancienne et la nouvelle bases diffèrent l'une de l'autre sur plusieurs plans, notamment en ce qui concerne le système de classification des industries. Le codage des unités de l'ancienne base reposait sur la Classification des activités économiques de 1970 (CAÉ 1970), tandis que le codage des unités de la nouvelle base repose sur la Classification type des industries de 1980 (CTI 1980). En outre, les codes de taille et les codes géographiques appliqués aux unités de la nouvelle base sont plus à jour que ceux appliqués aux unités de l'ancienne base. En octobre 1990, un premier échantillon était tiré de cette nouvelle base. Une analyse a montré que cet échantillon reflétait mal la modification qu'avaient subie les codes de taille dans le processus de conversion. En effet, tandis que l'échantillon comptait 1.8% d'unités avec un code de taille plus élevé qu'auparavant, la nouvelle base de sondage en comptait 2.3%. Les unités de l'échantillon dont le code de taille était plus élevé qu'auparavant avaient des effectifs inférieurs à la moyenne observée pour les unités du groupe de taille correspondant. Puisque ces unités de taille relativement moins élevée étaient sous-représentées et qu'on a utilisé l'estimateur de Horvitz-Thompson, un biais par excès conditionnel a été introduit dans les estimations. La nouvelle base de sondage comptait 1.9% d'unités avec un code de taille moins élevé qu'auparavant. Or, d'après l'échantillon, on estimait cette proportion à 6.1%. Les unités de l'échantillon dont le code

de taille était moins élevé qu'auparavant avaient des effectifs supérieurs à la moyenne observée pour les unités du groupe de taille correspondant. Puisque ces unités de taille relativement plus élevée étaient surreprésentées, un biais par excès conditionnel a été introduit dans les estimations. C'est pourquoi on a décidé de recourir à la stratification a posteriori pour produire des estimations, en se servant des nouvelles caractéristiques de la population. Pour plus de détails sur la méthode de stratification a posteriori, voir Gossen et Latouche (1992).

Les unités contenues dans chacune des trois strates à tirage partiel (selon la nouvelle classification) ont fait l'objet d'une stratification a posteriori. Celle-ci était basée sur une comparaison de l'ancien code de taille du RE en vigueur en septembre 1990 et du nouveau code de taille en vigueur à compter d'octobre 1990. Les unités ont été réparties entre les strates selon que leur code de taille avait augmenté, avait diminué ou était demeuré le même pour chaque groupe CTI2 x TAILLE au Canada.

### 3.3 Calage

La méthode de régression décrite dans la section 3.1 est un moyen d'intégrer des données auxiliaires dans les estimations. Il y a aussi la technique du calage, qui consiste à trouver de nouveaux poids, $w_k$, qui se rapprochent le plus possible des poids initiaux, $a_k$. Ces nouveaux poids sont assujettis aux contraintes

$$\sum_{s_p} w_k x_{pk} = X_p \quad , \tag{3.13}$$

pour $p = 1, ..., P$, où $X_p$ est le total auxiliaire observé pour le groupe de modèle $U_p$. Autrement dit, nous faisons en sorte que les poids $w_k$ reproduisent $X_p$ pour chaque groupe, de manière que le total **X** pondéré pour tout l'échantillon corresponde au total de groupe, observé, $\hat{X}_p$.

Le calage crée une classe d'estimateurs qui comprend, entre autres, l'estimateur GREG (3.4). Il permet aussi de déterminer de nouveaux poids dont la valeur est limitée par des bornes inférieure et supérieure. Par exemple, on peut exclure la possibilité de poids négatifs, bien que l'addition de telles contraintes contribue à accroître l'écart entre les nouveaux poids et les poids initiaux.

Une métrique doit être spécifiée dans le but de quantifier la distance entre $w_k$ et $a_k$. Plusieurs fonctions de distance possibles sont considérées dans Deville et Särndal (1992). Notons-en deux en particulier :

1. la fonction de distance des moindres carrés généralisés (MCG)

$$F(w_k/a_k) = (w_k/a_k - 1)^2 / 2 \tag{3.15}$$

2. la fonction de distance de la méthode itérative du quotient (MIQ)

$$F(w_k/a_k) = (w_k/a_k)\log(w_k/a_k) - w_k/a_k + 1 \qquad (3.16)$$

D'autres fonctions de distance envisagées par ces auteurs garantissent l'existence de limites inférieure et supérieure pour les poids. On peut donc éliminer les poids négatifs et les poids positifs très élevés.

Une fonction de distance $F(w_k/a_k)$ doit répondre à une condition comme $F(1) = 0$, de sorte que si $w_k = a_k$, la distance sera nulle. Soit $f(z) = F'(z)$, la dérivée première de F. Nous devons appliquerla condition f(1) = 0. Nous minimisons donc la distance pondérée totale pour l'échantillon s, $\sum_s a_k F(w_k/a_k)$, étant donné la contrainte (3.14). Autrement dit, nous minimisons

$$\sum_s a_k c_k F(w_k/a_k) - \lambda'(\sum_s w_k x_k - X)$$

par rapport à $w_k$, où le vecteur $\lambda$ est un multiplicateur de Lagrange. La constante $c_k$ est nécessaire dans l'équation ci-dessus pour tenir compte des résidus qui découlent de l'ajustement de y par rapport à x et qui peuvent avoir des variances différentes. La pondération uniforme ($c_k = 1$) est susceptible d'être la plus courante dans les applications. En calculant la dérivée par rapport à $w_k$, en posant le résultat égal à zéro, puis en résolvant l'équation en fonction de $w_k$, nous obtenons $w_k = a_k\, g(\lambda' x_k/c_k)$, où $g = f^{-1}$ est la fonction inverse de f. Pour calculer les poids, nous commençons par déterminer la valeur de $\lambda$ en résolvant le système d'équations de calage déduit de (3.14),

$$\sum_s a_k g(\lambda' x_k/c_k) x_k = X \qquad (3.17)$$

En ce qui concerne la fonction de distance des MCG, cela donne $w_k = a_k g_k$, où $g_k$ désigne les poids g pour l'estimateur GREG définis en (3.1).

En ce qui a trait à la fonction de distance de la MIQ, $g(u) = e^u$, les équations de calage (3.17) peuvent être résolues itérativement. Il existe des logiciels conçus à cette fin. Par exemple, le programme CALMAR (Deville, Särndal et Sautory, 1993) résout les équations de calage par la méthode de Newton et calcule les poids $w_k = a_k g_k = a_k g(\lambda' x_k/c_k)$ pour plusieurs fonctions de distance, dont les deux mentionnées plus haut (MCG et MIQ). D'autres programmes conçus à cette fin sont M-WEIGHT, de Huang et Fuller (1978), et BASCULA (Göttgens et coll., 1991).

La théorie du calage peut s'appliquer lorsque l'information supplémentaire consiste en des fréquences marginales connues dans un tableau de fréquences de n'importe quelle dimension. La famille des fonctions de distance produit alors des estimateurs "généralisés de la méthode itérative du quotient". Lorsque la fonction de distance MIQ est utilisée, nous

obtenons l'estimateur de la méthode itérative du quotient défini par Deming et Stephan (1940). Prenons l'exemple d'un tableau à double entrée formé de r lignes et de c colonnes où les effectifs marginaux de population $r \times c$ sont , et $N_{+j} = \Sigma'_{i=1} N_{ij}, j = 1, ..., c$. Dans ces équations les $N_{ij}$ représentent les effectifs de case inconnus.

Le $\mathbf{x}_k$-vecteur correspondant peut s'écrire $x_k = (\delta_{1.k}, ..., \delta_{r.k}, \delta_{.1k}, ..., \delta_{.ck})'$ , où $\delta_{i.k} = 1$ égale 1 si l'unité k fait partie de la ligne i et zéro dans le cas contraire. De même, $\delta_{.jk}$ égale 1 si l'unité k fait partie de la colonne j et zéro dans le cas contraire. Dans ce cas, $X = \sum_{U} x_k = (N_{1.}, ..., N_{r.}, N_{.1}, ..., N_{.c})'$, ce qui correspond au vecteur des effectifs marginaux de population connus qui ont servi au calage.

Si nous posons $\lambda = (\lambda_{1.}, ..., \lambda_{r.}, \lambda_{.1}, ..., \lambda_{.c})'$ , alors $\lambda' x_k = \lambda_{i.} + \lambda_{.j}$ si l'unité k appartient à la case (i, j). De plus, si nous posons $c_k = 1$ pour tous k, alors $g(\lambda' x_k / c_k) = g(\lambda_{i.} + \lambda_{.j})$ et (3.17) amène le système d'équations suivant qu'il faut résoudre en fonction de $\lambda_{i.}$ et $\lambda_{.j}$ :

$$\sum_{j=1}^{c} \hat{N}_{ij} g(\lambda_{i.} + \lambda_{.j}) = N_{i.}; i = 1, ..., r$$

$$\sum_{i=1}^{r} \hat{N}_{ij} g(\lambda_{i.} + \lambda_{.j}) = N_{.j}; j = 1, ..., c$$

où $\hat{N}_{ij} = \sum_{s_{ij}} a_k$ est l'effectif estimé pondéré ordinaire de la case (i,j), $s_{ij}$ représentant la portion de l'échantillon incluse dans la case (i,j). Si nous résolvons (3.18) en fonction de $\lambda$, comme dans Deville et Särndal (1992), nous obtenons l'estimateur généralisé de la méthode itérative du quotient :

$$\hat{Y}_{RR1} = \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{N}_{ij}^{III} \bar{y}_{s_{ij}} \qquad (3.19)$$

où $\hat{N}_{ij}^{W} = \hat{N}_{ij} g(\lambda_{i.} + \lambda_{.j})$ est l'effectif de case estimé révisé et $\bar{y}_{s_{ij}} = \sum_{s_{ij}} a_k y_k / \sum_{s_{ij}} a_k$

Notons que les poids g sont définis $g_k = \hat{N}_{ij}^{W} / \hat{N}_{ij}$ pour toutes les unités k incluses dans la case (i,j).

Dans le cas particulier où la fonction de distance MIQ est utilisée, de sorte que $g(\lambda_{i.} + \lambda_{.j}) = \exp(\lambda_{i.} + \lambda_{.j})$, l'estimation $\hat{N}_{ij}^{W}$, contenue dans (3.19), peut être calculée par la méthode de l'ajustement proportionnel itératif de Deming et Stephan (1940). Cependant, la méthode de Newton converge plus rapidement que l'ajustement proportionnel itératif.

**Exemple 3.3** : Estimation par la méthode itérative du quotient dans l'Enquête sur le commerce de détail au Canada

L'estimation par la méthode itérative du quotient a été expérimentée dans l'Enquête mensuelle sur le commerce de détail au Canada. On forme un échantillon stratifié d'entreprises. Les strates sont définies selon trois critères : province, branche d'activité et taille de l'entreprise. Chaque province constitue un groupe de modèle (dans l'exposé qui suit, nous omettrons l'indice de la province). On cherche à estimer le total pour la province, $\Sigma\, y_k$. Les branches d'activité économique sont identifiées par l'indice $i = 1, ..., r$ tandis que les tranches de taille le sont par l'indice $j = 1, ..., c$. De plus, posons x comme une variable auxiliaire pour laquelle il existe des totaux par branche d'activité $X_{i.}$, i=1,...,r, et des totaux par tranche de taille $X_{.j}$, $j = 1, ..., c$. (Même s'ils sont connus, les totaux $X_{ij}$ pour chaque combinaison de branche d'activité et de tranche de taille n'ont pas servi au calage à cause du trop petit nombre d'unités que peuvent contenir certaines cases.) Nous voulons utiliser cette information supplémentaire dans la régression décrite dans la section 3.1. Le vecteur x figurant en (3.3) est de dimension r x c et est défini

$$x_k = (x_{1,k}, ..., x_{r,k}, ..., x_{1k}, ..., x_{ck})'$$

où $x_{i.k} = x_k$ si k fait partie de la branche d'activité i et $x_{ik} = 0$ dans le cas contraire, et où $x_{.jk} = x_k$ si k est compris dans la tranche de taille j et $x_{.jk} = 0$ dans le cas contraire. Posons $\beta = (\rho_1, ..., \rho_r, \gamma_1, ..., \gamma_c)'$. Nous avons $x_k'\beta = (\rho_i + \gamma_j) x_k$ pour $k \in U_{ij}$. Notons que $\sum_U x_k = (X_{1.}, ..., X_{r.}, X_{.1}, ..., X_{.c})'$, ce qui correspond exactement à l'information utilisée pour le calage dans cet exemple. Dans le cas présent, la formule (3.3) s'écrit

$$y_k = (\rho_i + \gamma_j) x_k + \epsilon_k$$

pour $k \in U_{ij}$, où nous supposons $E_\xi(\epsilon_k) = 0$, $V_\xi(\epsilon_k) = \sigma^2$ que pour tous $k \in U$. Le "balayage" (ou calage) produit des poids g qui peuvent être définis par l'expression

$$g_k = \hat{X}_{RAK,ij} / \sum_{s_{ij}} a_k \; x_k$$

$k \in U_{ij}$, où $\hat{X}_{RAK,ij}$ est l'estimation du total par case $X_{ij}$ calculée au moyen du balayage. Le cas se complique par le fait qu'une correction de poids est nécessaire pour tenir compte des disparitions d'entreprises dans la base de sondage.

Nous multiplions chaque poids g dans la case ij par $\hat{N}_{ij} / N_{ij}$ ,où $\hat{N}_{ij} = \sum_{s_{ij}} a_k$ est l'estimation du nombre réel d'unités, $N_{ij}$, contenues dans la case ij et $N^*_{ij}$ est l'effectif de la case d'après la base de sondage, si des disparitions sont survenues parmi les unités d'une case, $N_{ij} < N^*_{ij}$. Le poids final pour chaque $k \in U_{ij}$ est $\{\hat{N}_{ij} / N_{ij}\} \{\hat{X}_{RAK,ij} / (\sum_{s_{ij}} a_k x_k)\}$, et l'estimateur correspondant du total pour la province est

$$\hat{Y}_{RAK,ij} = \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{N}_{ij} / N_{ij} \} \{\hat{X}_{RAK,ij} / (\sum_{s_{ij}} a_k x_k)\} \sum_{s_{ij}} a_k y_k$$

Cet estimateur se révèle beaucoup plus efficace que l'estimateur HT habituel. En effet, le coefficient de variation passe de 0.08% à 0.05% au niveau national.

### 3.4 Non-réponse

Même si l'on fait des efforts raisonnables pour obtenir un taux de participation de 100% dans les enquêtes, on enregistre toujours un certain niveau de non-réponse dans la pratique. Dans certains cas, la non-réponse est totale (non-réponse au questionnaire) alors que dans d'autres cas, il manque des données sur quelques-unes des variables étudiées (non-réponse partielle). En règle générale, la non-réponse partielle est compensée au moyen de l'imputation (voir le chapitre de Kovar). Dans le présent rapport, nous concentrons notre attention sur la non-réponse au questionnaire. Même dans les cas de non-réponse totale, il arrive souvent que l'on dispose d'une information supplémentaire utile pour améliorer l'estimation. Il peut s'agir d'une information d'ordre typologique ou géographique, d'une information relative à la taille des unités ou encore d'une information sur la base de sondage. Cette information sert habituellement à réduire le biais de non-réponse. Dans le cas d'un échantillon aléatoire stratifié, la méthode de compensation de la non-réponse la plus simple consiste à modifier les fractions de sondage de strate en fonction du nombre total de répondants. Cela revient à imputer des moyennes de strate ou à effectuer une correction par pondération à l'intérieur des strates. Cette méthode équivaut aussi à la stratification a posteriori, où les strates formées a posteriori correspondent aux strates initiales. D'une manière générale, toutes les méthodes de traitement de la non-réponse sont en quelque sorte une forme de correction par pondération. Par exemple, dans les enquêtes à passages répétés, on recourt souvent à l'imputation par quotient; dans ce cas, on applique le quotient (ou le ratio) à une valeur historique relative à l'unité non répondante. Le ratio est estimé à l'aide des données fournies par les unités qui ont participé à l'enquête dans la période courante et pour lesquelles il existe des données relatives aux périodes précédentes. Là encore, ce procédé peut être assimilé à une correction de poids.

En analysant la correction de poids pour non-réponse, Oh et Scheuren (1983) recommandent de considérer le mécanisme de réponse comme un autre volet du plan de sondage probabiliste. Cette idée est reprise par Särndal et Swensson (1987), qui voient le mécanisme de réponse comme une seconde phase d'échantillonnage. Comme modèle

d'échantillonnage simple et efficace, supposons un échantillonnage de Bernoulli (résultats indépendants et identiquement distribués) dans des classes de pondération pour les résultats de réponse. Cette hypothèse nous amène à repondérer chaque classe de pondération en fonction de la probabilité de réponse estimée. Särndal et Swensson (1987) examinent l'estimation de la variance suivant ce scénario en se référant à la théorie de l'échantillonnage à deux phases. Oh et Scheuren (1983) soulignent aussi que cette repondération équivaut à des ajustements de stratification a posteriori.

Bethlehem (1988) soutient que l'utilisation de l'estimateur $\hat{Y}_{GREG}$ en (3.4), où les probabilités de sélection sont corrigées en fonction des probabilités de réponse, aura pour effet de réduire le biais de non-réponse. De fait, si les coefficients de régression sont les mêmes pour les répondants et les non-répondants, le biais disparaît. Ce phénomène est également souligné par Thomsen (1973) en ce qui concerne l'estimateur de stratification a posteriori $\hat{Y}_{POST}$.

Ces observations sont importantes dans la pratique puisque la modélisation du mécanisme de non-réponse est une notion inconnue pour de nombreuses enquêtes. On suppose plutôt que les probabilités de réponse sont les mêmes pour tout l'échantillon, ce qui explique que l'on doit compter sur des estimateurs de régression comme $Y_{GREG}$ pour réduire le biais. Little (1986) souligne qu'une modélisation explicite des probabilités de réponse peut avoir pour effet d'amplifier la variance des estimations. Il propose qu'on examine la relation entre les variables étudiées et la probabilité de réponse estimée et qu'on utilise une méthode empirique de Bayes pour l'estimation. C'est une approche peu courante dans la pratique. De toute évidence, on obtiendrait ainsi des poids différents selon les variables, une pratique peu recommandée pour les enquêtes de grande envergure.

La modélisation des probabilités de réponse est utile lorsqu'on n'est pas sûr de la validité du modèle de régression tant pour les répondants que pour les non-répondants. Elle est aussi commode lorsqu'il s'agit de vérifier si les méthodes plus simples peuvent donner des résultats satisfaisants. En règle générale, les méthodes de régression logistique offrent une classe variée de modèles pour l'estimation de la probabilité de réponse; l'échantillonnage de Bernoulli dans des classes de pondération est un exemple particulier de ces modèles. Ceux-ci sont spécialement utiles lorsque les données auxiliaires offrent un bon pouvoir discriminatif par rapport aux probabilités de réponse. L'utilisation des poids initiaux dans l'ajustement de ces modèles pour les probabilités de réponse n'a pas fait l'objet de suffisamment de recherches jusqu'à maintenant.

## 4. Estimation de totaux de domaines

Les *domaines* sont des sous-populations pour lesquelles on cherche à obtenir des estimations ponctuelles de totaux, de moyennes ou d'autres paramètres et les mesures de précision correspondantes. Il ne faut pas confondre les domaines avec les groupes de modèle ou les strates. Les groupes de modèle et les strates sont aussi des sous-populations mais ils diffèrent des domaines sur le plan conceptuel. Comme avant, s désigne l'échantillon prélevé dans la population finie $U = \{1, ..., k, ..., N\}$ selon un plan d'échantillonnage donné. Les probabilités de sélection sont $\pi_k$ et $\pi_{kl}$ et, comme précédemment, $a_k = 1/\pi_k$ désigne le poids d'échantillonnage de l'unité k.

Désignons par $s_{(d)} = s \cap U_{(d)}$ la portion de l'échantillon s qui fait partie du domaine $U_{(d)}$. Sauf dans de rares situations, où les conditions sont contrôlées (par ex., lorsque $U_{(d)}$ équivaut à une strate), la taille de $s_{(d)}$ sera aléatoire.

Les données y observées dans le domaine sont $\{y_k : k \in s_{(d)}\}$. Dans beaucoup de cas, on peut allier de l'information supplémentaire à ces données dans le but de produire des estimations plus précises. Nous considérons ici le problème d'estimation suivant. Supposons que $x_{pk}$ est un vecteur auxiliaire dont le total $X_p = \sum_{U} x_{pk}$ est connu pour

des groupes de modèle déterminés $U_p$, $p = 1, ..., P$, qui forment une partition de la population U. Nous servons des données $\{(y_k, \boldsymbol{x}_k) : k \in s_{(d)}\}$ et des totaux $X_p$, $p = 1, ..., P$, pour estimer le total de domaine $Y_{(d)} = \sum_{U_{(d)}} y_k$.

Un domaine $U_{(d)}$ peut se rapporter aux groupes de modèle de diverses façons. Par exemple, supposons qu'une enquête nationale vise à établir des estimations pour plusieurs domaines définis comme des divisions de recensement. Supposons aussi qu'un échantillon national s est tiré dans la population entière.

Quatre situations sont possibles :

a) domaine = division de recensement = groupe de modèle;

b) domaine = division de recensement; groupe de modèle = population entière = pays;

Dans la situation (a), le total auxiliaire connu se rapporte à un niveau de détail particulier, notamment au domaine proprement dit, tandis que dans la situation (b), le total auxiliaire se rapporte au niveau d'agrégation maximum, c'est-à-dire à la population totale. Entre ces deux extrêmes, il existe des situations intermédiaires comme les suivantes :

c) domaine = division de recensement; groupe de modèle = région qui englobe la division de recensement;

d) domaine = division de recensement; groupes de modèle = deux régions non chevauchantes qui, si elles sont réunies, englobent la division de recensement.

Bien que, dans les situations (c) et (d), l'information supplémentaire ne porte pas sur les groupes de modèle proprement dits, cette information est utile et ne doit pas être laissée de côté. La question est de savoir comment l'utiliser le mieux possible dans le calcul d'estimations pour le domaine (division de recensement).

Dans de nombreuses applications, on trouve D domaines $U_{(d)}$, d = 1, ..., D, qui forment une partition de U. L'échantillon global s peut donc, lui aussi, être divisé :

$$s = \bigcup_{d=1}^{D} s_{(d)}$$

On peut aussi définir les cellules d'échantillon. La cellule d'échantillon *dp* est définie comme $s_{(d)_p} = s \cap U_{(d)} \cap U_p = s_p \cap U_{(d)}$, c'est-à-dire l'ensemble des unités de l'échantillon qui appartiennent au domaine $U_{(d)}$ et au groupe de modèle $U_p$.

Passons maintenant à l'estimation du total de domaine, $Y_{(d)} = \Sigma\, y_k$. Une pratique courante dans l'estimation pour domaine est d'introduire une variable de domaine, désignée par $y_{(d)}$, dont la valeur pour l'unité k est définie comme suit :

$$y_{(d)k} = \begin{cases} y_k & \text{if } k \in U_{(d)} \\ 0 & \text{if } k \in U_{(d)} \end{cases} \qquad (4.1)$$

On peut alors exprimer le total de domaine $Y_{(d)}$ comme la somme, pour la population U, des valeurs de la variable de domaine $y_{(d)}$, c'est-à-dire,

$$Y_{(d)} = \Sigma_U y_{(d)k}$$

On peut alors exprimer le total de domaine $Y_{(d)}$ comme la somme, pour la population U, des valeurs de la variable de domaine $y_{(d)}$, c'est-à-dire

$$Y_{(d)} = \Sigma_U y_{(d)k}$$

La procédure d'estimation de $Y_{(d)}$ décrite ci-dessous est tirée de la communication de Estevao, Hidiroglou et Särndal (1992); il s'agit d'une estimation de type GREG fondée sur un plan. On calcule d'abord une série de poids g, $g_k$, selon la formule (3.1) pour chaque groupe de modèle, $p = 1, ..., P$. On applique ensuite les poids $a_k g_k$ aux valeurs $y_k$ observées pour le domaine afin de calculer l'estimateur GREG pour domaine

$$\hat{Y}_{(d)GREG} = \Sigma_{p=1}^{P} \Sigma_{s_p} a_k g_k y_{(d)k} \qquad (4.2)$$

Notons que les poids g sont des fonctions de totaux auxiliaires connus pour un niveau d'agrégation quelconque (niveau du groupe de modèle); ce peut être un niveau supérieur au niveau du domaine. Une manière simple de décrire l'expression (4.2) est de dire qu'elle équivaut à (3.2), à la différence que $y_k$ est remplacée par la valeur de la variable de domaine, $y_{(d)k}$. On peut aussi décrire (4.2) de la façon suivante, en supposant que tous les poids g ont été calculés à l'avance selon la formule (3.1) :

1. Déterminer les groupes de modèle sécants pour le domaine $U_{(d)}$, c'est-à-dire les groupes de modèle $U_p$ tels que $U_{(d)} \cap U_p$ est non vide;
2. Si $U_p$ est un groupe de modèle sécant pour $U_{(d)}$, appliquer le poids $a_k g_k$ à la valeur $y_{(d)k}$, faire la somme pour tous les éléments $k \in s_p = s \cap U_p$;
3. Faire la somme pour tous les groupes de modèle sécants; on obtient alors l'estimateur GREG pour domaine, $\hat{Y}_{(d)GREG}$, défini en (4.2).

Le concept des groupes de modèle qui se recoupent est important pour l'estimation de la variance de $\hat{Y}_{(d)GREG}$. L'estimateur de la variance $V(\hat{Y}_{(d)GREG})$ est désigné en abrégé par $\hat{V}_{(d)}$. On le calcule par la formule

$$\hat{V}_{(d)} = \Sigma\Sigma_s (\Delta_{kt}/\pi_{kt})\{g_k e_{(d)k}/\pi_k\}\{g_t e_{(d)t}/\pi_t\} \qquad (4.3)$$

où $e_{(d)k} = y_{(d)k} - x'_{pk}\hat{B}_{(d)p}$ pour $k \in s$. Nous reconnaissons dans cette formule l'expression (3.5), à la différence que $y_k$ est remplacée par la valeur de la variable de domaine, $y_{(d)k}$. Notons que cette substitution implique que l'on remplace $\hat{B}_p$ par $\hat{B}_{(d)p}$, qui est défini comme la solution de

$$(\sum_{s_p} a_k x_{pk} x'_{pk} / c_k)\,\hat{B}_{(d)p} = \sum_{s_p} a_k x_{pk} y_{(d)k} / c_k \qquad (4.4)$$

Trois types de résidus $e_{(d)k}$ entrent dans le calcul de (4.3). Les deux premiers types se rapportent aux unités-échantillon k qui appartiennent à des groupes de modèle sécants; le troisième type se rapporte aux unités k qui appartiennent à des groupes de modèle non sécants. Plus précisément, étant donné $s_p = s \cap U_p$, nous avons

$$e_{(d)k} = \begin{cases} y_k - x'_{pk}\hat{B}_{(d)p} & \text{si } k \in s_p,\ U_{(d)} \cap U_p \text{ si non vide, et } k \in U_{(d)}; \\ -x'_{pk}\hat{B}_{(d)p} & \text{si } k \in s_p,\ U_{(d)} \cap U_p \text{ si non vide, et } k \notin U_{(d)}; \\ 0 & \text{si } k \in s_p,\ U_{(d)} \cap U_p \text{ si vide} \end{cases} \qquad (4.5)$$

Le fait que $e_{(d)k}$ est nul pour tous les k appartenant à des groupes de modèle non sécants simplifie le calcul de $\hat{V}_{(d)}$. Par exemple, si l'échantillon s est prélevé par ÉASSR, (4.3) devient

$$\hat{V}_{(d)} = N^2 (1-f)/n \sum_s (g_k e_{(d)k} - \overline{ge}_{(d)})^2/(n-1)$$

où $\overline{ge}_{(d)} = \sum g_k e_{(d)k} / n$.

Le coefficient de variation de plan est calculé exactement comme en (3.8), c'est-à-dire,

$$cv_{(d)} = \{\hat{V}_{(d)}\}^{1/2} / \hat{Y}_{(d),GREG}$$

Dans la pratique, il est important de calculer $cv_{(d)}$ pour tous les domaines étudiés. Dans certains cas, il arrive que le coefficient de variation excède le maximum acceptable pour publication, par exemple $cv_{(d)} > 25\%$. Cela peut se produire lorsque le domaine contient peu d'observations ou que l'information supplémentaire n'est pas suffisante. Si on décide de ne pas publier les estimations $\hat{Y}_{(d)GREG}$ pour tous les domaines ou pour quelques-uns, on peut envisager des méthodes d'estimation non fondées sur un plan, comme l'estimation synthétique. Cependant, si on publie des estimations ponctuelles et des estimations de la variance qui ne sont pas basées sur un plan, il faut prendre soin de mentionner que des méthodes non standard ont été utilisées.

Plusieurs remarques s'imposent ici.

1. Principe de calcul. Les calculs faits pour un domaine imitent ceux qui sont effectués pour la population entière. Dans le cas de l'estimation ponctuelle, la substitution de $y_{(d)k}$ à $y_k$ pour $k \in s$ implique que (3.2) devient (4.2). En ce qui concerne l'estimation de la variance, la substitution de $y_{(d)k}$ à $y_k$ pour $k \in s$ suppose

automatiquement le remplacement de $e_k$ par $e_{(d)k}$ pour $k \in s$, et (3.5) devient (4.3). Autrement dit, le calcul de l'estimateur pour domaine (4.2) et de l'estimateur de la variance correspondant (4.3) s'effectue de façon formelle en remplaçant la variable étudiée y par la variable de domaine $y_{(d)}$, définie en (4.1). Le calcul gagne ainsi en simplicité.

2. Nature des équations normales. Les équations normales (4.4) correspondent systématiquement à l'ajustement de la droite de régression de $y_{(d)}$ (la variable dépendante pour domaine) en $x_p$ (la variable explicative) au moyen des données-échantillon du groupe p. Cet ajustement peut être médiocre parce que $y_{(d)}$ n'est pas une variable dépendante naturelle : elle équivaut à la variable y à l'intérieur du domaine mais elle a toujours une valeur nulle à l'extérieur. Cependant, la qualité de l'ajustement au niveau du domaine n'est pas notre principale préoccupation ici. Ce que nous visons plutôt, c'est de pouvoir utiliser des poids g qui, premièrement, produisent des estimations pour domaine additives (voir la remarque 4) et, deuxièmement, ne varient pas d'un domaine à l'autre, ce qui crée des avantages sur le plan du calcul et permet de calculer d'autres estimations pour domaine que celles publiées par l'organisme statistique. Pour connaître d'autres estimateurs pour domaine, veuillez vous référer à Särndal, Swensson et Wretman (1992, p. 408).

3. Convergence selon le plan. Si l'on obtient des estimations précises pour les domaines, c'est grâce à la propriété de convergence selon le plan. Nous savons que $\hat{Y}_{\text{GREG}}$, défini en (3.2), est un estimateur convergent selon le plan du total Y pour la population. Cela signifie en gros que, peu importe la configuration des valeurs de population finie $(y_1, \ldots, y_k, \ldots, y_N)$, $\hat{Y}_{\text{GREG}}$ se rapprochera très vraisemblablement de Y si la taille de l'échantillon est grande, parce que $g_k$ tend vers 1 pour de grands échantillons. Cette propriété vaut donc en particulier pour le vecteur pour domaine $(y_{(d)1}, \ldots, y_{(d)k}, \ldots, y_{(d)N})$. Ainsi, $\hat{Y}_{(d)\text{GREG}}$, défini en (4.2), est un estimateur convergent selon le plan du total pour domaine $Y_{(d)}$. De la même manière, $\hat{V}$, défini en (3.5), est un estimateur de la variance convergent selon le plan, c'est-à-dire que $\hat{V}$ se rapprochera très vraisemblablement de la variance de $\hat{Y}_{\text{GREG}}$ si l'échantillon est grand, quelle que soit la configuration des valeurs y. Par conséquent, si nous calculons la formule $V$ pour le vecteur pour domaine $(y_{(d)1}, \ldots, y_{(d)k}, \ldots, y_{(d)N})$, ce qui donne $\hat{V}_{(d)}$, en (4.3), nous avons un estimateur de la variance convergent selon le plan pour $\hat{Y}_{(d)\text{GREG}}$.

4. Propriété d'additivité. Supposons que nous voulons estimer un total pour l'un et l'autre de D domaines $U_{(d)}$, $d = 1, \ldots, D$, formant une partition de U. Alors, $\hat{Y}_{GREG} = \sum_{d=1}^{D} \hat{Y}_{(d)GREG}$, où $\hat{Y}_{GREG}$ et $\hat{Y}_{(d)GREG}$ sont définis en (3.2) et en (4.2) respectivement. Cette relation signifie que la somme des estimations par domaine est égale à l'estimation calculée pour la population. La propriété d'additivité a été introduite pour répondre aux besoins des utilisateurs de statistiques officielles. On la déduit facilement de la relation $\sum_{d=1}^{D} y_{(d)k} = y_k$, pour tous $k \in U$.

Pour connaître d'autres estimateurs utiles qui répondent à cette propriété, veuillez vous référer à Särndal, Swensson et Wretman (1992, pp. 397-413).

**Exemple 4.1** Échantillonnage à deux phases de dossiers fiscaux pour les enquêtes économiques

On établit les estimations annuelles de la production économique au Canada en combinant des estimations qui proviennent de deux sources : les grandes entreprises et les petites entreprises. En ce qui concerne les grandes entreprises, on procède par enquête postale. Dans le cas des petites entreprises, l'estimation repose sur un échantillonnage à deux phases de dossiers fiscaux, que nous décrivons ici et que traitent Choudhry, Lavallée et Hidiroglou (1989) et Armstrong, Block et Srinath (1993). Les principales caractéristiques du plan de sondage sont les suivantes :

1. échantillonnage de Bernoulli dans chacune des deux phases;
2. stratification a posteriori des échantillons prélevés dans chacune des phases;
3. calcul d'estimations pour la population des entreprises et pour divers domaines définis selon la CTI, la province, le revenu et l'actif.

Il y a un poids d'échantillonnage et un poids g pour chacune des phases. Nous verrons plus bas que ces poids sont indispensables pour l'estimation ponctuelle comme pour l'estimation de la variance.

L'échantillon de première phase, désigné par $s_1$, est un échantillon aléatoire stratifié de déclarants tiré d'une base de sondage qui a été créée à partir de données de Revenu Canada. Les strates de première phase sont définies selon la province (PROV), la branche d'activité (CTI2 ou CTI3) et la taille de l'entreprise (TAILLE). Pour effectuer l'échantillonnage de Bernoulli, on attribue à chaque déclarant un nombre aléatoire compris dans l'intervalle (0,1). Ce nombre ne change pas d'une année à l'autre. Les probabilités de sélection de la première phase, désignées par $\pi_{1k}$, peuvent être mises à jour d'année en année pour tenir compte des créations d'entreprises et des changements dans la composition des strates. L'échantillon de première phase est longitudinal, c'est-à-dire qu'il dure d'une année à l'autre. L'échantillonnage de Bernoulli facilite la formation d'un tel échantillon. On peut ajouter des déclarants dans l'échantillon de première phase à chaque année dans le but d'accroître la précision et de remplacer des unités déclarantes qui ne font plus partie du champ de l'enquête.

Désignons par $U_p$, $p = 1, ..., P$, un ensemble de strates formées a posteriori dans la première phase. Ces strates sont formées par le morcellement des strates d'échantillonnage de la première phase. Soit $N_p$ le nombre de déclarants connu dans la strate a posteriori $U_p$. Le poids du déclarant k pour la première phase est

$$w_{1k} = a_{1k}\, g_{1k} = (1/\pi_{1k})\,(N_p/\hat{N}_p)$$

pour $k \in s_{1p} = s_1 \cap U_p$, ou $a_{1k} = 1/\pi_{1k}$ est le poids d'échantillonnage de l'unité k pour la première phase et $g_{1k} = N_p/\hat{N}_p$, avec $\hat{N}_p = \sum_{s_{1p}} 1/\pi_{1k}$ est le poids g de première phase pour chaque $k \in s_{1p}$. La taille effective de l'échantillon dans une strate a posteriori est aléatoire à cause de l'échantillonnage de Bernoulli, et les poids g jouent un rôle important en ce qui concerne la stabilisation de la variance.

Nous cherchons à obtenir des estimations pour la population des entreprises et non pour celle des déclarants. Certaines entreprises sont des sociétés en nom collectif, et les estimateurs exigent un ajustement par lequel des données sur les sociétés en nom collectif sont couplées à des données sur les déclarants. Mais nous n'avons pas, ici, à nous occuper de ce détail technique puisque nous cherchons surtout à illustrer l'estimation pour domaine appuyée

d'information supplémentaire. Dans les formules ci-dessous, chaque déclarant est considéré comme une entité commerciale.

L'échantillon de seconde phase, désigné par $s_2$, est un sous-échantillon aléatoire stratifié, $s_2^r \subset s_1$. Les strates de seconde phase sont définies selon la province, la branche d'activité (CTI4) et la taille de l'entreprise. Les codes CTI4 sont attribués par Statistique Canada à l'échantillon de première phase. Posons $\pi_{2k}$ comme la probabilité de sélection de la seconde phase pour l'unité k.

Désignons par $U_q$, $q = 1, ..., Q$, un ensemble de strates formées a posteriori dans la seconde phase. Celles-ci sont définies selon la branche d'activité (CTI4), la province et la taille de l'entreprise. Désignons par $N_q$ le nombre (inconnu) de déclarants dans la strate a posteriori de seconde phase $U_q$. On peut estimer ce nombre de deux manières. Si nous utilisons l'échantillon de première phase, nous obtenons l'estimation $\hat{N}_q = \sum_{s_{1q}} w_{1k}$ où $s_{1q} = s_1 \cap U_q$ et $w_{1k}$ est le poids de première phase. Si nous n'utilisons que les unités de l'échantillon de seconde phase, nous obtenons une autre estimation, $\tilde{N}_q = \sum_{s_{2q}} w_{1k} a_{2k}$. Dans ce dernier cas, $s_{2q} = s_2 \cap U_q$ et $a_{2k} = 1/\pi_{2k}$ est le poids d'échantillonnage de l'unité k pour la seconde phase. Le poids du déclarant k pour la seconde phase est donc $w_{2k} = a_{2k} g_{2k} = (1/\pi_{2k})(\hat{N}_q / \tilde{N}_q)$ pour $k \in s_{2q} = s_2 \cap U_q$, où $g_{2k} = \hat{N}_q / \tilde{N}_q$ est le poids g de seconde phase pour chaque $k \in s_{2q}$. Notons que les poids sont étalonnés à la première phase, de sorte que $\sum_{s_{1p}} w_{1k} = N_p$ ,pour $p = 1, ..., P$. De plus, ils sont "étalonnés conditionnellement" à la seconde phase, étant donné $s_{1q}$, de sorte que $\sum_{s_{2q}} w_{1k} w_{2k} = \sum_{s_{1q}} w_{1k} = \hat{N}_q$ , pour $q = 1, ..., Q$. Le poids total de l'unité k, désigné par $w_k$, est $w_k = w_{1k} w_{2k}$ et le total y pour le domaine $U_{(d)}$, c.-à-d. $Y_{(d)}$, est estimé au moyen de la formule

$$\hat{Y}_{(d)} = \sum_{k \in s_2} w_k y_{(d)k} = \sum_p \sum_q (N_p / \hat{N}_p)(\hat{N}_q / \tilde{N}_q) \sum_{k \in s_{2pq}} a_{1k} a_{2k} y_{(d)k}$$

où $s_{2pq} = s_2 \cap U_p \cap U_q$, $y_{(d)k} = y_k$ si $k \in U_{(d)}$ et 0 dans le cas contraire. L'estimateur de la variance correspondant est défini par l'expression

$$\hat{V}_{(d)} = \sum_p \sum_q (N_p / \hat{N}_p)^2 (\hat{N}_q / \tilde{N}_q)^2 \sum_{k \in s_{2pq}} a_{1k} (a_{1k} - 1) a_{2k} (e_{(d)pk})^2$$

$$- \sum_p \sum_q (N_p / \hat{N}_p)^2 (\hat{N}_q / \tilde{N}_q)^2 \sum_{k \in s_{2pq}} (a_{1k})^2 a_{2k} (a_{2k} - 1) (e_{(d)qk})^2\}$$

Les résidus contenus dans cette expression sont une extension des résidus définis en (4.5) pour le cas de l'échantillonnage à deux phases. Nous avons

$$e_{(d)pk} = y_{(d)k} - \bar{Y}_{(d)p} \text{ pour } k \in s$$

où

$$\tilde{y}_{(d)p} = (\sum_{s_{2p}} w_k y_{(d)k} / (\sum_{s_{2p}} w_k)$$

et

$$e_{(d)pk} = y_{(d)k} - \tilde{y}_{(d)p} \qquad \text{pour k} \in s_{2q} = s_2 \cap U_q$$

où

$$\tilde{y}_{(d)p} = (\sum_{s_{2p}} w_k y_{(d)k} / (\sum_{s_{2p}} w_k)$$

L'estimateur de la variance, $\hat{V}_{(d)}$ , peut être considéré comme une extension de (4.3) pour le cas de l'échantillonnage à deux phases, étant donné, en l'occurrence, un échantillonnage de Bernoulli dans chacune des phases.

## 5. Estimation dans le temps

Les données qui servent à constituer les séries économiques, et que recueillent périodiquement les organismes gouvernementaux, sont le plus souvent des données mensuelles, trimestrielles ou annuelles. Deux types de mesures très courants qui résument ces données sont les mesures de niveau et les mesures de variation. La variation peut se définir comme l'écart entre des totaux pour deux périodes différentes ou comme le rapport de totaux de périodes différentes. L'estimation de la variance pour les mesures de niveau a été traitée dans les sections précédentes. En ce qui concerne l'estimation de la variance pour les mesures de variation, il faut calculer des covariances pour deux périodes qui nous intéressent. Ces covariances doivent refléter la nature changeante de l'univers (créations et disparitions d'entreprises) comme de l'échantillon (créations, disparitions et renouvellement). Tam (1984) a défini des formules de covariance suivant des plans d'échantillonnage répété -- échantillonnage aléatoire simple -- en conservant la même population finie. Hidiroglou et Laniel (1986) ont étendu les résultats de Tam à des échantillons avec renouvellement (selon un plan d'échantillonnage aléatoire simple en grappes stratifié) dans une population changeante. Pour une période t donnée et un domaine d, posons l'estimateur du total pour domaine comme

$$\hat{Y}_d(t) = \sum_{h=1}^{H} \sum_{k=1}^{n_h(t)} \frac{N_h(t)}{n_h(t)} y_{(d)hk}(t) \tag{5.1}$$

où $N_h(t)$ et $n_h(t)$ sont, respectivement, l'effectif de la population et de l'échantillon au temps t; $y_{(d)hk}(t)$ est définie comme en (4.1). La variance de $Y_d(t)$ est estimée selon la formule énoncée plus tôt. La covariance estimée des estimations $\hat{Y}_{(d)}(t)$ et $\hat{Y}_{(d)}(s)$, $s < t$ , est définie

$$Cov\{\hat{Y}_d(t), \hat{Y}_{(d)}(s)\} = \sum_{h=1}^{L} \left(1 - \frac{n_h(t)n_h(s)}{N_h(t)N_h(s)} \frac{N_h(t,s)}{n_h(t,s)}\right)$$
$$\frac{n_h(t,s)}{n_h(t,s)-1} \sum_{k=1}^{n_h(t,s)} \{z_{(d)hk}(t) - \bar{z}_{(d)}(t)\}\{z_{(d)k}(s) - \bar{z}_{(d)}(s)\} \tag{5.2}$$

où $N_h$ *(t,s)* et $n_h$ *(t,s)* représentent, respectivement, le nombre d'unités de la population et le nombre d'unités échantillonnées présentes dans l'échantillon aux deux périodes, t et s. Notons que

$$z_{(d)hk}(t) = \frac{N_h(t)}{n_h(t)} y_{(d)hk}(t)$$ et $\bar{z}_{(d)}(t)$ est la moyenne d'échantillon de $z_{(d)hk}$ *(t)* basée sur $n_h(t)$

observations. Les résultats ci-dessus sont le prolongement d'un des scénarios d'échantillonnage élaborés par Tam (1984). Laniel (1988) pousse encore plus loin un autre des scénarios d'échantillonnage de Tam (1984).

Montrons maintenant comment ces covariances entrent dans l'estimation de la variance pour les mesures de variation. La différence de totaux estimés pour deux périodes t et s est

$$\hat{D}_d(t,s) = \hat{Y}_d(t) - \hat{Y}_d(s) \tag{5.3}$$

et la variance estimée de cette différence est

$$v\left(\hat{D}(t,s)\right) = v\left(\hat{Y}_d(t)\right) - 2cov\left(\hat{Y}_d(t), \hat{Y}_d(s)\right) - v\left(\hat{Y}_d(s)\right) \tag{5.4}$$

Pour le rapport , la variance estimée est

$$v\left(\hat{R}_d(t,s)\right) = v\left(\hat{D}_d(t,s)\right) / \hat{Y}_d(s)^2 \tag{5.5}$$

Le calcul de ces covariances est aussi nécessaire pour des méthodes qui, comme l'estimation composite par exemple, exploitent le caractère temporel des estimations tirées d'enquêtes répétées. On a aussi besoin des covariances pour le calcul de facteurs de pondération optimaux qui réunissent des estimations de totaux pour plusieurs périodes ainsi que les variances estimées correspondantes. Le raisonnement sur lequel s'appuient ces méthodes est que s'il existe une bonne corrélation entre les données tirées de passages répétés d'une enquête pour de mêmes unités, les estimateurs qui réunissent ces données seront plus fiables que ceux qui ne les réunissent pas.

Une série économique donnée peut être construite avec des valeurs infra-annuelles ou annuelles à l'aide de méthodes de collecte de données différentes. Ces deux sources de données différeront très vraisemblablement si les données infra-annuelles sont agrégées dans le but de produire des données annuelles. On appelle habituellement "étalonnage" l'opération qui consiste à corriger des données infra-annuelles tirées d'une source donnée dans le but de les faire concorder avec des données annuelles tirées d'une autre source. La source de données annuelles est considérée comme sûre. À titre d'exemple, Statistique Canada publie des estimations mensuelles des ventes au détail pour un certain nombre d'industries du secteur. En outre, l'organisme effectue une enquête annuelle indépendante qui permet de connaître le total des ventes au détail annuelles.

Si la source de données annuelles est considérée comme sûre, l'étalonnage permet de corriger la série infra-annuelle de manière que la somme des éléments de la série corrigée (pour une période donnée) corresponde à une valeur donnée de la série annuelle (repère) pour la même période. Cette question a été traitée par Denton (1971), Helfand.

Monsour et Trager (1977), Monsour et Trager (1979), Fernandez (1981) et Cholette (1984). Supposant que les données infra-annuelles sont des données mensuelles, ces auteurs révisent ces chiffres en recourant à une minimisation avec contrainte de forme quadratique des différences entre la série révisée et la série non révisée. Ils visent ainsi à limiter au maximum les variations d'un mois à l'autre et à réduire le plus possible la distorsion que peut introduire dans les données le mouvement saisonnier.

L'approche de ces auteurs fait abstraction de ce que i) les deux séries (infra-annuelle et annuelle) peuvent être entachées d'erreurs (conséquence de la variabilité d'échantillonnage et de la variabilité non due à l'échantillonnage) et que ii) la série infra-annuelle peut être biaisée. Reconnaissant que les séries peuvent comporter des erreurs, Hillmer et Trabelsi (1987) ont utilisé des méthodes d'analyse chronologique pour produire une solution. Ils ont montré aussi qu'une fois étalonnées, les estimations avaient une erreur quadratique moyenne moins élevée. En élargissant la méthode de Hillmer-Trabelsi de manière à tenir compte de la possibilité de séries infra-annuelles biaisées, Laniel et Fyfe (1989) ont obtenu des séries étalonnées en appliquant la théorie des moindres carrés à un système d'équations qui contient des modèles reflétant le caractère stochastique des séries infra-annuelles. Ces auteurs s'étaient servis à cette fin de la procédure de Gauss-Newton, en tenant compte des contraintes non actives.

## 6. Conclusions

Cette étude nous a permis d'exposer un certain nombre de méthodes d'estimation et de pondération qui peuvent être appliquées dans les enquêtes-établissements. Nous avons montré comment l'utilisation de données auxiliaires pouvait s'inscrire dans un modèle d'estimation. À l'aide de ce modèle général, nous avons exposé plusieurs estimateurs parmi les plus courants. Nous sommes aussi servis de ce modèle pour traiter l'importante question de l'estimation par domaine et nous sommes intéressés à l'estimation de la covariance de totaux de population estimés pour deux périodes distinctes lorsque la composition de la population et celle de l'échantillon peuvent avoir changé. Le calcul de cette covariance est utile pour des estimations obtenues par étalonnage ou par des méthodes d'estimation composite.

## BIBLIOGRAPHIE


Armstrong, J., Block, C., and Srinath, K.P., (1993), "Two-phase sampling of tax records for Business Surveys," to appear in the *Journal of Business and Economic Statistics*.

Cholette, P.A. (1984), "Adjusting sub-annual series to yearly benchmarks," *Survey Methodology Journal*, **10**, pp. 35-49.

Choudhry, G.H., Lavallée, P., and Hidiroglou, M.A. (1989), " Two-phase sample design for tax data," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pp. 646-641.

Deming, W.E. and Stephan, F.F. (1940) , "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, pp. 427-444.

Denton, F.T. (1971), "Adjustment on monthly or quarterly series to annual totals: an approach based on quadratic minimization," *Journal of the American Statistical Association*, **46**, pp. 99-102.

Deville, J.-C., and Särndal, C.E. (1992), "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, **87**, pp. 376-382.

Deville, J.-C., Särndal, C.E., and Sautory, O. (1993), " Generalized raking procedures in survey sampling," To appear in *Journal of the American Statistical Association*.

Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1992), "Requirements on a generalized estimation system at Statistics Canada," paper presented at the Workshop on Uses of Auxiliary Information, Statistics Sweden, Orebro.

Fernandez, R.B. (1981), "A methodological note on the estimation of time series," *Review of Economics and Statistics*, **63**, pp. 471-476.

Gossen, M. and Latouche, M. (1992), "Post-stratification to reduce sample bias in an establishment survey," paper presented at the American Statistical Association Meetings, Business Surveys Section, in Boston.

Göttgens, R., Vellen, B., Odekerken, M., and Hofman, L. (1991), "Bascula, version 1.0. A Weighting Package under MS-DOS, User Manual," CBS-Report, Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Helfand, S.D., Monsour, N.J., and Trager, M.L. (1977), "Historical revision of current business survey estimates," *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 246-250.

Hidiroglou, M.A., and Laniel N. (1986), "Specifications for the estimation system of the Wholesale and Retail Trade Survey," Statistics Canada internal document.

Hidiroglou, M.A., Choudhry, G.H., and Lavallée, P. (1991), "A sampling and estimation methodology for sub-annual business surveys," *Survey Methodology*, **17**, pp. 195-210.

Hillmer, S.C. and Trabelsi, A. (1987), "Benchmarking of economic time series," *Journal of the American Statistical Association*, **82**, pp. 1064-1071.

Huang, E. and Fuller, W.A. (1978), "Nonnegative regression estimation for sample survey data," *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 330-305.

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

Laniel. N. (1988), "Variances for a rotating sample from a changing population," *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 246-250.

Laniel, N., and Fyfe, K. (1989), "Benchmarking of economic time series," *Analysis of Data In Time, Proceedings of the 1989 International Symposium, held at Statistics Canada*, pp 125-130.

Lee, H., and Croal, J. (1989), "A simulation study of various estimators which use auxiliary data in an establishment survey," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 336-341.

Monsour, N.J., and Trager, M.L. (1979), "Revision and benchmarking of business time series," *Proceedings of the Business and Economic Statistics Section, American Statistical association*, pp. 333-337.

Rao, J.N.K., (1985), " Conditional Inference in Survey Sampling," *Survey Methodology*, **11**, pp. 15-31.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1989), " The weighted residual technique for estimating the variance of the general regression estimator of the finite population total," *Biometrika*, **76**, pp. 527-537.

Särndal, C.E. and Hidiroglou, M.A. (1989), "Small domain estimation: a conditional analysis," *Journal of the American Statistical Association*, **84**, pp. 266-275.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1992), "*Model Assisted Survey Sampling*: New-york, Springer-Verlag.

Schiopu-Kratina, and Srinath, K.P. (1991), "Sample rotation and estimation in the Survey of Employment, Payrolls and Hours," *Survey Methodology*, **17**, pp. 79-90.

Sunter, A.B. (1977), " Response burden, Sample rotation and classification renewal in economic surveys," *International Statistical Review*, **45**, pp. 209-222.

Tam, S.M. (1984), "On covariances from overlapping samples," *The American Statistician*, **38**, pp. 288-292.

Valliant, R. (1993), "Poststratification and conditional variance estimation," *Journal of the American Statistical Association*, **88**, pp. 89-96.