Techniques d'enquête, décembre 1986
Vol. 12, n° 2, pp. 205–214
Statistique Canada

Comparaison de la pondération et de l'imputation pour des données non-échantillonnées

SYLVIE MICHAUD[1]

RÉSUMÉ

Le Recensement de la Construction (RC) au Canada se sert d'un plan de sondage complexe pour échantillonner les petites entreprises (les entreprises dont le revenu brut est inférieur à $750,000). Des échantillons stratifiés sont sélectionnés à partir de base de sondage qui se chevauchent. À partir d'un des échantillons, deux sous-échantillons sont sélectionnés de façon indépendante. De l'information plus détaillée est recueillie, pour les entreprises choisies dans les sous-échantillons. Deux stratégies pourraient être envisagées, pour estimer des totaux pour les variables recueillies dans les sous-échantillons. La première approche serait de déterminer des poids, basés sur les fractions de sondage. Cette approche nécessite l'utilisation de plusieurs poids différents. Une seconde approche serait d'imputer des valeurs aux entreprises sélectionnées dans l'échantillon mais pas dans les sous-échantillons. Cette approche crée un fichier "rectangulaire" complet au niveau de l'échantillon. Un seul poids peut ensuite être utilisé pour obtenir des estimés pour la population. Cette approche "d'imputation massive" est présentement utilisée par le Recensement de la Construction. L'étude vise à comparer les estimés qui pourraient être obtenus, en utilisant diverses stratégies d'estimation aux estimés obtenus lorsque l'approche d'imputation massive est employée.

MOTS CLÉS: Pondération; imputation massive; non-échantillonné.

1. INTRODUCTION

Le recensement de la construction (RC) est une enquête annuelle qui vise à estimer les dépenses dans le domaine de la construction. Bien que l'enquête porte le titre de "recensement", seules les entreprises ayant un revenu brut supérieurà $750,000 sont recensées. On envoie un questionnaire long à ces entreprises, dans lequel on recueille diverses informations financières et non financières. On estime les dépenses des entreprises dont le revenu brut se situe entre $10,000 et $750,000 à partir d'un échantillon de données administratives. Dans un premier temps, deux échantillons sont sélectionnés de facon indépendante, à partir de bases de sondage qui se chevauchent. Deux sous-échantillons sont ensuite choisis parmi un des échantillons, pour obtenir de l'information supplémentaire.

Deux stratégies pourraient être employées pour estimer des variables recueillies dans les sous-échantillons. Présentement, l'approche utilisée par le recensement de la construction est d'imputer des valeurs pour les entreprises non sélectionnées dans un sous-échantillon, mais choisies dans un échantillon. Cela crée un fichier "rectangulaire" complet, qui peut être pondéré au niveau de la population en n'utilisant qu'un seul poids. Une alternative serait de calculer des poids, basés sur les probabilités de sélection. Différents poids devraient être calculés pour différents sous-ensembles de données. Le but de cette étude est de comparer les estimés obtenus par pondération aux estimés obtenus par imputation.

L'étude est effectuée sur une population d'entreprises non incorporées seulement parce que pour l'année fiscale 1983, les stratégies de sélection des échantillons d'entreprises non incorporées et incorporées étaient différentes. Cependant, la stratégie utilisée pour les entreprises

[1] S. Michaud, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11ième étage, R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

incorporées sera modifiée pour l'année fiscale 1984 et deviendra équivalente à celle des entreprises non incorporées. C'est donc la stratégie des entreprises non incorporées qui a été étudiée. Il est espéré que les conclusions resteront semblables, pour les entreprises incorporées.

2. DESCRIPTION DU PLAN DE SONDAGE

Comme mentionné précédemment deux échantillons sont sélectionnés de facon indépendante, à partir de deux bases de sondage qui se chevauchent. Un premier échantillon, l'échantillon pré-spécifié, est un échantillon stratifié (par revenu brut (RB), province et code d'activité économique CAE 3 chiffres 1970), sélectionné pour le recensement de la construction. Il est sélectionné à partir d'une base de sondage qui n'est pas complètement à jour. En effet, la base de sondage contient des entreprises "mortes", i.e. des entreprises qui ne sont plus dans le champs de l'enquête pour construction (l'entreprise n'existe plus, elle n'est plus dans un domaine d'activité de construction ou l'entreprise a un revenu brut inférieur à $10,000). De la même façon, la base de sondage ne contient pas les naissances et les entreprises qui ont changé de domaine d'activité et qui sont maintenant en construction. De facon indépendante, un échantillon "transversal" est choisi par Revenu Canada. Cet échantillon stratifié (par des classes de Revenu Brut) est choisi parmi tous les domaines d'activité économique (pas seulement construction), à partir d'une base de données complète il sert à couvrir les naissances. La figure 1 illustre la situation.

À partir des unités de l'échantillon pré-spécifié, deux sous-échantillons sont sélectionnés de façon indépendante: un sous-échantillon financier et un sous-échantillon "autres caractéristiques" (A.C.). Le sous-échantillon A.C. est sélectionné directement à partir de l'échantillon pré-spécifié, alors que le sous-échantillon financier est sélectionné à partir des données transcrites de l'échantillon (on ne sous-échantillonne donc pas les "morts"). Plus de détails sur le plan de sondage peuvent être obtenus dans Giles(1983).

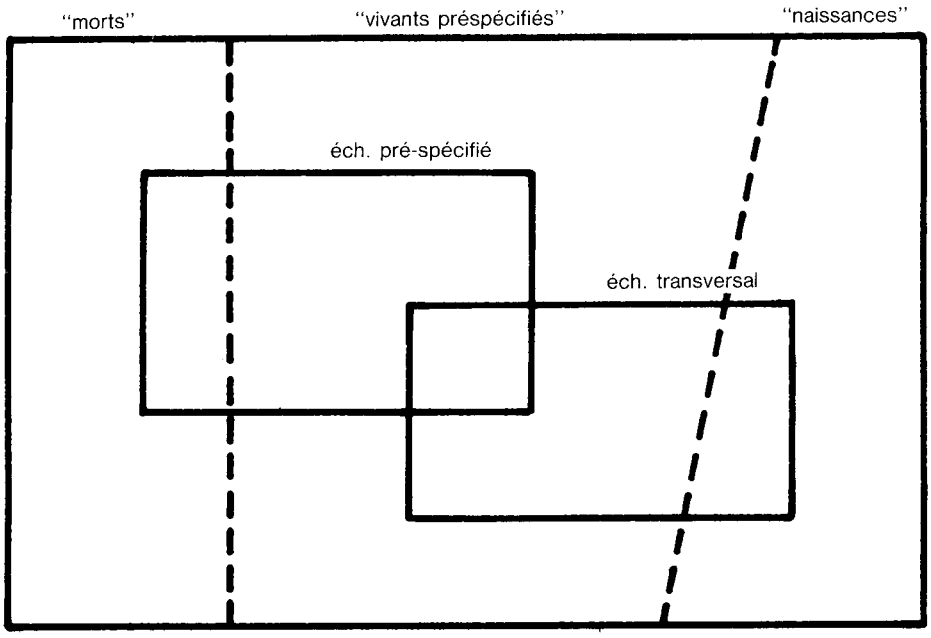

Figure 1. Schéma du plan de sondage du RC

3. STRATÉGIE D'IMPUTATION

Le RC utilise une stratégie d'imputation massive, pour estimer les variables sélectionnées dans un sous-échantillon donné (i.e. que pour tous les enregistrements non sélectionnés dans les sous-échantillons, des valeurs sont imputées pour chaque variable). Le processus d'imputation se fait de façon indépendante pour chaque sous-échantillon. (La procédure d'imputation se fait par phases. Les phases d'imputation des sous-échantillons sont indépendantes entre elles et elles utilisent des stratégies différentes). Dans chaque phase, le plus proche voisin est sélectionné comme donneur pour imputer les variables non- échantillonnées (le plus proche voisin est choisi parmi un sous-groupe de donneurs potentiels).

L'imputation se fait de façon différente pour chaque sous-échantillon.

Pour le sous-échantillon financier, la valeur imputée est la valeur du donneur, ajustée par le quotient d'une variable auxiliaire, disponible à la fois pour le donneur et pour le candidat (le candidat étant défini comme étant l'unité ayant besoin d'être imputée). (Note: La procédure réelle est plus complexe qu'une simple imputation ajustée par le quotient. En effet, les variables sont imputées de facon hiérarchique et des contraintes linéaires sont imposées aux valeurs imputées (la deuxième variable dépend de la valeur imputée à la première variable...). Plus de détails sur la procédure réelle peuvent être trouvés dans Philips et Emery (1976). Une vue générale plus détaillée se trouve aussi dans Colledge *et coll* (1978)).

Soit Y: la variable d'intérêt (qui doit être imputée pour le candidat, connue pour les donneurs),
X: une variable auxiliaire disponible à la fois pour le donneur et le candidat,
c: un indice référant au candidat,
d: un indice référant au donneur,
I: un indice référant à une valeur imputée.

Pour les variables du sous-échantillon financier la valeur imputée Y_c^I est définie comme étant:

$$Y_c^I = Y_d \frac{X_c}{X_d}$$

Pour les variables du sous-échantillon A.C., la valeur imputée est la valeur du donneur.

$$Y_c^I = Y_d$$

Après l'étape d'imputation, il existe un fichier rectangulaire complet (toutes les entreprises ayant été sélectionnées dans un des échantillons possèdent des valeurs pour toutes les variables des échantillons sous-échantillons). En pondérant l'échantillon, des estimés au niveau de la population peuvent être obtenus.

Le poids utilisé peut être décrit comme étant l'inverse de la probabilité de sélection, dans au moins un des échantillons.

Soit: $P(\text{presp}_h)$: la probabilité qu'une unité ait été sélectionnée dans la strate h de l'échantillon pré-spécifié.
$P(\text{trans}_k)$: la probabilité qu'une unité ait été sélectionnée dans la strate k de l'échantillon transversal.
hk : classification croisée des unités.
h : l'indice relié à la stratification de l'échantillon pré-spécifié.
k : l'indice relié à la stratification de l'échantillon transversal.

Le poids associé à chaque unité peut être écrit comme

$$W_{hk}^{-1} = 1 - [1 - P(\text{presp}_h)]\,[1 - P(\text{trans}_k)]$$

Les naissances et les morts ne peuvent être classifiés de façon croisée. Les morts ont un poids $W_h = 0$ et les naissances, un poids W_k inversement égal à la probabilité de sélection dans la strate k de l'échantillon transversal. Plus de details peuvent être obtenus dans Bankier (1982).

L'estimateur du total obtenu, lorsque la stratégie d'imputation est utilisée, est donc

$$\hat{Y} = \sum_{h,k} W_{hk} \sum_{j=1}^{n_{hk}} y_{jhk}^{*}$$

où $y_{jhk}^{*} = y_{jhk}$ si $j \in$ sous-échantillon

$= y_{jhk}^{I}$ si $j \notin$ sous-échantillon.

4. STRATÉGIE DE PONDÉRATION

Si une stratégie de pondération était utilisée pour estimer les variables des sous-échantillons, différents estimateurs pourraient être envisagés. Les estimateurs ont la même forme pour les deux sous-échantillons. Cependant les poids utilisés sont différents.

Un premier estimateur ($\hat{Y}_1$) serait un estimateur basé sur le plan d'échantillonnage utilisé, ajusté pour la sous-couverture de la population. Dans chaque strate de CAE, PROV et RB (code d'activité économique, province, revenu brut), un échantillon pré-spécifié est sélectionné. Lorsque transcrites (unités échantillonnées et encore vivantes), les unités sont classées en deux strates : "hors du champs de l'enquête" et "dans le champs de l'enquête". Les sous-échantillons sont choisis à partir de la strate d'unités "dans le champs de l'enquête". (On peut supposer que toutes les unités de la strate "hors du champs de l'enquête" ont été sous-échantillonnées et qu'elles ont une moyenne égale à zéro). L'estimateur contient un facteur de correction, pour compenser pour la sous-couverture de la base de sondage (obtenu à partir de l'information de l'échantillon transversal).

Un second estimateur possible ($\hat{Y}_2$) serait une version simplifiée du premier estimateur proposé. Au lieu de supposer un double échantillonnage pour déterminer les unités "dans le champs de l'enquête" et "hors du champs de l'enquête", on pourrait supposer qu'un échantillon stratifié pré-spécifié est choisi à partir des unités "dans le champs de l'enquête". Un sous-échantillon est sélectionné à partir de l'échantillon pré-spécifié. L'estimateur doit encore une fois, être ajusté pour tenir compte de la sous-couverture. S'il ne s'avère pas y avoir de différences significatives entre le premier et le second estimateur, le second serait préférable, parce que plus simple.

Un troisième estimateur possible ($\hat{Y}_3$) serait un estimateur basé sur l'information provenant de l'échantillon transversal seulement. On pourrait supposer que les unités choisies à la fois dans le sous-échantillon et dans l'échantillon transversal ont été choisies à partir de l'échantillon transversal. Le rationnel pour un tel estimateur serait que l'échantillon transversal est sélectionné à partir d'une base de sondage complète. Cependant, comme les sous-échantillons sont sélectionnés à partir de l'échantillon pré-spécifié et non à partir de l'échantillon transversal, la taille des sous-échantillons dans l'échantillon transversal sera petite.

Finalement un dernier estimateur possible ($\hat{Y}_4$) serait de supposer que le sous-échantillon a été sélectionné à partir de l'échantillon complet (échantillon pré-spécifié + échantillon transversal), et que l'échantillon complet provient de bases multiples. Ce quatrième estimateur est celui qui est le plus "semblable" à l'estimateur obtenu après l'imputation massive. En effet, ces deux estimateurs supposent que les naissances et les nouvelles entreprises "réagissent" comme le reste de la population. Aucun ajustement particulier n'est fait pour ces entreprises, dans la technique d'imputation, et l'estimateur pondéré n'est pas stratifié de façon à distinguer ces unités. De plus les deux estimateurs tiennent compte du fait que l'échantillon provient de bases multiples. Le même poids est donc utilisé dans les deux cas, pour pondérer l'échantillon à la population.

Comme mentionné précédemment, les variables recueillies dans le sous-échantillon financier sont ajustées par le quotient d'une variable auxiliaire lors de l'imputation. Pour cette raison, un autre type d'estimateur pourrait être envisagé pour les variables recueillies dans le sous-échantillon financier: soit un estimateur par le quotient. La variable auxiliaire utilisée serait la variable utilisée lors de l'imputation. Comme pour la pondération simple, différents estimateurs pourraient être calculés.

Les différents estimateurs et leur variance sont écrits de façon mathématiques dans l'annexe 1.

5. RÉSULTATS

Pour l'étude certaines variables ont été choisies parmi chacun des sous-échantillons.

Le sous-échantillon financier est composé de sept variables. L'étude s'est limitée à quatre de ces sept variables.

Pour le sous-échantillon autres caractéristiques, huit variables sont recueillies pour toutes les entreprises. Les autres variables sont disponibles pour différents groupes d'activité économique seulement. L'étude se limitera donc à ces huit variables.

Pour le sous-échantillon financier, les variables présentées dans ce rapport sont ADD (additions immobilisations) et RM (réparation et entretien). Pour le sous-échantillon A.C., la variable PCON (pourcentage de construction dans un domaine particulier) est présentée. Cependant, la variable PCON n'est pas publiée directement. Elle est multipliée par les dépenses totales, pour obtenir les dépenses dans un domaine particulier: PEXP. C'est cette deuxième variable qui a été étudiée.

Comme mentionné précédemment, les variables dans le sous-échantillon A.C. ne sont pas ajustées par un quotient, lors de l'imputation. Les estimateurs par le ratio ne s'appliqueront donc pas à ces variables.

Les tables 1, 2 et 3 nous présentent les différents estimateurs et les estimateurs de leurs variances respectives (basé sur les données fiscales de 1983, pour les entreprises non incorporées).

Tableau 1

Différentes estimations de PEXP et de l'écart-type de PEXP (%EXP*EXPCONS)

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_1$
estimation ($\times 10^{11}$)	3.44	3.43	3.96	3.66	3.70
écart-type ($\times 10^{9}$)	3.5	3.5	8.4	3.2	

Tableau 2

Différentes estimations de ADD et de l'écart-type de ADD

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_{Q2}$	$\hat{Y}_{Q3}$	$\hat{Y}_{Q4}$	$\hat{Y}_1$
estimation ($\times 10^{8}$)	2.08	2.10	2.14	1.84	7.82	5.06	5.2	1.4
écart-type ($\times 10^{7}$)	1.9	1.9	2.0	1.0	0.8	2.2	0.8	

Tableau 3

Différentes estimations de RM et de l'écart-type de RM

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_{Q2}$	$\hat{Y}_{Q3}$	$\hat{Y}_{Q4}$	$\hat{Y}_1$
estimation ($\times 10^8$)	1.5	1.5	1.43	1.55	0.9	1.63	1.67	1.75
écart-type ($\times 10^6$)	6.9	6.9	8.9	5.3	3.1	11.0	4.3	

On peut remarquer dans un premier temps, qu'il n'y a pas de différences significatives entre les deux premiers estimateurs. (Selon les définitions préalablement établies le second estimateur est une version simplifiée du premier estimateur). La version simplifiée sera donc conservée.

Pour les variables dans le sous-échantillon non financier, en général, l'imputation semble donner des résultats semblables à ceux obtenus par la pondération ($\hat{Y}_4$). L'estimateur obtenu en ne considérant que les unités provenant de l'échantillon transversal ($\hat{Y}_3$) semble plus variable que les autres estimateurs. Cette variabilité pourrait être expliquée par le plus petit nombre d'unités utilisées dans le calcul de cet estimateur. Il est à noter que ces observations ne sont basées que sur un échantillon observé et que les conclusions sont donc limitées. Cependant, à cause de la nature des données (souvent des pourcentages et des subdivisons d'activité dans le domaine de la construction), qui est relativement stable dans les strates (CAÉ 1970 3 chiffres, province et RB) il n'a pas eté jugé nécessaire de pousser plus à fond l'étude de ces variables.

Pour les variables dans le sous-échantillon financier il a été remarqué qu'en général, les estimateurs ajustés par le quotient ne semblent pas toujours applicables (ex: variable ADD). Ils donnent des estimés extrêmement biaisés. Une explication possible est que la variable ADD et la variable auxiliaire utilisée ont une grande fréquence de zéro. Un "mauvais" échantillon dans certaines strates peut donc faire gonfler excessivement les estimés.

Certains problèmes ont aussi été observés avec le système d'imputation, (données imputées alors qu'elles n'auraient pas dû, données non imputées), ce qui a parfois pu influencer les estimés obtenus par la stratégie d'imputation. Parce que les résultats étaient basés sur un échantillon observé seulement et parce qu'il était difficile d'estimer l'impact des problèmes reliés au système, sur les estimés, il a été décidé de faire une simulation.

6. SIMULATION

La simulation a été effectuée sur un sous-ensemble de données; soit les entreprises ayant été sélectionnées dans le sous-échantillon financier (pour ce sous-ensemble de données, toutes les variables étudiées sont présentes). Par la suite, il a été essayé de reprendre de façon simplifiée la stratégie utilisée par le recensement de la construction. Un échantillon stratifié a été sélectionné. Des fractions de sondage semblables à celles de l'enquête ont été utilisées dans l'échantillonnage. Les variables du sous-échantillon financier, pour les données non-sélectionnées dans l'échantillon, ont été mises manquantes, pour être ensuite imputées par le système. Le processus de sélection d'un échantillon, puis d'imputation a été répété trente fois.

Des estimés ont été produits, comparant les résultats obtenus en sommant les données non-imputées et imputées, par rapport aux estimés obtenus en pondérant l'échantillon par l'inverse de la fraction de sondage. Comme la valeur de la population est connue, le biais a été calculé, en plus de la variance des estimés. Les résultats pour les variables ADD et RM sont présentés dans les tables 4 et 5.

Tableau 4

Estimés de ADD obtenus par simulation

	Population	Pondération	Ratio	Imputation
estimé ($\times 10^7$)	1.41	1.43	1.24	1.41
écart-type ($\times 10^5$)		1.11	.85	1.15
biais ($\times 10^5$)		.22	−1.73	−0.07

Tableau 5

Estimés de RM obtenus par simulation

	Population	Pondération	Ratio	Imputation
estimé ($\times 10^7$)	1.06	1.06	1.07	1.04
écart-type ($\times 10^5$)		4.52	4.11	4.87
biais ($\times 10^5$)		−0.07	−0.95	−1.38

Pour la variable ADD, l'estimé obtenu par le ratio est significativement différent des estimés obtenus par imputation ou par pondération. Le biais de L'estimé est aussi significativement non nul. Pour la variable RM, tous les estimateurs sont équivalents (variance égales, biais non significatif à un seuil de 5%, estimés non significativement différents).

7. CONCLUSIONS

D'après l'étude, il ne semble pas y avoir de différences significatives entre une stratégie d'imputation massive et de pondération, pour les variables du sous-échantillon autres caractéristiques. Le résultat est prévisible dans la mesure où les variables étudiées semblent relativement stables à l'intérieur des strates.

Les conclusions pour les variables du sous-échantillon financier se basent sur les résultats obtenus par la simulation. De cette étude, il semble que la pondération par l'inverse de la probabilité de sélection et l'imputation massive donnent des estimés comparables. L'estimateur par le ratio ne semble pas approprié pour la variable ADD (ainsi que pour d'autres variables étudiées, non présentées ici). Une autre étude cherchera à étudier si l'estimateur par la régression serait plus approprié. L'étude visera aussi à évaluer l'impact de l'imputation sur la structure de corrélation des variables.

ANNEXE

Les estimateurs proposés pourraient être écrits de la façon suivante:

Notations

Soit h : les strates de l'échantillon pré-spécifié,

k : les strates de l'échantillon transversal,

N_h : la taille de la population "pré-spécifiée" dans la strate h,

$\hat{N}_{1h}$: la taille de la population "pré-spécifiée" avec des entreprises "vivantes" (dans le champs de l'enquête) dans la strate h (estimé),

$\hat{N}_{2h}$: la taille de la population "pré-spécifiée" avec des entreprises "hors du champs de l'enquête", dans la strate h (estimé),

$\hat{N}_k$: la taille de la population dans la strate k, estimée à partir de l'information de l'échantillon transversal,

$\hat{N}_k'$: la taille de la population dans la strate k, estimée à partir de l'information des deux échantillons (bases multiples),

n_h : le nombre de données échantillonnées dans la strate h de l'échantillon pré-spécifié,

$\hat{n}_{1h}$: le nombre de données échantillonnées et transcrites de la strate h de l'échantillon pré-spécifiée,

$\hat{n}_k'$: le nombre de données échantillonnées et transcrites dans la strate k,

$\hat{m}_{1h}$: le nombre de données sous-échantillonnées parmi les "vivants" de la strate h,

y : une variable d'un des sous-échantillons,

x : une variable auxiliaire disponible pour toutes les données des échantillons,

s_{yh}^2 : une estimé de la variance de y pour les données du sous-échantillon dans la strate h,

s_{xh}^2 : une estimé de la variance de x pour les données du sous-échantillon dans la strate h,

s_{yxh} : une estimé de la covariance entre x et y dans la strate h.

i)
$$\hat{Y}_1 = \left(\frac{\hat{N}_{1\text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec}}}\right) \sum_h \frac{N_h}{n_h} \frac{\hat{n}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} Y_{hj}$$

$$V(\hat{Y}_1) \simeq \left(\frac{\hat{N}_{1\text{ pre-spec}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec}}}\right)^2 \sum_h N_h n_h \left(\frac{N_h - 1}{n_h - 1}\right)$$

$$\times \left[W_{1h} S_h^2 \left(\frac{1}{\gamma_h} - \frac{1}{N_h}\right) + \frac{G_h}{n_h} S_h^2 \left(\frac{W_{1h}}{N_h} - \frac{1}{\gamma_h}\right) + \frac{G_h}{n_h} W_{1h}(1 - W_{1h})^2 \bar{y}_h^2 \right]$$

où
$$G_h = \left(\frac{N_h - n_h}{N_h - 1}\right), \ \gamma_h = n_h \frac{\hat{m}_{1h}}{\hat{n}_{1h}}, \text{ et } W_{1h} = \frac{\hat{n}_{1h}}{n_h}.$$

ii)
$$\hat{Y}_2 = \left(\frac{\hat{N}_{1\text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec.}}}\right) \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} \sum_{j=1}^{\hat{m}_{1h}} y_{hj}$$

$$V(\hat{Y}_2) \simeq \left(\frac{\hat{N}_{1\text{ pre-spec.}} + \hat{N}_{\text{births}}}{\hat{N}_{1\text{ pre-spec.}}}\right)^2 \sum_h \frac{\hat{N}_{1h}}{\hat{m}_{1h}} (\hat{N}_{1h} - \hat{m}_{1h}) s_{yh}^2$$

iii)
$$\hat{Y}_3 = \sum_k \frac{\hat{N}_k}{\hat{m}_{1k}} \sum_j Y_{kj}$$

$$V(\hat{Y}_3) = \sum_k \hat{N}_k \left(\frac{\hat{N}_k - \hat{m}_{1k}}{\hat{m}_{1k}}\right) S_{yk}^2$$

iv) $$\hat{Y}_4 = \sum_k \frac{\hat{N}'_k}{\hat{m}_{1k}} \sum_{j=1}^{\hat{m}_{1k}} y_{kj}$$

$$V(\hat{Y}_4) = \sum_k \hat{N}'_k \left(\frac{\hat{N}'_k - \hat{m}_{1k}}{\hat{m}_{1k}}\right) s^2_{yk}.$$

Des estimateurs par le quotient peuvent être calculés. Comme pour les estimateurs simples, ils peuvent prendre différentes forms, dépendant des hypothèses formulées. Par exemple, l'estimateur par le quotient équivalent à l'estimateur 4 serait:

$$\hat{Y}_{Q4} = \sum_k \hat{N}'_k \, \bar{Y}_{\text{sub}_k} \frac{\bar{X}_{\text{samp}_k}}{\bar{X}_{\text{sub}_k}}$$

où $\bar{X}_{\text{samp}_k}$ est la moyenne de la variable X pour les unités sélectionnées dans l'échantillon complet, qui sont dans la strate k

$\bar{X}_{\text{sub}_k}$ est la moyenne de X pour les unités sélectionnées dans le sous-échantillon, qui sont dans la strate k

$\bar{Y}_{\text{sub}_k}$ est la moyenne de la variable Y dans la strate k du sous-échantillon.

$$V(\hat{Y}_{Q4}) = \sum_k (\hat{N}'_k)^2 \left(\frac{1}{\hat{m}_{1k}} - \frac{1}{\hat{n}'_{1k}}\right) \left[s^2_{y_k} + \hat{R}^2_k s^2_{x_k} - 2\hat{R}_k s_{yx_k} + \left(\frac{1}{\hat{n}'_{1k}} - \frac{1}{\hat{N}'_k}\right) s^2_{y_k}\right]$$

où $\hat{R}_k = \dfrac{\bar{Y}_{\text{sub}_k}}{\bar{X}_{\text{sub}_k}}$.

REMERCIEMENTS

L'auteur tient à remercier P. Giles pour les commentaires et les suggestions fournis lors de cette étude, de même que les arbitres pour leurs remarques pertinentes.

BIBLIOGRAPHIE

BANKIER, M., (1982). Variance formula for an estimator based on any number of independant stratified samples of which some are Poisson samples. Document technique, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.

COLLEDGE, M.L., JOHNSTON, J.H., PARÉ, R., et SANDE, I.G.(1978). Large scale imputation of survey data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 721-726.

GILES, P. (1983). Construction division: Census of construction. Document Technique, Division des méthodes d'enquêtes-entreprises, Statistique Canada.

PHILIPS, J.L., et EMERY, D. (1976), FIBCOC documentation Document Technique, Division des Développements de système, Statistique Canada.

RAO, J.N.K. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.