11-617F
no.91-01
c.2

stics Statistique
ada Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

[redacted]

WORKING PAPER NO. BSMD-91-001F

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-91-001F

DIRECTION DE LA METHODOLOGIE

J9811

RÉPARTITION DE L'ÉCHANTILLON POUR UN PLAN DE SONDAGE À DEUX PHASES, AVEC APPLICATION AUX DONNÉES FISCALES

par

John Armstrong et Christel Le Petit

Janvier 1991

STATISTICS CANADA STATISTIQUE CANADA
OCT 18 1994
LIBRARY
BIBLIOTHÈQUE

Répartition de l'échantillon pour un plan de sondage à deux phases, avec application aux données fiscales

John Armstrong et Christel Le Petit

Abstract

As part of the Business Survey Redesign Project at Statistics Canada, the annual survey conducted in order to produce estimates of financial variables using tax data was modified. The survey involves a two-phase sample design. For the first phase of sampling, strata are defined using, among other variables, the first two digits of the Standard Industrial Classification (SIC2). Definition of second phase sampling strata involves use of four digits of the the Standard Industrial Clasification (SIC4). The objective of the survey is to produce estimates for SIC4 x province domains with coefficients of variation that do not exceed specified levels. The problem of sample allocation for this survey is examined.

1. Introduction

Le projet de remaniement des enquêtes-entreprises se déroule à Statistique Canada depuis 1985. Le but principal de ce projet est l'intégration et la standardisation des données et des systèmes pour toutes les enquêtes-entreprises. Afin d'atteindre cet objectif toutes les enquêtes utiliseront une seule base de sondage et une méthodologie généralisée. Colledge (1987) donne plus de détails concernant la stratégie de ce projet.

Statistique Canada effectue des enquêtes-entreprises de fréquence mensuelle, trimestrielle et annuelle. Cet article s'intéresse aux enquêtes annuelles traitant des variables financières, qui doivent utiliser la Base de Données du Registre Central. Cette base de données est composée de deux parties - la partie non-intégrée (PNI) et la partie intégrée (PI). Un recensement est effectué pour recueillir l'information concernant les variables financières des plus grandes entreprises, qui font partie de la PI. De plus, l'information concernant la structure de ces entreprises est maintenue sur la base. Les plus petites entreprises, quant à elles, font partie de la PNI. Des modifications à l'enquête annuelle utilisant les données fiscales de Revenu Canada ont été introduites afin d'obtenir l'information nécessaires concernant ces entreprises.

L'objectif de cette nouvelle enquête est de produire des estimations pour des variables financières pour les domaines définis par le croisement de la Classification Type des Industries, avec quatre chiffres de précision (CTI4) et des provinces. Cependant, seulement deux chiffres de la CTI sont codés avec un niveau suffisant d'exactitude à Revenu Canada.

Avant le début du projet de remaniement des enquêtes entreprises, Statistique Canada effectuait une enquête utilisant les données fiscales à Revenu Canada avec un plan de sondage à une phase. Les strates étaient définies utilisant la province et la taille (en terme de revenu brut). Le nombre d'unités échantillonnées qui tombaient dans une cellule CTI4 x province était aléatoire et dépendait de la distribution des unités dans les différents CTI4. Par conséquent, la précision des estimations variait d'une industrie à l'autre.

Afin de rendre la précision des estimations au niveau CTI4 x province plus uniforme, la nouvelle enquête utilise un plan de sondage à deux phases. La population cible de l'enquête pour la première phase est constituée de déclarants fiscaux. Il en existe deux types: les T1 et les T2. Un déclarant fiscal T1 est un individu, qui peut posséder une ou plusieurs entreprises. Par contre, un déclarant fiscal T2 est une entreprise constituée en société. Tout déclarant fiscal dont au moins une des entreprises a un revenu supérieur à vingt-cinq mille dollars est dans le champ de l'enquête, sauf s'il opère dans certains secteurs industriels qui sont exclus du champ de l'enquête. (On n'utilise pas de données fiscales concernant les services du gouvernement et on sert d'un autre plan de sondage pour les industries agricoles.)

On sélectionne l'échantillon de première phase constitué de déclarants fiscaux à Revenu Canada en utilisant des strates définies à l'aide des variables CTI2 (ou la CTI avec trois chiffres de précision pour certaines industries de construction), province et taille. L'échantillon de première phase est un échantillon longitudinal, c'est à dire que tous les déclarants fiscaux qui sont inclus dans l'échantillon de première phase de l'année Y, et qui sont encore dans le champ de l'enquête dans l'année $Y+1$, sont inclus dans l'échantillon de première phase de l'année $Y+1$. En général, on ajoutera quelques déclarants fiscaux à l'échantillon de première phase chaque année, soit pour améliorer la précision de certaines estimations, soit pour remplacer les déclarants fiscaux qui ne sont plus dans le champ de l'enquête.

L'échantillon de deuxième phase est un échantillon d'entreprises. Avant de les sélectionner, on détermine à Statistique Canada la CTI4 de chaque entreprise

appartenant aux déclarants fiscaux qui sont dans l'échantillon de première phase. Lors de la sélection de l'échantillon de deuxième phase, on utilise des strates définies au niveau CTI4 x province x taille.

La sélection de l'échantillon à chaque phase se fait par échantillonnage de Bernoulli (aussi appelé échantillonage de Poisson). On utilise des nombres aléatoires qui sont créés à partir des identificateurs des déclarants fiscaux et restent inchangés d'une année à l'autre. L'utilisation de ces nombres aléatoires permet de faciliter la sélection d'un échantillon longitudinal à la première phase. Cependant les tailles d'échantillon réalisées avec cette méthode sont aléatoires.

Pour chaque année, le processus de sélection de l'échantillon de première phase à Revenu Canada sur une période de deux ans. Donc, au moment de la spécification des fractions de sondage de l'année *Y*, le plus récent échantillon de première phase disponible et celui de l'année *Y-2*.

L'échantillon de deuxième phase de l'année *Y* est composé de deux parties: (i) un sous-échantillon des entreprises appartenant aux déclarants fiscaux qui étaient inclus dans l'échantillon de première phase de l'année *Y-2*; (ii) les entreprises appartenant aux déclarants fiscaux qui n'étaient pas dans l'échantillon de première phase de l'année *Y-2*, mais qui ont été sélectionnées dans l'échantillon de première phase de l'année *Y*. Subséquemment, on désignera les déclarants fiscaux dans la partie (ii) comme les naissances de première phase. Il est à noter qu'on ne fait pas de sous-échantillonage parmi les entreprises appartenant aux déclarants fiscaux qui sont des naissances dans l'échantillon de première phase. Choudhry, Lavallée et Hidiroglou (1989) donne plus de détails sur le plan de sondage.

Le but de cet article est d'examiner le problème de la répartition conjointe des échantillons de première et de deuxième phase. Une contrainte de précision pour une variable d'enquête, le revenu brut, a été spécifiée pour chaque domaine CTI4 x province. Par contre, il n'y a aucune contrainte de précision au niveau des strates de taille. Par conséquent, on utilise une approche liée à la méthode de répartition optimale (répartition de Neyman) pour un plan de sondage à une phase. On note que, s'il y avait des contraintes de précision associées aux strates de taille, il seraient préférable d'utiliser une approche basée, par exemple, sur la méthode de répartition exponentielle décrite par Bankier (1988).

La solution au problème de répartition pour un plan de sondage à deux phases lorsqu'il y a une seule contrainte sur l'estimation globale est bien connue et est donnée par Cochran (1977). Récemment, Smith (1989) a étudié le problème de répartition quand il y a une contrainte de précision sur une combinaison linéaire

d'estimations de domaines. Il est à noter que le problème que l'on considère dans le présent article est différent car il y a une contrainte de précision pour chacun des domaines.

Dans cet article on donne premièrement une formulation du problème de la répartition conjointe des échantillons de première et de deuxième phase. La solution de cette formulation, qu'on appelle la méthode exacte, donne la répartition optimale. On décrit en deuxième lieu une méthode d'approximation de la répartition optimale. Cette méthode a été utilisée pour répartir l'échantillon pour l'année fiscale 1990, et on discute par la suite de son application. Finalement on présente les résultats de quelques comparaisons de la méthode d'approximation et la méthode exacte. On termine par une brève conclusion.

Il y a quelques points à souligner afin de mettre la discussion suivante dans le contexte approprié. Premièrement, il y a chaque année une contrainte budgétaire portant sur la somme du nombre d'entreprises sélectionnées dans l'échantillon de deuxième phase et le nombre d'entreprises qui appartiennent aux naissances de première phase. Ce qui signifie qu'il ne sera peut-être pas possible d'augmenter suffisament l'échantillon de première phase pour satisfaire toutes les contraintes de précision en une seule année. Par conséquent, la répartition optimale sert comme objectif à long terme.

Deuxièmement, lors du développement algébrique des deux méthodes de répartition et de leur comparaison, on suppose que chaque déclarant fiscal correspond à une seule entreprise. Cette supposition n'est pas exacte car le nombre d'entreprises appartenant aux déclarants fiscaux T1 qui étaient inclus dans l'échantillon de première phase pour l'année 1988 était d'environ 5% supérieur au nombre de déclarants fiscaux inclus dans l'échantillon. On ignore également le fait qu'il n'y a pas de sous-échantillonage parmi les naissances de première phase. Finalement, afin de répartir l'échantillon de l'année Y, on utilise l'univers et l'échantillon de première phase de l'année $Y-2$. Les changements de caractéristiques des entreprises qui font partie de l'univers et de l'échantillon de première phase ont évidemment des effets sur la qualité de la répartition. Cependant ces aspects, qui seront discutés à la section quatre, ne changent pas la comparaison des deux méthodes de répartition.

2. Méthode exacte

Dans cette section, on décrit le problème de la répartition conjointe de l'échantillon de première et de deuxième phase. La solution de ce problème donne la répartition optimale pour le plan de sondage à deux phases. Afin d'expliquer la méthode, il suffit de considérer une cellule CTI2 x province contenant N unités. Chaque unité tombe dans l'une des H cellules CTI4 x province. La cellule CTI2 x province contient G strates de première phase CTI2 x province x taille et $G \cdot H$ strates de deuxième phase CTI4 x province x taille. Appelons N_g le nombre d'unités dans la strate CTI2 x province x taille g et υ_g la fraction de sondage de première phase correspondante. De même, appelons N_{gh} le nombre d'unités et υ_{gh} la fraction de sondage de deuxième phase, respectivement, pour la strate CTI4 x province x taille gh. On dénote le total, dans la population, de la variable étudiée pour la cellule CTI4 x province h par Y_h et l'estimation de Horvitz-Thompson du total par $\hat{Y}_h$.

Choudhry, Lavallée et Hidiroglou (1989) montrent qu'en conditionnant sur la taille d'échantillon obtenue, l'échantillonnage de Bernoulli est équivalent à l'échantillonnage aléatoire simple dans l'hypothèse ou la probabilité de sélection de chaque unité dans une strate est la même. Ainsi la variance de $\hat{Y}_h$ pour le plan de sondage à deux phases peut être représentée par

$$V_h = \sum_g \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot B_{gh},$$

où

$$A_{gh} = N_{gh} \cdot S^2_{gh},$$

$$B_{gh} = \left(\frac{N_g - N_{gh}}{N_g - 1} \right) \cdot \left(\frac{Y^2_{gh}}{N_{gh}} - S^2_{gh} \right),$$

et S^2_{gh} est la variance de la population dans la strate de deuxième phase CTI4 x province x taille gh.

Si on dénote par y_{ghj} la valeur de la variable à l'étude pour l'unité j de la population dans la strate de deuxième phase gh, on peut montrer que

$$A_{gh} + B_{gh} = \frac{N_g}{(N_g - 1)} \cdot \left(\sum_{j=1}^{N_{gh}} y^2_{ghj} - Y^2_{gh} / N_g \right).$$

Donc, si toutes les fractions de sondage de deuxième phase sont égales à un, la variance de $\hat{Y}_h$ correspond à la variance pour un estimation de domaine donnée, par exemple, dans Cochran (1977).

Le coût de sélection d'une unité dans l'échantillon de première phase est K_1, peu importe la strate dans laquelle se trouve l'unité, tandis que le coût de sélection d'une unité dans l'échantillon de deuxième phase est K_2, peu importe la strate. Le problème de la répartition conjointe des échantillons de première et de deuxième phase revient à trouver le minimum de la fonction de coût

$$F = K_1 \cdot \sum_g \upsilon_g \cdot N_g + K_2 \cdot \sum_g \sum_h \upsilon_g \cdot \upsilon_{gh} \cdot N_{gh}, \tag{1}$$

par rapport à υ_g, $g = 1,2,\ldots G$, et υ_{gh}, $g = 1,2,\ldots,G$, $h = 1,2,\ldots,H$ sous les contraintes

$$\sum_g \left(\frac{1}{\upsilon_g \cdot \upsilon_{gh}} - 1 \right) \cdot A_{gh} + \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot B_{gh} \leq C_h^2 \cdot Y_h^2, \quad h = 1,2,\ldots,H. \tag{2}$$

$$0 < \upsilon_g \leq 1, \quad g = 1,2,\ldots,G,$$

$$0 < \upsilon_{gh} \leq 1, \quad g = 1,2,\ldots,G, \quad h = 1,2,\ldots,H,$$

où C_h dénote le coefficient de variation cible pour le domaine d'intérêt h, équivalent à une cellule CTI4 x province.

Puisque la solution vérifiera certaines des contraintes de (2) avec inégalité stricte, des méthodes d'optimisation non linéaire doivent être utilisées pour résoudre ce problème. Toutefois, comme Wu (1989) l'a suggéré, le problème peut être sensiblement simplifié en le séparant en deux étapes principales résolues itérativement. D'abord, on trouve le minimum de (1) par rapport à υ_g, $g = 1,2,\ldots G$ sous réserve d'un ensemble de valeurs pour les fractions de sondage de deuxième phase. Cette étape exige l'emploi de techniques d'optimisation non linéaires. Il est à noter que les contraintes sont convexes et la fonction objective est concave. Donc, il s'agit d'un problème de programmation convexe, un forme particulière de problème d'optimisation non linéaire. Pour ce genre de problème, on peut montrer qu'il n'y a qu'une solution locale. Par conséquent, pour trouver la solution globale du problème, il suffit de trouver une solution locale.

La deuxième étape consiste à trouver le minimum de (1) par rapport aux fractions de sondage de deuxième phase, conditionnellement aux valeurs des fractions de

sondage de première phase obtenues à la première étape. L'obtention du minimum peut se faire indépendamment pour chaque $h = 1, 2, \ldots, H$. On recommence alors la première étape et ainsi de suite jusqu'à convergence.

Grâce à ce processus itératif, on obtient la répartition optimale en solutionnant une série de problèmes d'optimisation non linéaire, dont chacun ne comporte que G variables. Dénotons par $\upsilon_g^{(i)}$ et $\upsilon_{gh}^{(i)}$ les estimations des valeurs optimales de υ_g et de υ_{gh} obtenues après i itérations. Au début de l'itération $i+1$, les changements de variables définis par $X_g^{(i+1)} = 1/\upsilon_g^{(i+1)} - 1$ peuvent être utilisés pour transformer le problème d'optimisation en un problème comportant une fonction objective non linéaire et des contraintes linéaires convenant à l'application de méthodes de programmation non linéaire. En guise de valeurs de départ pour la première itération on pose $\upsilon_{gh}^{(0)} = 1$, $g = 1, 2, \ldots, G$, $h = 1, 2, \ldots, H$.

Plus précisément, chaque itération comporte les étapes suivantes:

(i) Recherche du minimum de

$$F = \sum_g \left(N_g + \frac{K_2}{K_1} \sum_h \upsilon_{gh}^{(i-1)} \cdot N_{gh} \right) / (X_g^{(i)} + 1)$$

par rapport à $X_g^{(i)}$, $g = 1, 2, \ldots, G$, sous les contraintes

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{X_g^{(i)} + 1}{\upsilon_{gh}^{(i-1)}} - 1 \right) \cdot A_{gh} - \sum_g X_g^{(i)} \cdot B_{gh} \geq 0, \quad h = 1, 2, \ldots, H$$

$$X_g^{(i)} \geq 0, \quad g = 1, 2, \ldots, G.$$

(ii) Calcul de $\upsilon_g^{(i)} = 1/(X_g^{(i)} + 1)$, $g = 1, 2, \ldots, G$

(iii) Recherche du minimum, pour $h = 1, 2, \ldots, H$, de

$$F_h = \sum_g \upsilon_g^{(i)} \cdot \upsilon_{gh}^{(i)} \cdot N_{gh}$$

par rapport à $\upsilon_{gh}^{(i)}$, $g = 1, 2, \ldots, G$ sous les contraintes

$$C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_g^{(i)} \cdot \upsilon_{gh}^{(i)}} - 1 \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g^{(i)}} - 1 \right) \cdot B_{gh} \geq 0,$$

$$0 < \upsilon_{gh}^{(i)} \leq 1, \quad g = 1, 2, \ldots, G.$$

On verra dans la section suivante que la solution de l'étape (iii) est analytique.

3. Méthode d'approximation

Dans cette section, on décrit une méthode de répartition qui donne une approximation de la répartition optimale. En supposant que toutes les fractions de sondage de deuxième phase sont égales à un, on calcule premièrement une approximation de la répartition optimale de l'échantillon de première phase. Ensuite, on fait la répartition de l'échantillon de deuxième phase, étant donné les fractions de sondage de première phase. Le coût de prélèvement d'une unité dans l'échantillon à chaque phase d'échantillonnage étant constant pour l'ensemble des strates, la réduction du coût équivaut à la réduction de la taille de l'échantillon à chaque étape de cette approche.

La première étape de la méthode consiste en la résolution approximative du problème de répartition pour un plan de sondage à une phase. Cette étape revient d'abord à trouver les minima, indépendamment pour chaque h, de

$$F^{(h)} = \sum_g \upsilon_{g|h} \cdot N_g \tag{3}$$

par rapport à $\upsilon_{g|h}$, $g = 1, 2, \ldots, G$, sous les contraintes

$$\sum_g \left(\frac{1}{\upsilon_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \tag{4}$$

$$0 < \upsilon_{g|h} \leq 1, \quad g = 1, 2, \ldots, G. \tag{5}$$

On peut montrer que le minimum de (3) est obtenu quand les membres de (4) sont égaux, de sorte que le problème défini par (3) et (4) équivaut à trouver le point critique du lagrangien

$$L = \sum_g \upsilon_{g|h} N_g + \lambda \cdot \left[C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \right].$$

En posant égales à zéro les dérivées par rapport aux fractions de sondage, on obtient

$$0 = N_g + \lambda \cdot (A_{gh} + B_{gh}) / \upsilon_{g|h}^2,$$

et, par conséquent,

$$\upsilon_{g|h} = ((A_{gh} + B_{gh}) / N_g)^{1/2} \cdot (-\lambda)^{1/2}, \quad g = 1, 2, \ldots, G. \tag{6}$$

En posant $\partial L / \partial \lambda = 0$ on a

$$(-\lambda)^{1/2} = \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} / \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \tag{7}$$

L'incorporation de (7) dans (6) donne

$$\upsilon_{glh} = ((A_{gh} + B_{gh})/N_g)^{1/2} \cdot \sum_g ((A_{gh} + B_{gh}) \cdot N_g)^{1/2} / \left(C_h^2 \cdot Y_h^2 + \sum_g (A_{gh} + B_{gh}) \right). \tag{8}$$

Si une ou plusieurs des fractions de sondage données par (8) sont plus grandes que un, on peut les poser égales à un et résoudre un problème modifié comportant un nombre réduit de strates. Cette méthode correspond au procédé de sur-répartition mentionné dans Cochran (1977). Enfin, on pose υ_g égal à la valeur la plus élevée de l'ensemble $\{\upsilon_{glh},\ h = 1, 2, \ldots, H\}$ pour $g = 1, 2, \ldots, G$, ce qui permet de satisfaire les contraintes pour chaque h.

Etant donné les fractions de sondage de première phase, on peut facilement trouver les fractions de sondage de deuxième phase optimales. Appelons n_{gh} le nombre d'unités choisies à la première phase se trouvant dans la strate de deuxième phase CTI4 x province x taille gh. Notons que n_{gh} ne sera pas nécessairement égal à son espérance, qui est donnée par $\upsilon_g \cdot N_{gh}$. Supposons que, pour la cellule CTI4 x province h, les strates de taille qu'on veut inclure dans le problème de répartition correspondent à un ensemble d'indices, Γ. On pose les fractions de sondage de deuxième phase égales à un pour les strates de taille qu'on n'inclut pas dans le problème de répartition. Habituellement on aura $\Gamma = \{1, 2, \ldots, G\}$ mais c'est possible, à cause de sur-répartition par exemple, que Γ ne renferme pas tous les indices entre 1 et G. Le problème de la répartition de l'échantillon de deuxième phase peut être considéré comme le problème consistant à trouver le minimum de

$$F_h = \sum_{g \in \Gamma} \upsilon_{gh} \cdot n_{gh} \tag{9}$$

par rapport à υ_{gh}, $g \in \Gamma$, sous les contraintes

$$\sum_{g \in \Gamma} \frac{1}{\upsilon_g} \cdot \left(\frac{1}{\upsilon_{gh}} - 1 \right) \cdot A_{gh} \leq M_h, \tag{10}$$

$$0 < \upsilon_{gh} \leq 1, \quad g \in \Gamma, \tag{11}$$

où

$$M_h = C_h^2 \cdot Y_h^2 - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Il est facile de montrer que (9) atteint un minimum quand l'égalité de la contrainte (10) est vérifiée. Par conséquent le problème équivaut à trouver le point critique du lagrangien

$$L_h = \sum_{g \in \Gamma} \upsilon_{gh} \cdot n_{gh} + \lambda \cdot \left(M_h - \sum_{g \in \Gamma} \frac{1}{\upsilon_g} \cdot \left(\frac{1}{\upsilon_{gh}} - 1 \right) \cdot A_{gh} \right)$$

par rapport à υ_{gh}, $g \in \Gamma$, et λ, sous les contraintes

$$0 < \upsilon_{gh} \leq 1, \quad g \in \Gamma.$$

En égalant à zéro les dérivées premières de L_h, on obtient

$$n_{gh} \cdot \upsilon_{gh}^2 + \lambda \cdot A_{gh} / \upsilon_g = 0, \quad g \in \Gamma,$$

$$M_h - \sum_{g \in \Gamma} \frac{1}{\upsilon_g} \cdot \left(\frac{1}{\upsilon_{gh}} - 1 \right) \cdot A_{gh} = 0,$$

et, après simplification,

$$\upsilon_{gh} = (-\lambda \cdot A_{gh} / (n_{gh} \cdot \upsilon_g))^{1/2}, \quad g \in \Gamma, \tag{12}$$

$$(-\lambda)^{1/2} = \sum_g (n_{gh} \cdot A_{gh} / \upsilon_g)^{1/2} / D_{\Gamma h}, \tag{13}$$

où

$$D_{\Gamma h} = C_h^2 \cdot Y_h^2 + \sum_{g \in \Gamma} \left(\frac{1}{\upsilon_g} \right) \cdot A_{gh} - \sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}).$$

Notons que le problème de répartition demeure sans solution à moins que $D_{\Gamma h}$ soit positif. En incorporant (13) dans (12), on a

$$\upsilon_{gh} = [A_{gh} / (\upsilon_g \cdot n_{gh})]^{1/2} \cdot \sum_{g \in \Gamma} (n_{gh} \cdot A_{gh} / \upsilon_g)^{1/2} / D_{\Gamma h}. \tag{14}$$

Si certaines valeurs de υ_{gh} sont plus grandes que un, on peut évidemment utiliser le procédé de sur-répartition décrit plus haut. En pratique, N_{gh}, Y_h et S_{gh}^2 sont, bien sûr, inconnus, et il est nécessaire d'utiliser des estimations fondées sur les données de l'année la plus récente disponible. Il est à noter que, pour un h donné, si on remplace n_{gh} dans (14) par son espérance par rapport à la première phase d'échantillonnage, $\upsilon_g \cdot N_{gh}$, on obtient la solution analytique de l'étape (iii) de la méthode exacte.

4. Répartition de l'échantillon pour l'année 1990

La répartition de l'échantillon des données fiscale pour l'année 1990 est effectuée selon la méthode d'approximation décrite à la section 3. Le budget total pour les deux phases est de 170 000 unités, comprenant les unités sélectionnées pour la deuxième phase et les naissances sélectionnées pour la première phase. Pour la première phase plusieurs ajustements sont nécessaires afin de tenir compte de la

longitudinalité de l'échantillon de première phase, appelé échantillon maître, du taux d'accroissement de la population et des migrations. Pour la deuxième phase la solution est directe.

4.1 Répartition de la première phase

Le problème de la répartition de la première phase pour l'année 1990 consiste à déterminer le nombre d'unités additionnelles qui devront être transcrites pour l'échantillon maître. Celui-ci comprend déjà les unités sélectionnées en 1987 et 1988, soit un total d'environ 200 000 unités. Les unités qui seront sélectionnées en 1990 pour l'échantillon maître sont choisies parmi les unités qui n'ont pas déjà été sélectionnées pour celui-ci en 1987 ou en 1988. Ce sont donc des naissances à l'échantillon maître. La répartition est effectuée à l'aide des données de 1988. La taille de la population est de 1 million en 1988, répartie parmi 12 provinces ou territoires, 5 groupes de taille, 80 CTI2 et 700 CTI4. Le niveau de publication étant le croisement CTI4 x province, cela donne un total de 8400 domaines. Bien que 5200 domaines soient non-vides, satisfaire un coefficient de variation (CV) de 3% pour chacun de ces domaines demanderait une augmentation de l'échantillon de première phase beaucoup trop grande pour le budget alloué. Afin de réduire la taille d'échantillon, les domaines sont séparés en deux groupes. Les domaines importants correspondent aux CTI4 x province qui contiennent plus que 10% du revenu du CTI2 x province correspondant, et les domaines moins importants sont ceux qui contiennent moins que 10% du revenu. Parmi nos 5200 domaines non-vides, 40% sont considérés important. Le CV cible est fixé à 3% pour les domaines importants, et 20% pour les domaines moins importants.

Lors de la répartition de l'échantillon, certaines des fractions de sondages obtenues sont supérieures à un. Cette situation se produit lorsque la fraction de sondage globale pour une cellule CTI2 x province est grande, et que les variances sont très différentes entre les strates de taille. Dans cette situation, on utilise le procédé de sur-répartition.

On obtient donc des fractions de sondages pour chaque CTI4 x province x taille. Mais la stratification pour la première phase étant seulement au niveau CTI2 x taille x province on prendra comme fraction pour une CTI2 la fraction maximale parmi tous les CTI4 de cette CTI2, afin de satisfaire toutes les contraintes pour tous les domaines.

Pour certaines combinaisons CTI2 x province il est possible de ne pas avoir une estimation de la variance pour aucune des CTI4, la taille d'échantillon étant un ou zéro dans toutes les strates de taille. La fraction de sondage est alors fixée à un pour toutes les strates de taille dans la cellule CTI2 x province, dans l'objectif d'avoir à long terme assez d'unités dans l'échantillon maître pour pouvoir obtenir des estimations.

Il est nécessaire d'appliquer une série d'ajustements à ces fractions de sondage. Premièrement il faut concilier les fractions de sondage obtenues par la répartition (qu'on va appeler les fractions de sondage calculées) avec le nombre d'unités qui a déjà été sélectionné dans l'échantillon maître en 1987 et 1988. Une solution serait de comparer la fraction calculée avec le maximum des fractions de 1987 et 1988. Cependant, la population étant très mobile entre les strates, la différence entre la fraction de sondage calculée et le nombre d'unités qui est réellement sélectionné pour une strate donnée est assez grande pour que cette approche ne soit pas retenue.

Une solution plus exacte est de calculer, pour chaque strate, le pourcentage d'unités dans la population qui fait déjà partie de l'échantillon maître, qu'on appelle la fraction de sondage réalisée. Les fractions de sondages calculées sont alors comparées aux fractions de sondages réalisées. Si la fraction de sondage calculée est inférieure à la fraction réalisée en 1987 et 1988, on fixe à 2% la fraction de sondage de 1990 afin de sélectionner seulement une faible proportion de naissances. Par contre, si la fraction de sondage calculée est supérieure à la fraction de sondage réalisée en 1987 et 1988, la fraction de sondage de 90 est déterminée par la différence entre les deux, soit $v_g - f_{87-88}$, ou v_g représente la fraction calculée, et f_{87-88} la fraction réalisée dans l'échantillon maître en 1987 et 1988. On ne sélectionne donc que la proportion nécessaire pour obtenir un pourcentage global correspondant à la fraction calculée.

Deuxièmement, il est nécessaire de prendre en compte le taux d'accroissement de la population. Ce taux est estimé à 9% par année, soit 18.81% pour deux ans. Compte tenu de ce fait, on estime qu'environ 50 000 unités seront sélectionnées selon les fractions de sondages. Cependant ce nombre de naissance est sous-estimé puisque l'on utilise un taux d'accroissement net de la population au lieu d'un taux de naissance qui serait le taux d'accroissement net plus le taux de mortalité.

Troisièmement il faut prendre en compte le taux de migrations des unités. En effet, quand une unité qui n'était pas sélectionnée auparavant change de strate (changement de groupe de taille, de CTI2 ou de province), elle peut être sélectionnée dans sa nouvelle strate si la nouvelle fraction de sondage est supérieure à

l'ancienne fraction de sondage. Une estimation du pourcentage d'unité qui seront sélectionnées à cause d'une migration a été effectuée avec les données de 1986 et 1988. On estime que les migrations vont causer une augmentation de 56% du nombre d'unités sélectionnées, soit 28 000 unités supplémentaires. Ce qui donne un nombre total de 78 000 unités pour la première phase. Ces unités sont également sélectionnées pour la deuxième phase avec probabilité un.

Toute l'information nécessaire à la sélection de l'échantillon, soit la définition des strates, les fractions de sondages, etc... est envoyée à Revenu Canada par le biais de fichiers de paramètres. La sélection des unités est effectuée à Revenu Canada, au fur et à mesure que les dossiers sont traités par leur système.

4.2 Répartition de la deuxième phase

La répartition de l'échantillon de deuxième phase est effectuée de façon directe, selon la méthode décrite à la section 3. Le CV cible est de 3% pour tous les domaines. Il n'est évidemment pas toujours possible d'obtenir un CV de 3%, si la taille de l'échantillon maitre n'est pas suffisante. Cette situation s'est produite pour 12% des 5200 domaines non-vides. Dans le cas où on obtient des fractions de sondages supérieures à un, on utilise encore le processus de sur-répartition. De plus, un ajustement pour les petites cellules est effectué afin d'avoir une taille d'échantillon minimum.

Lors de la répartition de la deuxième phase on ne tient pas compte des unités sélectionnées en première phase comme naissances qui feront aussi partie de la deuxième phase. Ces unités sont un "bonus" qui nous aidera à obtenir des estimations encore plus précises. Cependant ce facteur devrait être incorporé dans le processus de la répartition de l'échantillon de deuxième phase pour les années subséquentes afin d'effectuer une meilleure allocation des ressources.

Une fois les fractions de sondage calculées, les unités sont sélectionnées à Statistique Canada parmi les unités faisant partie de l'échantillon maitre de 1987 et 1988. La liste des unités choisies est envoyée à Revenu Canada, et comprend 92 000 unités. Ces 92 000 unités représentent donc l'échantillon nécessaire pour obtenir un CV de 3% dans les domaines où la taille de l'échantillon maitre le permet. Cependant un certain nombre de ces unités préspécifiées seront mortes en 1990, ne permettant plus d'obtenir un CV de 3%. Mais à ce sous-échantillon de l'échantillon maitre de 1987 et 1988 on ajoute les unités qui sont sélectionnées pour la première phase comme naissance à l'échantillon maitre et qui font aussi

partie de l'échantillon de deuxième phase avec probabilité un, soit 78 000 unités supplémentaires. Donc, la taille totale de l'échantillon de deuxième phase devrait permettre d'obtenir des CV de 3% et même inférieurs à 3% pour plusieurs domaines.

5. Comparaison des deux méthodes

On effectue ici une comparaison entre la méthode exacte et la méthode approximative décrites dans les sections 2 et 3. On considère entre autres le cas particulier où $G=2$ et $H=2$. Un bref rappel des deux méthodes est fait, suivi de l'analyse des situations où elles sont équivalentes et des situations où la méthode approximative donne de moins bons résultats que la méthode exacte. Une étude des coûts engendrés par l'utilisation de la méthode approximative est aussi présentée.

La méthode approximative consiste à résoudre H problèmes d'optimisation linéaire, soit à minimiser pour chaque h la fonction

$$F^{(h)} = \sum_g \upsilon_{g|h} \cdot N_g \tag{3}$$

soumise aux $G+1$ contraintes

$$\sum_g \left(\frac{1}{\upsilon_{g|h}} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2,$$

$$0 < \upsilon_{g|h} \leq 1, \quad g = 1, \ldots, G.$$

La notation utilisée est la même que dans les sections précédentes.

On compare la méthode approximative avec la première itération de la méthode exacte. Pour cette première itération on pose $\upsilon_{gh} = 1$ pour tout g et h. La fonction à minimiser qui était donnée par (1) devient alors

$$F = K_1 \cdot \sum_g \upsilon_g \cdot N_g + K_2 \cdot \sum_g \upsilon_g \cdot N_g,$$

soumise aux $G+H$ contraintes

$$\sum_g \left(\frac{1}{\upsilon_g} - 1 \right) \cdot (A_{gh} + B_{gh}) \leq C_h^2 \cdot Y_h^2, \quad h = 1, \ldots, H,$$

$$0 < \upsilon_g \leq 1, \quad g = 1, \ldots, G.$$

Comme dans la section 2 on effectue un changement de variables afin de transformer un problème d'optimisation d'une fonction linéaire sur une région définie par des fonctions non-linéaires en un problème d'optimisation non-linéaire sur une région définie par des fonctions linéaires.

De plus, afin de simplifier la discussion, on étudie le cas où $G=2$ et $H=2$.

Posons

$$X_1 = \frac{1}{\upsilon_1} - 1, \quad X_2 = \frac{1}{\upsilon_2} - 1.$$

Comme $H=2$, on a deux problèmes à résoudre pour la méthode approximative. Pour $h=1$ on doit premièrement minimiser

$$F^{(1)} = \frac{N_1}{X_{111}+1} + \frac{N_2}{X_{211}+1},$$

sous les contraintes

$$X_{111} \cdot (A_{11} + B_{11}) + X_{211} \cdot (A_{21} + B_{21}) \leq C_1^2 \cdot Y_1^2, \tag{17}$$

$$X_{111} \geq 0, \quad X_{211} \geq 0. \tag{18}$$

Deuxièmement, pour $h=2$, on doit minimiser

$$F^{(2)} = \frac{N_1}{X_{112}+1} + \frac{N_2}{X_{212}+1},$$

sous les contraintes

$$X_{112} \cdot (A_{12} + B_{12}) + X_{212} \cdot (A_{22} + B_{22}) \leq C_2^2 \cdot Y_2^2, \tag{19}$$

$$X_{112} \geq 0, \quad X_{212} \geq 0. \tag{20}$$

Une fois que X_{111}, X_{112}, X_{211} et X_{212} qui minimisent $F^{(1)}$ et $F^{(2)}$ sont obtenus on calcule υ_1 et υ_2 de la façon suivante:

$$X_1 = \min(X_{111}, X_{112}), \quad X_2 = \min(X_{211}, X_{212}),$$

$$\upsilon_1 = \frac{1}{X_1+1}, \quad \upsilon_2 = \frac{1}{X_2+1}.$$

Pour la solution exacte, la fonction à minimiser dans le cas $G=2$ et $H=2$ est donnée par

$$F = \left(1 + \frac{K_1}{K_2}\right) \cdot \left\{ \frac{N_1}{X_1 + 1} + \frac{N_2}{X_2 + 1} \right\}. \tag{21}$$

Celle-ci est équivalente à $F^{(1)}$ et à $F^{(2)}$ à une constante près. On minimise F simultanément sous les contraintes de $F^{(1)}$ et de $F^{(2)}$ soit:

$$X_1 \cdot (A_{11} + B_{11}) + X_2 \cdot (A_{21} + B_{21}) \leq C_1^2 \cdot Y_1^2,$$

$$X_1 \cdot (A_{12} + B_{12}) + X_2 \cdot (A_{22} + B_{22}) \leq C_2^2 \cdot Y_2^2,$$

$$X_1 \geq 0, \quad X_2 \geq 0.$$

La région définie par les contraintes étant convexe il s'agit d'un problème de programmation convexe. Dans cette situation, si on trouve un minimum local il s'agira aussi d'un minimum global. Une fois que X_1 et X_2 qui minimisent F sont trouvés on obtient υ_1 et υ_2 comme

$$\upsilon_1 = \frac{1}{X_1 + 1}, \quad \upsilon_2 = \frac{1}{X_2 + 1}.$$

On voit donc que la fonction à minimiser pour les deux approches est la même, la différence étant dans le nombre de contraintes. Ainsi, pour l'approche approximative on trouve deux solutions sous deux séries de contraintes puis on prend le minimum des deux solutions, alors que pour la solution exacte on minimise la même fonction sous toutes les contraintes simultanément.

On peut représenter les contraintes graphiquement. Les schémas 1 et 2 représentent deux cas avec deux ensembles de contraintes différents.

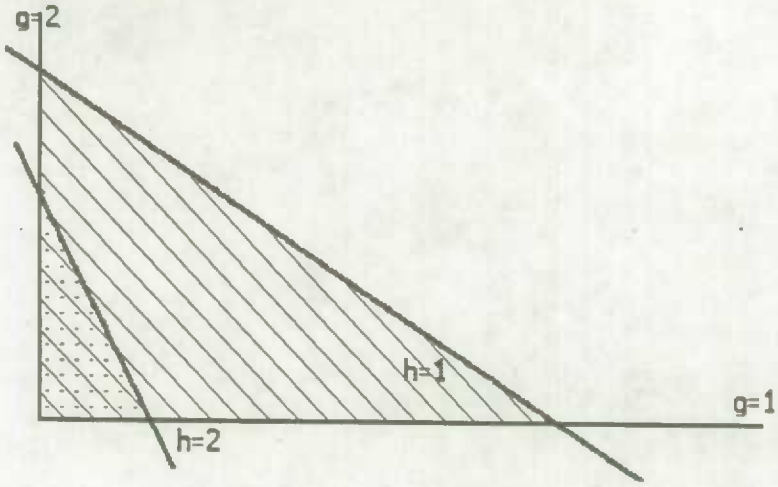

Schéma 1: Type de région

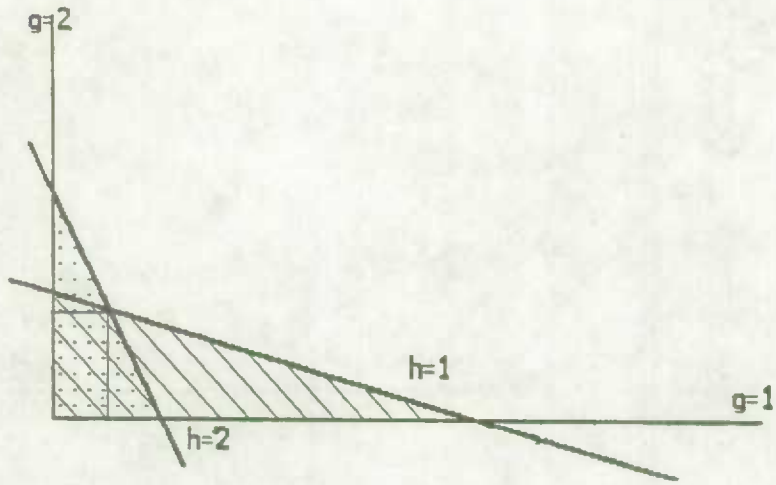

Schéma 2: Type de région

Sur ces schémas les strates de taille sont représentées par les axes. Les droites *h=1* et *h=2* représentent les contraintes liées à la précision du CV pour chaque *h*. Les contraintes pour le premier problème de la solution approximative (équations 17 et 18) définissent la région hachurée. Celles du deuxième problème (équations 19 et 20) définissent la région identifiée par des points. Quant à la solution exacte, elle consiste à minimiser F sous toutes les contraintes, soit sur la région représentée par l'intersection des zones hachurées et à points.

Les deux schémas représentent deux types de région différentes. Dans le cas du schéma 1 la contrainte liée à *h=1* est plus restrictive que celle liée à *h=2* pour toutes les valeurs de X_1 et X_2, c'est à dire qu'elle est plus proche des axes. Dans le cas du schéma 2, aucune des deux contraintes n'est plus restrictive que l'autre pour toutes les valeurs de X_1 et X_2. La contrainte liée à *h=1* est plus restrictive pour les valeurs de X_1 allant de 0 à 2 et pour les valeurs de X_2 allant de 4 à 4.5, alors que la contrainte liée à *h=2* est plus restrictive pour les valeurs de X_1 allant de 2 à 4 et les valeurs de X_2 allant de 0 à 4.

Étudions en premier lieu la solution approximative dans le cas où on a une région similaire à celle du schéma 1, soit la situation où une contrainte est plus restrictive que l'autre pour toutes les valeurs de X. L'équation (3) définie une famille de fonctions pour chaque h et chaque membre de cette famille correspond à une valeur fixe de $F^{(h)}$. On veut trouver la fonction de cette famille qui est aussi loin que possible de l'origine tout en restant dans la région pour maximiser X_1 et X_2. Quelque unes des fonctions de cette famille sont représentées au schéma 3.

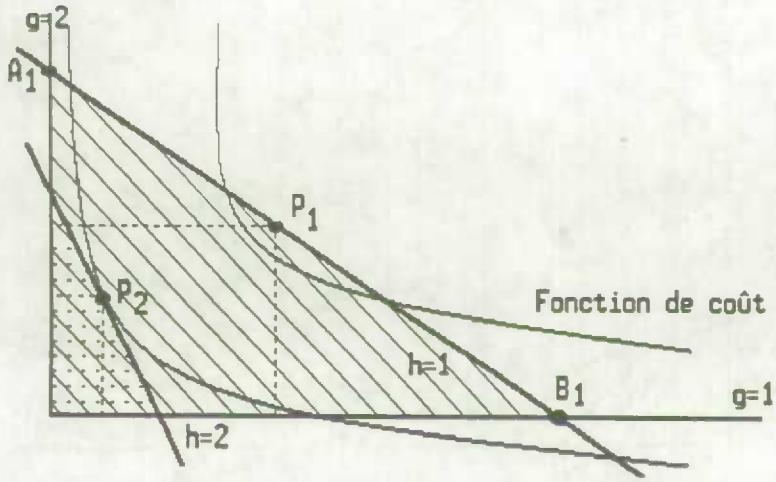

Schéma 3: Méthode approximative

La solution du premier problème ($h=1$) de la solution approximative est représentée par P_1 alors que la solution du deuxième problème ($h=2$) est représentée par P_2. Une fois ces deux points obtenus on prend les fractions de sondage maximales qui satisferont les deux contraintes, soit les valeurs minimales de X_1 et X_2. Dans notre cas, le point représentant ce minimum est P_2 car il a une plus petite abcisse et une plus petite ordonnée.

Si on utilise maintenant l'approche exacte sur ce même type de région, la première itération consiste à minimiser la même fonction F sur la région définie par l'intersection des zones hachurées et à points (voir schéma 4).

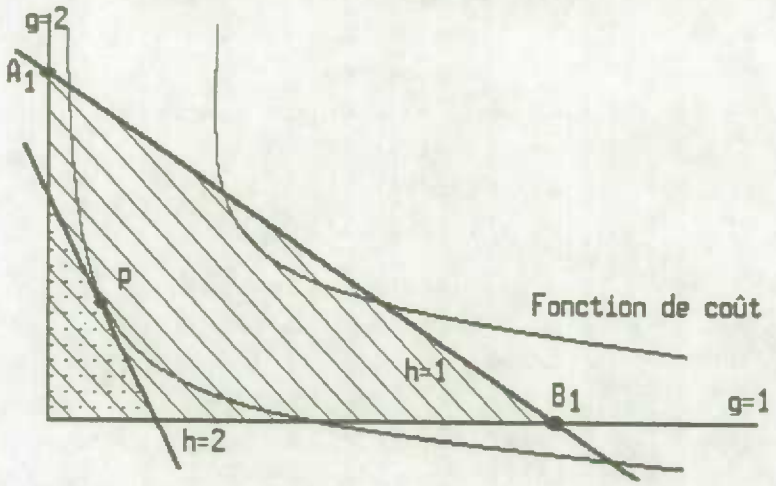

Schéma 4: Méthode exacte, première itération

Le point optimal est représenté par le point P, qui correspond exactement au point P_2 du schéma 3. Donc, dans le cas où le même h détermine la fraction de sondage maximale dans l'ensemble $\{\upsilon_{gh}, h = 1, \ldots, H\}$ pour tous les g, la solution approximative et la première itération de la solution exacte donne la même solution. Une fois les υ_g obtenus à la première itération, on calcule les υ_{gh} selon la méthode décrite à la section 3.

Pour la deuxième itération, la fonction à minimiser ainsi que les contraintes changent car les valeurs de υ_{gh} sont différentes de 1. On obtient une autre solution P'', qui correspond à d'autres fractions de sondages de première phase qui sont utilisées pour calculer un autre ensemble de fractions de deuxième phase, et ainsi de suite jusqu'à convergence. Notons qu'à chaque itération la valeur de la fonction de coût F diminue.

Il y a cependant des régions où ce n'est pas le même h qui détermine la fraction de sondage maximale pour chaque g. On se trouve alors dans la situation présentée par le schéma 5.

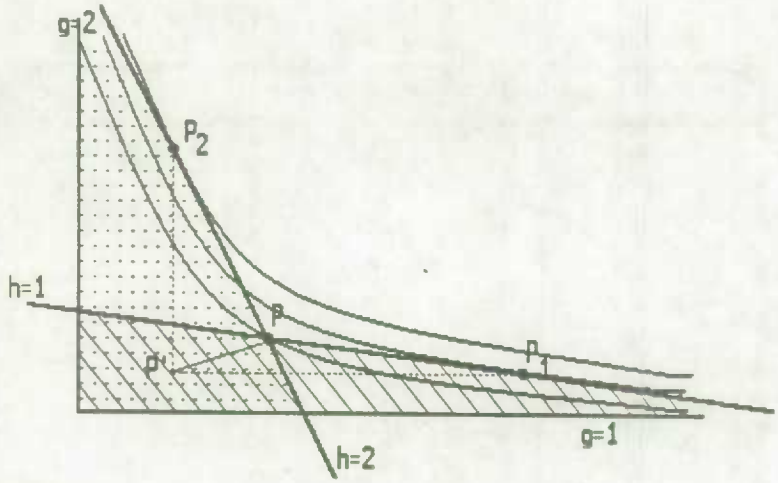

Schéma 5: Comparaison des méthodes exacte et approximative

Le point P_1 représente la solution du premier problème de la méthode approximative et le point P_2 la solution du deuxième problème. Quand on prend le minimum de ces deux points on obtient le point P' qui est la solution de la méthode approximative. La méthode exacte quant à elle nous donne le point P comme solution qui est plus loin de l'origine que P' et correspond donc à des fractions de sondages inférieures. Dans cette situation la méthode exacte donne de meilleurs résultats dès la première itération. La distance entre P et P' montre qu'il y

a une perte d'efficacité due à l'utilisation de la méthode approximative. On calcule ensuite les υ_{gh} pour la deuxième phase, et on continue le processus itératif jusqu'à convergence.

La discussion des paragraphes précédents peut être généralisée aux dimensions $g=2,\ldots,5$ et $h=1,\ldots,H$ pour chaque combinaison CTI2 x province. Le groupe de taille $g=1$ représente la partie intégrée où la fraction de sondage est un pour toutes les combinaisons CTI2 x province et ne fait donc pas partie de la discussion.

Afin de comparer les deux méthodes on discrimine les cellules CTI2 x province en trois groupes selon le critère suivant:

GROUPE A: Le même h détermine la fraction de sondage maximale pour les quatre g.

GROUPE B: Le même h détermine la fraction de sondage maximale pour trois g parmi quatre.

GROUPE C: Le même h ne détermine pas la fraction de sondage maximale pour plus que deux g parmi quatre.

PROVINCE	POPULATION	ECHANTILLON
TN	10 551	4 197
IPE	2 563	1 032
NE	22 341	7 335
NB	18 753	6 229
QC	230 040	18 269
ON	349 557	15 737
MB	33 245	8 482
SK	26 863	6 886
AB	97 321	13 427
CB	129 748	15 190
TY	964	544
TNO	692	388
TOTAL	922 638	97 716

Tableau 1: Taille de population et d'échantillon par province

Le tableau 1 donne les tailles de la population et de l'échantillon pour chaque province et territoire. La répartition de cet échantillon a été effectuée avec la méthode approximative.

Le schéma 6 donne la répartition des unités de l'échantillon entre les groupes A, B et C définis plus haut. Pour l'ensemble du Canada la répartition est 17% des unités dans le groupe A, 28% dans le groupe B et 55% dans le groupe C.

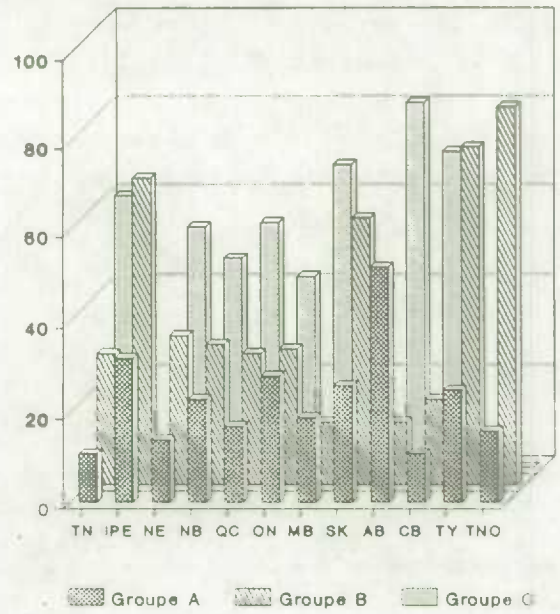

Schéma 6: Répartition des unités de l'échantillon entre les strates des groupes A, B et C

Afin d'étudier l'ampleur du gain obtenu en utilisant la solution exacte au lieu de la solution approximative, l'étude suivante a été faite pour le Québec. Pour chaque CTI2 non-vide, les fractions de sondages de première et deuxième phase ont été calculées avec les deux approches. Pour obtenir les coûts associés à ces fractions de sondage il faut déterminer le rapport entre K_1 et K_2 les coûts de sélection de première et deuxième phase. Les coûts de sélection sont identiques à la première et à la deuxième phase, soit environ sept dollars. Cependant, l'échantillon de première phase étant longitudinal, une unité sélectionnée une année donnée reste dans l'échantillon de première phase les années subséquentes,

alors que l'échantillon de deuxième phase est sélectionné chaque année indépendamment des années précédentes. Le rapport entre les coûts de première phase et de deuxième phase a été fixé pour l'étude à $K_2/K_1 = 5$, ce qui suppose qu'une unité de première phase reste en moyenne cinq ans dans l'échantillon avant de mourir. Lors de la comparaison des deux méthodes avec les données du Québec, on a posé $K_1 = 7/5$ et $K_2 = 7$. On note que les coûts d'entretien de l'échantillon de première phase (révision cyclique des codes CTI4, élimination des unités mortes, etc...) n'ont pas été pris en compte dans notre étude.

Afin de solutionner le problème de programmation non-linéaire qui fait partie de la méthode exacte, on a utilisé des sous-routines de la librairie IMSL (1987) qui sont basées sur l'algorithme de Schittkowski (1985). Cet algorithme est sensible aux points de départ. La solution approximative a été utilisée comme point de départ, et des essais avec d'autres points de départ ont également été effectuées. L'algorithme n'a pas réussi à atteindre la solution lors de tous les essais.

Le tableau 2 présente, pour chaque groupe, la différence nette entre les deux coûts et la différence en terme de pourcentage de perte obtenue utilisant la solution approximative par rapport à la solution exacte. Le détail de ces coûts pour chaque CTI2 est donné à l'annexe A. La description des codes correspondants aux CTI2 est présentée à l'annexe B.

Groupe	Coût solution exacte	Coût solution approximative	Différence nette	Différence %
A	19 444.60	22 111.45	2 666.85	12.1%
B	32 722.92	34 787.20	2 064.28	5.9%
C	58 643.67	60 493.93	1 850.26	3.1%
TOTAL	110 811.19	117 392.58	6 581.39	5.6%

Tableau 2: Coût des deux méthodes pour les groupes A, B et C

Le gain total obtenu grâce à l'utilisation de la solution exacte est de 6% seulement. Le gain maximal avec la solution exacte a été réalisé pour le groupe A, soit le groupe où les deux méthodes donnaient le même résultat après une itération. Ceci implique que le gain provient des itérations subséquentes. Pour les deux autres groupes, bien que la méthode exacte donne de meilleurs résultats

après une itération, les itérations suivantes n'ont permis qu'une amélioration inférieure à celle du groupe A. Ceci peut être expliqué par la nature des données. Il y a moins de groupes de taille vides pour les domaines du groupe A. Par contre pour les strates des groupes B et C, on avait davantage de domaines vides pour des groupes de taille. Or ces groupes de taille vides peuvent être considérés comme des contraintes supplémentaires, qui limitent ainsi les gains pouvant être obtenus avec la méthode exacte.

Même si le coût total avec la méthode exacte est toujours inférieur à celui de la méthode approximative, l'examen des résultats a montré que le coût pour la première phase était souvent supérieur avec la méthode exacte et que celui de la deuxième phase était inférieur. Ceci vient du fait que le rapport entre K_1 et K_2 a été fixé à 5. Il revient donc à moins cher de prendre d'avantage d'unités à la première phase, ce qui donne une meilleure base de sondage pour la deuxième phase et permet de sélectionner moins d'unités.

La simulation a été effectuée pour le Québec, soit la deuxième plus grande province en terme de population d'entreprises au Canada et le gain total en terme de coût grâce à l'utilisation de la méthode exacte est seulement de 6%. Comme on l'a vu, la solution exacte semble donner de meilleurs résultats pour les domaines complets et homogènes. Ceci laisse à penser que si on appliquait cette méthode aux petites provinces, les gains en terme de coût ne seraient pas importants.

6. Conclusion

Cette étude nous a permis de comparer deux approches qui peuvent être utilisées pour résoudre le problème de la répartition de l'échantillon des données fiscales. L'une de ces approches, la solution approximative, est facile à implanter et a déjà été utilisée pour la répartition de l'échantillon de 1990. L'autre approche, la solution exacte, offre une réduction en terme du coût, mais pose également des problèmes d'application avec des données réelles. On constate que l'algorithme arrête souvent avant de trouver une solution et semble très sensible aux points de départ qui lui sont offerts. De plus, lorsque les domaines ne sont pas complets pour chaque groupe de taille, ce qui est souvent le cas avec les données réelles, le gain obtenu peut-être considéré comme marginal. Ces dernières remarques nous incitent donc à continuer d'utiliser la solution approximative pour la répartition de l'échantillon des données fiscales.

Remerciements

Les auteurs tiennent à remercier Pierre Lavallée, Jean Leduc, Georges Lemaître et Alain Théberge pour leurs commentaires utiles sur les versions intermédiaires du présent document.

Annexe A

Coût des deux méthodes pour chaque CTI2

CTI2[1]	Coût solution exacte	Coût solution approximative	Différence nette	Différence %
04	1 863.69	2 633.15	769.46	29.2
05	49.11	49.11	0.00	0.0
06	54.28	58.62	4.34	7.4
28	1 752.03	2 097.52	345.49	16.5
31	483.62	544.19	60.57	11.1
36	70.78	73.64	2.86	3.9
40	180.36	181.76	1.40	0.8
422	852.26	1 149.85	297.59	25.9
426	164.89	164.89	0.00	0.0
44	975.83	1 009.51	33.68	3.3
48	1 164.60	1 184.75	20.15	1.7
51	46.81	46.81	0.00	0.0
52	257.33	259.35	2.02	0.8
54	888.59	988.06	99.47	10.1
61	3 206.70	3 603.28	396.58	11.0
64	358.06	524.23	166.17	31.7
69	1 891.22	2 142.15	250.93	11.7
70	126.80	126.80	0.00	0.0
72	3 580.25	3 580.25	0.00	0.0
73	8.95	8.95	0.00	0.0
76	249.12	249.12	0.00	0.0
98	1 219.32	1 435.46	216.14	15.1
TOTAL	19 444.60	22 111.45	2 666.85	12.1

Tableau 3: Coûts pour chaque CTI2 du groupe A (même *h* pour les quatre *g*)

1 La méthode exacte et la méthode d'approximation sont équivalentes pour les CTI2 05, 426, 51 et 76 car il y a un seul CTI4 parmi la CTI2.

CTI2[2]	Coût solution exacte	Coût solution approximative	Différence nette	Différence %
02	1 245.31	1 245.31	0.00	0.0
07	10.76	10.76	0.00	0.0
08	376.33	427.73	51.40	12.0
09	338.04	381.44	43.40	11.4
12	8.40	8.40	0.00	0.0
18	96.99	98.13	1.14	1.2
19	600.59	675.40	74.81	11.1
26	1 248.77	1 248.77	0.00	0.0
27	156.88	207.54	50.66	24.4
30	1 897.46	2 010.69	113.23	5.6
39	1 450.30	1 529.19	78.89	5.2
41	287.18	288.79	1.61	0.6
421	2 916.09	2 997.81	81.72	2.7
423	2 426.56	2 595.75	169.19	6.5
427	3 381.49	4 078.02	696.53	17.1
429	963.16	1 012.73	49.57	4.9
47	357.59	374.25	16.66	4.5
53	742.59	742.59	0.00	0.0
57	2 129.41	2 281.29	151.88	6.7
60	7 482.17	7 482.17	0.00	0.0
62	2 270.67	2 296.66	25.99	1.1
75	1 242.84	1 456.04	213.20	14.6
85	1 093.34	1 337.74	244.40	18.3
TOTAL	32 722.92	34 787.20	2 064.28	5.9

Tableau 4: Coûts pour chaque CTI2 du groupe B (même *h* pour trois *g* parmi quatre)

2 La méthode exacte et la méthode d'approximation sont équivalentes pour les CTI2 07 et 12 car il y a un seul CTI4 parmi la CTI2.

CTI2	Coût solution exacte	Coût solution approximative	Différence nette	Différence %
03	659.68	682.70	23.02	3.4
10	656.99	674.93	17.94	2.7
11	48.94	48.94	0.00	0.0
15	50.96	50.96	0.00	0.0
16	268.25	268.25	0.00	0.0
17	342.38	362.49	20.11	5.5
24	1 271.40	1 271.40	0.00	0.0
25	871.51	879.59	8.08	0.9
29	124.81	126.06	1.25	1.0
32	558.74	563.18	4.44	0.8
33	324.45	338.42	13.97	4.1
35	508.41	512.13	3.72	0.7
37	208.57	208.57	0.00	0.0
424	966.06	971.87	5.81	0.6
425	694.88	699.66	4.78	0.7
45	14 513.53	14 817.28	303.75	2.0
49	425.96	430.39	4.43	1.0
55	252.07	253.06	0.99	0.4
56	830.07	834.88	4.81	0.6
59	3 119.76	3 119.76	0.00	0.0
63	3 223.33	3 338.13	114.80	3.4
65	9 650.59	9 982.38	331.79	3.3
71	392.84	409.28	16.44	4.0
74	1 095.26	1 122.03	26.77	2.4
77	8 810.59	8 893.70	83.11	0.9
91	1 868.38	1 868.38	0.00	0.0
92	6 905.26	7 765.51	860.25	11.1
TOTAL	58 643.67	60 493.93	1 850.26	3.1

Tableau 5: Coûts pour chaque CTI2 du groupe C (même h pour deux g ou moins)

Annexe B

Descriptions des codes de classification type des industries à deux chiffres

03: Industries de la pêche et du piégeage
04: Industries de l'exploitation forestière et des services forestiers
05: Industries des services forestiers
06: Industries des mines
07: Industries du pétrole brut et du gaz naturel
08: Industries des carrières et sablières
09: Industries des services miniers
10: Industries des aliments
11: Industries des boissons
12: Industries du tabac
15: Industries des produits en caoutchouc
16: Industries des produits en matière plastique
17: Industries du cuir et des produits connexes
18: Industries textiles de première transformation
19: Industries des produits textiles
24: Industries de l'habillement
25: Industries du bois
26: Industries du meuble et des articles d'ameublement
27: Industries du papier et produits connexes
28: Imprimerie, édition et industries connexes
29: Industries de première transformation des métaux
30: Industries de la fabrication des produits métalliques (sauf les industries de la machinerie et du matériel de transport)
31: Industries de la machinerie (sauf électrique)
32: Industries du matériel de transport
33: Industries des produits électriques et électroniques
35: Industries des produits minéraux non métalliques
36: Industries des produits raffinés du pétrole et du charbon
37: Industries chimiques
39: Autres industries manufacturières
40: Industries des constructeurs, promoteurs et entrepreneurs généraux
41: Industries de la construction lourde et industrielle

421: Travaux sur chantier
422: Travaux de charpente et travaux connexes
423: Travaux extérieurs de finition
424: Installations mécaniques, plomberie, chauffage, et climatisation
425: Travaux de mécanique spécialisée
426: Travaux d'électricité
427: Travaux d'intérieur et de finition
429: Autres travaux spécialisés
44: Industries des services relatifs à la construction
45: Industries du transport
46: Industries du transport par pipelines
47: Industries de l'entreposage et de l'emmagasinage
48: Industries des communications
49: Autres industries de services publics
50: Industries des produits agricoles, commerce de gros
51: Industries des produits pétroliers, commerce de gros
52: Industries des produits alimentaires, boissons, médicaments et tabac, commerce de gros
53: Industries du vêtement et de la mercerie, commerce de gros
54: Industries des articles ménagers, commerce de gros
55: Industries des véhicules automobiles, pièces et accessoires, commerce de gros
56: Industries des métaux, articles de quincaillerie, matériel de plomberie et de chauffage, matériaux de construction, commerce de gros
57: Industries des machines, matériel et fournitures, commerce de gros
59: Industrie de produits divers, commerce de gros
60: Industries des aliments, boissons et médicaments, commerce de détail
61: Industries des chaussures, vêtements, tissus et filés, commerce de détail
62: Industries des meubles, appareils et accessoires d'ameublement de maison, commerce de détail
63: Industrie des véhicules automobiles, pièces et accessoires, vente et service
64: Industries des magasins de marchandises diverses
65: Autres industries de magasins de détail
69: Industries de commerce de détail hors magasins
70: Industries des intermédiaires financiers de dépôts
71: Industries du crédit à la consommation et aux entreprises
72: Industries des intermédiaires d'investissements
73: Industries des assurances

74: Autres industries d'intermédiaires financiers
75: Industries des services immobiliers (sauf les lotisseurs)
76: Industries des agences d'assurances et agences immobilières
77: Industries des services aux entreprises
85: Industries des services d'enseignement
86: Industries des services de soins de santé et des services sociaux
91: Industries de l'hébergement
92: Industries de la restauration
96: Industries de services de divertissement et loisirs
97: Industries des services personnels et domestiques
98: Associations
99: Autres industries de services

Ca 603

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010175212

Bibliographie

BANKIER, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

CHOUDHRY, G.H., LAVALLÉE P., et HIDIROGLOU, M. (1989). Two-phase sample design for tax data. Communication présentée lors de la réunion annuelle de l'American Statistical Association.

COCHRAN, W.G. (1977). Sampling Techniques, (3e éd.). New York: John Wiley.

COLLEDGE, M. (1987). The business survey redesign project: implementation of a new strategy at Statistics Canada. U.S. Bureau of the Census, Proceedings of the Third Annual Research Conference, 550-576.

IMSL (1987). Math/Library FORTRAN Subroutines for Mathematical Applications. Houston: IMSL Inc.

SCHITTKOWSKI, K. (1985). NLPQL: A FORTRAN subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5, 485-500.

SMITH, P.J. (1989). Is two-phase sampling really better for estimating age composition?. *Journal of the American Statistical Association*, 84, 916-921.

WU, J. (1989). Procès-verbal de la réunion du Comité consultatif des méthodes statistiques, les 3-4 avril, 1989. Document interne, Statistique Canada.