LE SONDAGE INDIRECT APPLIQUÉ AUX POPULATIONS ASYMÉTRIQUES

Pierre LAVALLÉE, Sébastien LABELLE-BLANCHET

Division des méthodes d'enquête auprès des entreprises, Statistique Canada

Introduction

Les enquêtes économiques diffèrent à plusieurs égards des enquêtes sociales. L'une des différences tient au fait que la base de sondage des firmes contient surtout des variables fortement asymétriques, tandis que celle utilisée pour les enquêtes sociales est beaucoup plus homogène. Habituellement, le sondage des firmes est effectué en transformant les structures opérationnelles en unités normalisées appelées unités statistiques, qui sont ordinairement représentées sous forme d'une hiérarchie ou d'une série de niveaux permettant l'intégration subséquente des divers éléments de données disponibles aux divers niveaux de l'organisation. Le nombre de niveaux que compte la hiérarchie diffère selon l'organisme statistique. Par exemple, le Registre des entreprises du Canada comprend quatre niveaux : l'entreprise, la compagnie, l'établissement et l'emplacement. Dans la plupart des cas, l'entreprise statistique correspond à l'entité juridique. L'établissement statistique équivaut, dans la plupart des cas, à un centre de profit et fournit des données sur la valeur de la production, et sur le coût des intrants et de la main-d'œuvre. Cela donne suffisamment de données pour calculer la valeur ajoutée (bénéfices, traitements et salaires). Dans le présent article, nous ne prenons en considération que ces deux niveaux (c.-à-d. entreprise et établissement), en nous fondant sur leurs définitions canadiennes. Pour plus de renseignements, le lecteur pourra consulter [9].

Pour le tirage de l'échantillon, la stratification est souvent effectuée au niveau de l'établissement. Cela permet de contrôler la représentativité géographique (p. ex., en stratifiant par province), la représentativité industrielle (p. ex., en stratifiant par activité industrielle) et la représentativité de taille (p. ex. en stratifiant par catégorie de revenu ou par nombre d'employés). Le contrôle de ces types de représentativité est impossible si la stratification est effectuée au niveau de l'entreprise. Cependant, outre les statistiques au niveau de l'établissement, il est souvent nécessaire de produire des statistiques au niveau de l'entreprise. Par conséquent, pour réaliser ces deux objectifs, nous sélectionnons un échantillon au niveau de l'établissement, puis nous l'étendons à l'ensemble des établissements appartenant aux entreprises propriétaires de ceux sélectionnés au départ. Soulignons que la sélection des entreprises par la voie de la sélection des établissements peut être associée à la sélection de grappes par la voie de leurs composantes. Cela permet de produire des estimations au niveau de l'établissement ainsi que de l'entreprise, et de réduire dans une certaine mesure les coûts de collecte en sélectionnant des groupes d'établissements, plutôt qu'un ensemble d'établissements non reliés.

Une façon d'envisager la production d'estimations au niveau de l'entreprise en se servant d'un échantillon d'établissements consiste à visualiser séparément la base de sondage et la population cible. La première est un ensemble d'établissements, tandis que la seconde est un ensemble d'entreprises correspondant à des groupes d'établissements. Lorsque la base de sondage ne coïncide pas avec la population cible, mais qu'elle est quand même reliée à cette dernière, nous nous trouvons dans une situation de *sondage indirect* (voir [4] et [5]). Plus formellement, nous souhaitons produire une estimation pour une population cible U^B, mais nous ne disposons que d'une base de sondage U^A, qui est reliée d'une certaine façon à U^B. Nous tirons alors un échantillon s^A de U^A pour produire une estimation pour U^B en utilisant les liens existants entre les deux populations. Afin de produire des estimations sans biais des quantités d'intérêt (p. ex., totaux ou moyennes) pour la population cible U^B en utilisant s^A, nous obtenons des poids d'estimation par la *méthode généralisée du partage des poids* (MGPP).

Quoique la théorie de la MGPP soit bien développée (voir [4] et [5]), son application aux enquêtes économiques pose certaines difficultés. En effet, bien qu'elle produise des estimations sans biais, la MGPP a tendance à donner lieu à de fortes variances. Ce manque de précision est dû à l'asymétrie de la population, un petit nombre d'établissements couvrant la majeure partie de l'économie.

Le but du présent article est de proposer certaines corrections des poids d'estimation en vue de réduire la variance des estimations dans le contexte des populations asymétriques, tout en préservant l'absence de biais de la méthode. Après un bref aperçu du sondage indirect et de la MGPP, nous décrirons les corrections qu'il est proposé d'apporter à cette dernière. Nous comparerons ensuite les estimations produites moyennant ces corrections à celles obtenues en appliquant la MGPP originale en nous servant d'un petit exemple numérique et de données réelles provenant du Registre des entreprises de Statistique Canada.

1. Le sondage indirect et la MGPP

À la présente section, nous donnons un aperçu du sondage indirect et de la MGPP. Le sondage indirect a été élaboré pour tout type de plan de sondage, mais nous nous concentrerons ici sur le sondage aléatoire simple sans remise (SASSR) puisque ce plan de sondage est celui utilisé le plus souvent pour les enquêtes économiques.

Soit la population U^A de M^A établissements stratifiée en H strates, où la strate h contient M_h^A établissements. Dans chaque strate h, nous tirons un échantillon s_h^A de m_h^A établissements par SASSR. Soit $s^A = \bigcup_{h=1}^{H} s_h^A$ et $m^A = \sum_{h=1}^{H} m_h^A$. La population cible U^B contient N^B entreprises, où l'entreprise i contient les M_i^B établissements de U^A. Cette population peut alors être considérée comme une population de M^B établissements, où chaque établissement k appartient à une entreprise i, avec $M^B = \sum_{i=1}^{N^B} M_i^B$.

Nous souhaitons produire une estimation pour la population cible U^B en utilisant la base de sondage U^A, ainsi que les liens qui existent entre les deux populations. Les liens entre la population U^A et la population U^B sont identifiés par la variable indicatrice $I_{j,i}$, où $I_{j,i} = 1$ s'il existe un lien entre l'établissement $j \in U^A$ et l'entreprise $i \in U^B$, et 0 autrement. Ici, $I_{j,i} = 1$ si l'établissement j de U^A appartient à l'entreprise i de U^B, et 0 autrement. Comme chaque établissement ne peut appartenir qu'à une seule entreprise, les liens entre U^A et U^B sont de type plusieurs à un, ou un à un (surjectivité). Par conséquent, nous avons $L_j^A = \sum_{i=1}^{N^B} I_{j,i} = 1$, $L_i^B = \sum_{j=1}^{M^A} I_{j,i} = M_i^B$, pour tous les établissements $j \in U^A$ et pour toutes les entreprises $i \in U^B$.

Conformément à la méthode de sondage indirecte, pour chaque établissement j sélectionné dans s^A, nous identifions l'entreprise correspondante i de U^B. Pour chaque entreprise i identifiée, nous supposons que nous pouvons dresser la liste U_i^B des M_i^B établissements de cette entreprise. Nous réalisons ensuite l'enquête auprès de **chacun des** M_i^B établissements de cette entreprise. À la fin, nous obtenons un échantillon s^B de n^B entreprises, et cet échantillon contient $m^B = \sum_{i=1}^{n^B} M_i^B$ établissements.

Pour chacun des établissements k des entreprises $i \in s^B$, nous mesurons une variable d'intérêt y_{ik}. Nous voulons alors estimer le total $Y = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{N^B} Y_i$ pour la population cible U^B. Selon la MGPP, pour estimer le total Y, nous utilisons l'estimateur

$$\hat{Y} = \sum_{i=1}^{n^B} w_i Y_i \tag{1}$$

où n^B est le nombre d'entreprises sondées. Les poids produits par la MGPP sont donnés par

$$w_i = \sum_{j=1}^{M^A} \frac{t_j^A}{\pi_j^A} \frac{I_{j,i}}{L_i^B} \tag{2}$$

où $t_j^A = 1$ si $j \in s^A$, 0 autrement, et π_j^A est la probabilité de sélection de l'établissement *j*. Dans le cas qui nous occupe, nous avons $\pi_j^A = m_h^A / M_h^A$ pour $j \in h$. Il convient de souligner qu'en général, les poids (2) ne correspondent pas aux probabilités de sélection π_i^B des entreprises *i*. En utilisant (2), nous pouvons récrire l'estimateur (1) sous la forme

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \tag{3}$$

où

$$Z_j = \sum_{i=1}^{N^B} \frac{Y_i}{L_i^B} I_{j,i} \ . \tag{4}$$

Étant donné la correspondance surjective entre U^A et U^B, nous avons

$$w_i = \frac{1}{M_i^B} \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A} \tag{5}$$

En outre, la variable Z_j de (4) peut s'écrire $Z_j = Y_i / M_i^B = \bar{Y}_i$, pour $j \in i$, ce qui est la moyenne des M_i^B établissements appartenant à l'entreprise *i*. Nous obtenons donc

$$\hat{Y} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj} \tag{6}$$

où $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, pour $j \in i$.

On peut prouver que l'estimateur (1), et par conséquent (3) et (6), est sans biais pour *Y* (voir [4] et [5]). Notons que l'estimateur $\hat{Y}$ est en fait simplement un estimateur de Horvitz-Thompson où la variable d'intérêt est la variable Z_{hj}. Dans le cas du SASSR stratifié, sa variance est donnée par

$$Var(\hat{Y}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{Z,h}^2 \tag{7}$$

où $S_{Z,h}^2 = \frac{1}{M_h^A - 1} \sum_{j=1}^{M_h^A} \left(Z_{hj} - \bar{Z}_h \right)^2$ et $\bar{Z}_h = \frac{1}{M_h^A} \sum_{j=1}^{M_h^A} Z_{hj}$.

La variance $Var(\hat{Y})$ peut être estimée en se servant de l'estimateur classique pour le SASSR stratifié, ou au moyen d'autres estimateurs de la variance proposés dans la littérature scientifique, tels que les estimateurs par le jackknife et le bootstrap. Voir [11] ou [8].

1.1. Utilisation de liens pondérés

La variable indicatrice $I_{j,i}$ indique simplement s'il existe ou non un lien entre l'établissement *j* et l'entreprise *i* provenant des populations U^A et U^B, respectivement. Il est cependant possible de remplacer la variable indicatrice $I_{j,i}$ par une variable quantitative $\theta_{j,i}$ représentant l'importance que nous voulons accorder au lien $I_{j,i}$. Autrement dit, la généralisation de la variable indicatrice *I* définie sur {0,1} au moyen d'une variable quantitative θ définie sur $[0,+\infty[$, l'ensemble de nombres réels non négatifs, ne pose aucun problème. Ici, une valeur de $\theta_{j,i} = 0$ équivaut à un lien $I_{j,i} = 0$. La théorie articulée sur la MGPP demeure valide. Par exemple, l'estimateur résultant demeure sans biais. Comme nous le verrons plus loin, choisir les valeurs appropriées pour les liens pondérés $\theta_{j,i}$ sera le fondement des méthodes visant à réduire la variance des estimations obtenues en appliquant la MGPP.

Soit $\tilde{\theta}_{j,i} = \theta_{j,i} / \theta_i^B$ où $\theta_i^B = \sum_{j=1}^{M^A} \theta_{j,i}$. Partant de (2), nous définissons

$$w_i^\theta = \frac{1}{\theta_i^B} \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A} \theta_{j,i} \tag{8}$$

En utilisant (8), nous pouvons récrire l'estimateur (6) sous la forme

$$\hat{Y}_\theta = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^\theta \tag{9}$$

où

$$Z_{hj}^\theta = \sum_{i=1}^{N^B} \frac{Y_i}{\theta_i^B} \theta_{j,i} \tag{10}$$

pour $j \in h$. Étant donné la correspondance surjective entre U^A et U^B, la variable Z_{hj}^θ dans (10) est une portion pondérée du total Y_i des M_i^B établissements appartenant à l'entreprise *i*. La variance de (9) s'obtient en remplaçant Z_j par Z_j^θ dans (7) :

$$Var(\hat{Y}_\theta) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{\theta Zh}^2 \tag{11}$$

où $S_{\theta Zh}^2 = \frac{1}{M_h^A - 1} \sum_{j=1}^{M_h^A} \left(Z_{hj}^\theta - \bar{Z}_h^\theta \right)^2$ et $\bar{Z}_h^\theta = \frac{1}{M_h^A} \sum_{j=1}^{M_h^A} Z_{hj}^\theta$.

1.2. Utilisation de liens pondérés optimaux

La MGPP offre une solution simple pour obtenir un poids d'estimation w_i pour chaque entreprise *i* sondée. Cependant, l'estimateur $\hat{Y}$ donné par (1) ou (3) résultant de l'application par défaut de la MGPP n'est pas toujours celui possédant la variance la plus faible. Il est possible de l'améliorer en déterminant les poids optimaux pour les liens $\theta_{j,i}$. Ce problème a été résolu par Deville et Lavallée (voir [2]).

Dans [2], Deville et Lavallée ont obtenu un estimateur dont la variance est inférieure ou égale à celle de l'estimateur original $\hat{Y}$. Comme nous l'avons mentionné plus haut, l'estimateur $\hat{Y}_\theta$ donné par (9) produit encore des estimations sans biais. Or, la variance (11) de cet estimateur dépend des liens pondérés $\theta_{j,i}$. Le problème consiste alors à trouver au moins un jeu de valeurs $\theta_{j,i}$ tel que la variance de l'estimateur $\hat{Y}_\theta$ soit minimale. Autrement dit, pour les $\theta_{j,i}$ qui sont plus grands que zéro, nous voulons trouver les valeurs que ces $\theta_{j,i}$ devraient avoir pour obtenir l'estimateur $\hat{Y}$ le plus précis possible. La solution de ce problème s'obtient en minimisant la variance (11) par rapport aux liens pondérés $\theta_{j,i}$, ce qui est un problème relativement standard et simple à résoudre. Cependant, la solution n'est pas simple à écrire, et elle dépend souvent de la variable d'intérêt *y*.

Si les liens pondérés optimaux $\theta_{j,i}^{opt}$ dépendent de la variable d'intérêt *y*, les poids w_i^θ dépendront également de *y*. Cela signifie qu'un jeu différent de poids devra être calculé pour chaque variable d'intérêt. Pour contourner ce problème, Deville et Lavallée ont défini une *optimalité faible*, qui correspond à la minimisation de la variance (11) pour un choix très précis d'une variable d'intérêt : $Y_i = 1$ pour une entreprise *i* de U^B et $Y_{i'} = 0$ pour toutes les autres entreprises i' de U^B ($i' \neq i$). Les liens pondérés faiblement optimaux résultants ne font pas intervenir, à proprement parler, la variable *y* et ils s'avèrent relativement faciles à calculer, c'est-à-dire qu'ils peuvent être obtenus sous forme d'une solution explicite, sans que des calculs numériques soient nécessaires. En outre, si certaines conditions énoncées par Deville et Lavallée dans [2] sont satisfaites, l'optimalité faible correspond à une *optimalité forte indépendante de y*. Autrement dit, les liens pondérés $\theta_{j,i}^{f-opt}$ obtenus par optimalité faible correspondent aux liens pondérés optimaux $\theta_{j,i}^{opt}$ obtenus en minimisant (11), et ils ne

dépendent pas de la variable d'intérêt *y*. Malheureusement, ces conditions sont rarement satisfaites en pratique, même pour des plans de sondage simples tels que le SASSR.

Sous SASSR **sans** stratification, on peut montrer que les liens pondérés faiblement optimaux sont donnés par $\tilde{\theta}_{j,i}^{f-opt} = 1/M_i^B$ pour l'établissement $j \in U^A$ appartenant à l'entreprise $i \in U^B$, et 0 sinon. Cette solution va dans le sens de celle conjecturée par Kalton et Brick dans [3]. Ils ont obtenu ce résultat en se fondant sur la situation simplifiée où M^A=2 et avec s^A sélectionné par sondage équiprobabiliste. Dans leurs conclusions, ils suggéraient d'utiliser les valeurs optimales $\theta_{j,i}^{opt} = 1$ quand $\theta_{j,i} > 0$, et $\theta_{j,i}^{opt} = 0$ quand $\theta_{j,i} = 0$. Lavallée (dans [4]), et Lavallée et Caron (dans [6]) ont obtenu des résultats allant dans le même sens en utilisant des simulations. Comme nous l'avons mentionné plus haut, malheureusement, les poids faiblement optimaux $\tilde{\theta}_{j,i}^{f-opt} = 1/M_i^B$ ne correspondent pas aux poids fortement optimaux indépendants de *y*.

2. Le problème des populations asymétriques

Comme il est mentionné dans l'introduction, l'application de la MGPP aux enquêtes économiques peut produire des estimations dont la variance est importante. Ce manque de précision est dû à l'asymétrie de la population. Nous proposons d'illustrer le problème à l'aide d'un petit exemple donné à la figure 1.

Nous voulons étudier le revenu *y* de la population U^B de la figure 1 contenant $N^B = 3$ entreprises, où l'entreprise 1 contient $M_1^B = 4$ établissements, l'entreprise 2 contient $M_2^B = 4$ établissements, et l'entreprise 3, $M_3^B = 3$ établissements. Comme on peut le voir, le revenu *y* des $M^B = 11$ établissements peut être considéré comme une population asymétrique.

Pour l'enquête, nous construisons une base de sondage U^A contenant les $M^A = 11$ établissements et nous décidons de répartir les établissements en trois strates de taille, à savoir la strate *h*=1 contenant les établissements pour lesquels $y \geq 750$, la strate *h*=2 contenant ceux pour lesquels $100 \leq y < 750$ et la strate *h*=3 contenant ceux pour lesquels $y < 100$ [1]. Dans la strate *h*=1, nous utilisons une fraction de sondage de 1 (c.-à-d. $f_1 = m_1^A / M_1^A = 1$); pour *h*=2, la taille d'échantillon est de 1 (c.-à-d. $f_2 = m_2^A / M_2^A = 1/3$), et pour *h*=3, la taille d'échantillon est de 2 (c.-à-d. $f_3 = m_3^A / M_3^A = 2/6 = 1/3$).

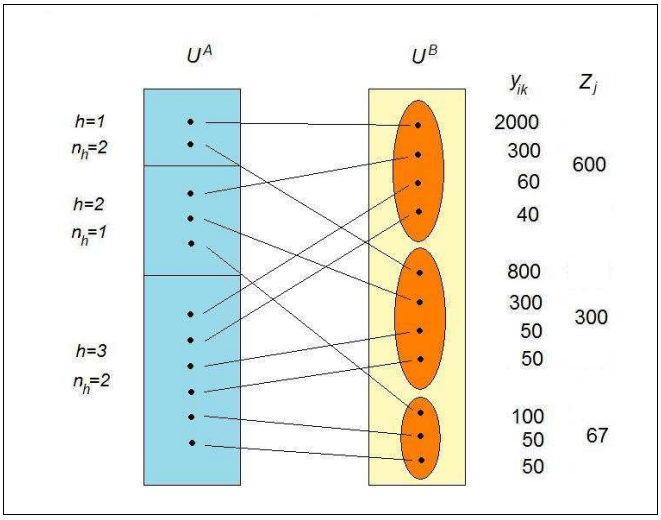


Figure 1. Petit exemple

[1] En pratique, une telle stratification est impossible puisque la variable de stratification *y* est la même que la variable d'intérêt. Nous pourrions alors utiliser une variable de taille *x* fortement corrélée à la variable d'intérêt *y*.

Il existe $1\times 3\times 15 = 45$ échantillons possibles s^A qui peuvent être tirés de U^A afin de produire une estimation du vrai total $Y = 3\,800$. Pour chacun de ces 45 échantillons, nous calculons $\hat{Y}$ en utilisant (1), ou (3). Les résultats sont présentés en annexe.

Nous calculons aussi les estimations de Y en supposant que l'on utilise un SASSR stratifié **sans** sondage indirect. Autrement dit, dans chaque strate *h*, nous sélectionnons un échantillon s_h^A de m_h^A établissements par SASSR et nous mesurons la variable d'intérêt y_{ik} uniquement pour les établissements *ik* de U^B directement liés aux établissements échantillonnés *j* de U^A. En d'autres termes, nous mesurons la variable d'intérêt y_j pour les établissements échantillonnés *j* de U^A. À l'opposé du sondage indirect, nous ne mesurons pas les variables d'intérêt pour les autres établissements des entreprises contenant les établissements échantillonnés au départ. Cela correspond à la théorie classique de l'échantillonnage. Donc, nous avons estimé Y en utilisant

$$\hat{Y}_{classique} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} y_{hj} \tag{12}$$

On peut prouver que l'estimateur (12) est sans biais et que sa variance est donnée par

$$Var(\hat{Y}_{classique}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{y,h}^2 \tag{13}$$

où $S_{y,h}^2 = \frac{1}{M_h^A - 1} \sum_{j=1}^{M_h^A} \left(y_{hj} - \bar{Y}_h \right)^2$ et $\bar{Y}_h = \frac{1}{M_h^A} \sum_{j=1}^{M_h^A} y_{hj}$. Les résultats sont présentés en annexe.

L'examen du tableau en annexe montre que les estimations produites par sondage indirect varient assez bien d'un échantillon à l'autre. Sans sondage indirect (c.-à-d. en utilisant l'approche classique), la variabilité est beaucoup plus faible. Ce résultat peut être observé directement d'après les variances de $\hat{Y}$ et $\hat{Y}_{classique}$. En utilisant les formules (7) et (13), nous obtenons la variance $V(\hat{Y}_{classique}) = 80\,480$, tandis que $V(\hat{Y}) = 1115\,111$!

À la section suivante, nous présentons les méthodes en vue de réduire la variabilité des estimations produites sous sondage indirect.

3. Méthodes proposées

Les méthodes proposées ici pour réduire la variance des estimations sont fondées principalement sur l'emploi de liens pondérés pour le calcul des estimations de Y sous sondage indirect. Nous utiliserons par conséquent l'estimateur (9) plutôt que l'estimateur (3). Un premier ensemble de méthodes est fondé sur l'emploi des liens pondérés $\theta_{j,i}$ qui sont proportionnels à une certaine mesure de taille des établissements. Le deuxième ensemble de méthodes correspond à l'utilisation des solutions optimales obtenues à la section 2.2, sous diverses hypothèses. Finalement, le dernier ensemble de méthodes correspond à l'utilisation des probabilités de sélection exactes plutôt que les poids d'estimation obtenus par la MGPP, sous deux scénarios d'échantillonnage.

3.1. Méthodes fondées sur l'emploi de liens pondérés

Méthode 1 : $\theta_{j,i}$ proportionnel à π_j^A

Nous proposons d'abord de réduire la variance (11) en prenant $\theta_{j,i}$ proportionnel à π_j^A. Formellement, cela peut s'écrire $\theta_{j,i}^{\pi} = \pi_j^A I_{j,i}$. Dans les enquêtes économiques, comme la stratification est habituellement effectuée selon la taille (en s'appuyant sur une certaine mesure de taille), prendre $\theta_{j,i}$

proportionnel à π_j^A revient à attribuer des poids élevés aux liens des grands établissements et des poids faibles à ceux des petits établissements, ce qui est une approche naturelle.

Selon cette méthode, nous avons $\tilde{\theta}_{j,i}^{\pi} = \theta_{j,i}^{\pi} \big/ \theta_i^{\pi B} = \pi_j^A I_{j,i} \,/\, \sum_{j=1}^{M^A} \pi_j^A I_{j,i}$. Étant donné la correspondance surjective entre U^A et U^B, nous obtenons $\tilde{\theta}_{j,i}^{\pi} = \theta_{j,i}^{\pi} \big/ \theta_i^{\pi B} = \pi_j^A I_{j,i} \Big/ \sum_{j=1}^{M_i^B} \pi_j^A$. Donc, partant de (8), nous avons

$$w_j^{\pi} = \frac{\sum_{j=1}^{M_i^B} t_j^A}{\sum_{j=1}^{M_i^B} \pi_j^A} \quad . \tag{14}$$

En utilisant (14), nous pouvons récrire l'estimateur (9) sous la forme

$$\hat{Y}_{\pi} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{\pi} \tag{15}$$

où

$$Z_{hj}^{\pi} = \frac{\pi_j^A Y_i}{\sum_{j=1}^{M_i^B} \pi_j^A} \tag{16}$$

pour $j \in h$ et $j \in i$. Il convient de souligner que si tous les établissements *j* d'une entreprise donnée appartiennent à la même strate *h*, nous avons $\tilde{\theta}_{j,i} = 1/M_i^B$, et l'estimateur (15) est alors équivalent à l'estimateur (1), et (3).

Pour calculer la variance de $\hat{Y}_{\pi}$, nous utilisons la formule (11) avec les valeurs (16). Pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_{\pi}) = 439\,111$, valeur qui représente une réduction importante comparativement à $V(\hat{Y}) = 1115\,111$, mais qui est encore relativement éloignée de $V(\hat{Y}_{classique}) = 80\,480$.

Méthode 2 : $\theta_{j,i}$ proportionnel à une mesure de taille x_j

Nous proposons de réduire la variance (11) en prenant $\theta_{j,i}$ proportionnel à une mesure de taille *x* corrélée à la variable d'intérêt *y*. Nous supposons que la variable x_j est disponible pour tous les établissements $j \in U^A$. Cette variable pourrait être utilisée, par exemple, pour stratifier la base de sondage U^A selon la taille. Comme dans la méthode 1, prendre $\theta_{j,i}^{x} = x_j I_{j,i}$ peut être considéré comme attribuer des poids élevés aux liens des grands établissements et des poids faibles à ceux des petits, ce qui de nouveau est une approche naturelle. Selon cette méthode, nous avons $\tilde{\theta}_{j,i}^{x} = \theta_{j,i}^{x} \big/ \theta_i^{xB} = x_j I_{j,i} \Big/ \sum_{j=1}^{M^A} x_j I_{j,i}$.

En raison de la correspondance surjective entre U^A et U^B, nous avons $\tilde{\theta}_{j,i}^{x} = \theta_{j,i}^{x} \big/ \theta_i^{xB} = x_j I_{j,i} \Big/ \sum_{j=1}^{M_i^B} x_j = x_j I_{j,i} \big/ X_i$. Par conséquent, partant de (8), nous obtenons

$$w_i^{x} = \frac{1}{X_i} \sum_{j=1}^{M_i^B} \frac{t_j^A x_j}{\pi_j^A} \quad . \tag{17}$$

En utilisant (17), nous pouvons récrire (9) sous la forme

$$\hat{Y}_x = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{x} \tag{18}$$

où

$$Z_{hj}^{x} = \frac{Y_i}{X_i} x_j \qquad (19)$$

pour $j \in h$ et $j \in i$. Il convient de souligner que, si tous les établissements *j* d'une entreprise donnée appartiennent à la même strate *h*, nous avons de nouveau $\tilde{\theta}_{j,ij} = 1/M_i^B$, et l'estimateur (18) est alors équivalent à l'estimateur (1), et (3).

Pour calculer la variance de $\hat{Y}_x$, nous utilisons la formule (11) avec les valeurs (19). Pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_x) = 686\,540$, valeur qui représente de nouveau une réduction importante comparativement à $V(\hat{Y}) = 1115\,111$, mais qui est encore relativement éloignée de $V(\hat{Y}_{classique}) = 80\,480$.

Méthode 3 : $\theta_{j,i}$ proportionnel à la variable d'intérêt y_j

La troisième méthode proposée consiste à réduire la variance (11) en prenant $\theta_{j,i}$ proportionnel à la variable d'intérêt *y* mesurée pour l'établissement *j* appartenant à l'entreprise *i*. Manifestement, prendre $\theta_{j,i}^{y} = y_j I_{j,i}$ attribue des poids élevés aux liens des grands établissements et des poids faibles à ceux des petits, ce qui de nouveau est une approche naturelle. Comme y_j est inconnue au début de l'enquête, cette méthode pourrait paraître impossible à appliquer puisque $\theta_{j,i}^{y}$ dépend de y_j. Or, puisque la correspondance entre U^A et U^B est surjective, chaque quantité entrant dans $\theta_{j,i}^{y}$ est mesurée selon le processus de sondage indirect. Par conséquent, la méthode est applicable en pratique.

Selon cette méthode, nous avons $\tilde{\theta}_{j,i}^{y} = \theta_{j,i}^{y} / \theta_i^{yB} = y_j I_{j,i} / \sum_{j=1}^{M_i^B} y_j = y_j I_{j,i} / Y_i$. Les poids w_i^y sont donnés directement par (17) en remplaçant *x* par *y*. L'estimateur $\hat{Y}_y$ donné par (9) se réduit à

$$\hat{Y}_y = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} y_{hj} \ , \qquad (20)$$

ce qui n'est autre que l'estimateur (12) découlant de la théorie classique de l'échantillonnage.

Pour calculer la variance de (20), nous employons simplement la formule (13). Naturellement, pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_y) = V(\hat{Y}_{classique}) = 80\,480$, ce qui représente une réduction très importante comparativement à $V(\hat{Y}) = 1115\,111$.

3.2. Méthodes utilisant des liens pondérés faiblement optimaux

Méthode 4 : Utilisation de liens pondérés faiblement optimaux $\theta_{j,i}^{f-opt,SAS}$ sous SASSR stratifié

Cette méthode consiste à utiliser les liens pondérés faiblement optimaux $\theta_{j,i}^{f-opt,SAS}$ de Deville et Lavallée décrits à la section 2.2. Comme nous l'avons mentionné plus haut, ceux-ci sont obtenus en minimisant la variance (11) pour un choix très précis de la variable d'intérêt, à savoir $Y_i = 1$ pour une entreprise *i* de U^B et $Y_{i'} = 0$ pour toutes les autres entreprises i' de U^B ($i' \neq i$). Les liens pondérés faiblement optimaux résultants ne font pas intervenir la variable *y* à proprement parler. Écrire les valeurs de $\theta_{j,i}^{f-opt,SAS}$ comprend des expressions qui peuvent être exprimées relativement facilement en

notation matricielle. En utilisant des sommations, les expressions deviennent beaucoup plus compliquées, principalement parce qu'elles font intervenir un mélange de probabilités de sélection conjointes $\pi^A_{jj'}$ des établissements j et j' qui peuvent appartenir ou non à la même strate.

Définissons la matrice carrée $\mathbf{\Delta}^A = \left[\Delta^A_{j,j'}\right]$ de taille M^A où $\Delta^A_{j,j'} = (\pi^A_{jj'} - \pi^A_j \pi^A_{j'}) / \pi^A_j \pi^A_{j'}$. Soit $\mathbf{\Gamma}^A = \left[\gamma^A_{j,j'}\right]$ l'inverse de la matrice $\mathbf{\Delta}^A$, c.-à-d. $(\mathbf{\Delta}^A)^{-1} = \mathbf{\Gamma}^A$. Soit $\mathbf{\Gamma}^A_i = \left[\gamma_{ij,ij'}\right]$ la sous-matrice carrée de $\mathbf{\Gamma}^A$ contenant tous les éléments (établissements) (j, j') appartenant à l'entreprise *i*. En suivant [2], nous obtenons $\tilde{\theta}^{f-opt,SAS}_{j,i} = I_{j,i} \sum_{j'=1}^{M^B_i} \gamma_{ij,ij'} / \sum_{j=1}^{M^B_i} \sum_{j'=1}^{M^B_i} \gamma_{ij,ij'}$. Malheureusement, ici, la correspondance surjective entre U^A et U^B n'aide pas à obtenir une forme plus simple de $\tilde{\theta}^{f-opt,SAS}_{j,i}$.

Nous avons

$$w^{f-opt,SAS}_i = \sum_{j=1}^{M^B_i} \frac{t^A_j}{\pi^A_j} \tilde{\theta}^{f-opt,SAS}_{j,i} \quad . \tag{21}$$

En utilisant (21), nous pouvons récrire l'estimateur (9) sous la forme

$$\hat{Y}_{f-opt,SAS} = \sum_{h=1}^{H} \frac{M^A_h}{m^A_h} \sum_{j=1}^{m^A_h} Z^{f-opt,SAS}_{hj} \tag{22}$$

où

$$Z^{f-opt,SAS}_{hj} = \sum_{i=1}^{N^B} Y_i \tilde{\theta}^{f-opt,SAS}_{j,i} \tag{23}$$

pour $j \in h$ et $j \in i$.

Pour calculer la variance de $\hat{Y}_{f-opt,SAS}$, nous utilisons la formule (11) avec les valeurs (23). Pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_{f-opt,SAS}) = 23\,111$, ce qui correspond à une réduction énorme de la variance comparativement à $V(\hat{Y}) = 1115\,111$, ainsi que $V(\hat{Y}_{classique}) = 80\,480$.

Méthode 5 : Utilisation de liens pondérés faiblement optimaux $\theta^{f-opt,SP}_{j,i}$ sous sondage de Poisson

Dans le contexte des enquêtes économiques, le sondage de Poisson consiste à tirer l'échantillon s^A en passant un à un les M^A établissements de la population U^A et en sélectionnant l'établissement *j* si $u_j \leq \pi^A_j$, où $u_j \sim U(0,1)$. Les probabilités de sélection sont simplement données par $\pi^A_j = m^A_h / M^A_h$ pour $j \in h$. Dans ce contexte, ce plan de sondage peut également être considéré comme un sondage de Bernoulli stratifié (voir [8]).

Le sondage de Poisson (ou le sondage de Bernoulli stratifié) est un plan de sondage très simple. Comme on peut le constater, la sélection de chaque établissement de s^A est faite indépendamment d'un établissement à l'autre. Cela signifie que la probabilité de sélection conjointe $\pi^A_{jj'}$ de deux établissements différents j et j' de U^A est donnée simplement par $\pi^A_{jj'} = \pi^A_j \pi^A_{j'}$. Bien que les tailles d'échantillon de strate réalisées $\tilde{m}^A_h$ soient aléatoires, en conditionnant sur ces $\tilde{m}^A_h$, on peut montrer que le sondage de Bernoulli stratifié correspond au SASSR stratifié. Étant donné la proximité relative de ces deux plans, supposer que l'on procède à un sondage de Poisson peut être une approche raisonnable pour calculer les liens pondérés faiblement optimaux $\theta^{f-opt,SP}_{j,i}$.

Les liens pondérés faiblement optimaux $\theta^{f-opt,SP}_{j,i}$ s'obtiennent en calculant $\tilde{\theta}^{f-opt,SAS}_{j,i}$ comme dans la méthode 4, mais en supposant que cette sélection d'échantillon est faite sous sondage de Poisson.

Cette hypothèse permet de simplifier considérablement les calculs parce que la matrice $\mathbf{\Delta}^A$ devient alors une matrice diagonale, facile à inverser. Notons que, sous sondage de Poisson, la variance de (9) est donnée par

$$Var_{SP}(\hat{Y}^{\theta}) = \sum_{h=1}^{H}\left(\frac{M_h^A - m_h^A}{m_h^A}\right)\sum_{j=1}^{M_h^A}\left(Z_{hj}^{\theta}\right)^2 , \tag{24}$$

forme qui ressemble à la variance (11).

En raison de la correspondance surjective entre U^A et U^B, après le processus de minimisation, nous obtenons

$$\tilde{\theta}_{j,i}^{f-opt,SP} = \frac{\pi_j^A I_{j,i}}{(1-\pi_j^A)\tau_i} \tag{25}$$

où $\tau_i = \sum_{j=1}^{M_i^B}\frac{\pi_j^A}{(1-\pi_j^A)}$. Par conséquent, partant de (8), nous obtenons

$$w_i^{f-opt,SP} = \frac{1}{\tau_i}\sum_{j=1}^{M_i^B}\frac{t_j^A}{(1-\pi_j^A)} \ . \tag{26}$$

En utilisant (26), nous pouvons récrire (9) sous la forme

$$\hat{Y}_{f-opt,SP} = \sum_{h=1}^{H}\frac{M_h^A}{m_h^A}\sum_{j=1}^{m_h^A} Z_{hj}^{f-opt,SP} \tag{27}$$

où

$$Z_{hj}^{f-opt,SP} = \frac{\pi_j^A}{(1-\pi_j^A)}\frac{Y_i}{\tau_i} \tag{28}$$

pour $j \in h$ et $j \in i$. Notons que les résultats précédents reposent sur l'hypothèse que $0 < \pi_j^A < 1$ pour tous les établissements *j* de U^A. Pour le cas où $\pi_j^A = 1$ pour un établissement donné j_0, nous prenons $\tilde{\theta}_{j_0,i}^{f-opt,SP} = 1$, et $\tilde{\theta}_{j',i}^{f-opt,SP} = 0$ pour $j' \neq j_0$. Pour calculer la variance de $\hat{Y}_{f-opt,SP}$, nous utilisons la formule (11) avec les valeurs (28).

Pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_{f-opt,SP}) = 22\,857$. De nouveau, ceci correspond à une réduction énorme de la variance comparativement à $V(\hat{Y}) = 1115\,111$, ainsi que $V(\hat{Y}_{classique}) = 80\,480$.

Méthode 6 : Utilisation de liens pondérés faiblement optimaux $\theta_{j,i}^{f-opt,grp}$ sous sondage poissonnien de groupes d'établissements

Cette méthode consiste de nouveau à utiliser les liens pondérés faiblement optimaux de Deville et Lavallée décrits à la section 2.2, mais pour des groupes d'établissements. Pour commencer, nous construisons des groupes d'établissements dans la population U^A où un groupe d'établissements *j** comprend tous les établissements qui font partie de la même strate *h* et qui appartiennent à la même entreprise *i*. Cela crée une nouvelle population U^{A^*} contenant M^{A^*} groupes d'établissements. L'échantillon s^{A^*} de m^{A^*} groupes d'établissements contient tous les groupes d'établissements formés au moyen des établissements de l'échantillon s^A. La probabilité de sélection du groupe d'établissements *j** est donnée par

$$\pi_{j^*}^{A^*} = 1 - \frac{\binom{M_h^A - M_{j^*}}{m_h^A}}{\binom{M_h^A}{m_h^A}} = 1 - \frac{(M_h^A - M_{j^*})(M_h^A - M_{j^*} - 1)\ldots(M_h^A - M_{j^*} - m_h^A + 1)}{M_h^A(M_h^A - 1)\ldots(M_h^A - m_h^A + 1)} \tag{29}$$

pour $j^* \in h$, où M_{j^*} est le nombre d'établissements dans le groupe d'établissements j^*.

La logique de l'utilisation de groupes d'établissements est de n'avoir qu'une seule unité appartenant à une entreprise donnée par strate. Comme, par construction, les établissements du groupe j^* d'une entreprise *i* appartiennent à différentes strates, leur sélection est effectuée indépendamment d'un groupe d'établissements à l'autre. Cela implique que la solution de l'optimalité faible est alors similaire à celle obtenue à la section 4.5 pour le sondage de Poisson, mais avec des groupes d'établissements. Par conséquent, nous avons

$$\tilde{\theta}_{j^*,i}^{f-opt,grp} = \frac{\pi_{j^*}^A I_{j^*,i}}{(1-\pi_{j^*}^A)\tau_i^*} \tag{30}$$

où $\tau_i^* = \sum_{j^*=1}^{M_i^{B^*}} \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})}$ et $M_i^{B^*}$ est le nombre de groupes d'établissements contenus dans l'entreprise *i.*

L'utilisation de groupes d'établissements peut être considérée comme une étape intermédiaire dans le processus de sondage indirect allant de la population U^A à la population U^B. Autrement dit, le processus de sondage indirect va de la population U^A à la population U^{A^*}, puis de la population U^{A^*} à la population U^B, ce qui peut être considéré comme un processus transitif. Ici, nous avons $j \in j^* \in i$ pour tous les établissements. En suivant les règles de transitivité définies par Deville et Lavallée dans [2], nous pouvons montrer que les liens pondérés faiblement optimaux $\tilde{\theta}_{j,i}^{w-opt,grp}$ pour $j \in j^*$ et $j^* \in i$ (et donc, $j \in i$) sont donnés par

$$\tilde{\theta}_{j,i}^{f-opt,grp} = \frac{\pi_{j^*}^A I_{j,i}}{(1-\pi_{j^*}^A)\tau_i^* M_{j^*}} \ . \tag{31}$$

Par conséquent, en partant de (8), nous obtenons

$$w_i^{f-opt,grp} = \frac{1}{\tau_i^*} \sum_{j^*=1}^{M_i^{B^*}} \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})M_{j^*}} \sum_{j=1}^{M_{j^*}} \frac{t_j^A}{\pi_j^A} \tag{32}$$

En utilisant (32), nous pouvons récrire (9) sous la forme

$$\hat{Y}_{f-opt,grp} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{f-opt,grp} \tag{33}$$

où

$$Z_{hj}^{f-opt,grp} = \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})M_{j^*}} \frac{Y_i}{\tau_i^*} \tag{34}$$

pour $j \in h$ et $j^* \in h$. Soulignons que les résultats qui précèdent reposent sur l'hypothèse que $0 < \pi_{j^*}^{A^*} < 1$ pour tous les groupes d'établissements j^* de U^{A^*}. Pour le cas où $\pi_{j^*}^{A^*} = 1$ pour un groupe d'établissements donné j_0^*, nous prenons $\tilde{\theta}_{j,i}^{f-opt,grp} = I_{j,i} / M_{j^*}$ pour tous les établissements *j* de ce groupe d'établissements j_0^*, et $\tilde{\theta}_{j,i}^{f-opt,grp} = 0$ pour tous les autres établissements ne faisant pas partie du groupe d'établissements j_0^*. Nous avons $\pi_{j^*}^{A^*} = 1$ quand au moins un établissement *j* appartenant au groupe j^* a une probabilité de sélection $\pi_j^A = 1$. Pour calculer la variance de $\hat{Y}_{f-opt,grp}$, nous utilisons la formule (11) avec les valeurs (34).

Pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_{f-opt,grp}) = 23\,000$. De nouveau, ceci correspond à une réduction énorme de la variance comparativement à $V(\hat{Y}) = 1115\,111$, ainsi que $V(\hat{Y}_{classique}) = 80\,480$.

3.3. Autres méthodes

Méthode 7 : Utilisation d'un établissement désigné

Comme nous l'avons mentionné plus haut, la logique de l'utilisation de groupes d'établissements dans la méthode 6 est de n'avoir qu'une seule unité appartenant à une entreprise donnée par strate. Dans le même ordre d'idée, on peut choisir un seul établissement qui représentera l'entreprise complète. Autrement dit, pour chaque entreprise de U^B, nous identifions un établissement de U^A qui sera utilisé pour sélectionner l'entreprise qui en est propriétaire. Un choix naturel pour l'*établissement désigné* est celui ayant la plus grande valeur pour une variable donnée *x*. Par exemple, *x* peut être le revenu de l'établissement.

En choisissant un seul établissement désigné, nous obtenons une nouvelle base de sondage U^{A+} qui contient le même nombre d'unités que la population cible U^B, c.-à-d. $M^{A+} = N^B$. Puisqu'il existe une correspondance bijective entre l'établissement désigné et l'entreprise propriétaire, l'établissement désigné de l'entreprise *i* peut également être étiqueté en utilisant *i*. La nouvelle base de sondage U^{A+} peut garder la même définition de stratification que la base de sondage originale U^A. C'est-à-dire que si la stratification de U^A était faite par province et par catégorie d'industrie en se fondant sur les valeurs pour les établissements, la stratification de U^{A+} serait faite selon les mêmes catégories en se fondant sur les valeurs pour les établissements désignés.

Dans la base de sondage U^{A+}, nous sélectionnons un échantillon s^{A+} de m^{A+} établissements désignés sous SASSR en utilisant des fractions de sondage égales aux fractions originales, c.-à-d. $\pi_i^{A+} = m_h^{A+} / M_h^{A+} = m_h^A / M_h^A$, pour $i \in h$. Le total Y est estimé en utilisant l'estimateur qui suit, découlant de la théorie classique :

$$\hat{Y}_+ = \sum_{h=1}^{H} \frac{M_h^{A+}}{m_h^{A+}} \sum_{i=1}^{m_h^{A+}} Y_i \quad . \tag{35}$$

On peut démontrer que l'estimateur (35) est sans biais, et que sa variance est donnée par

$$Var(\hat{Y}_+) = \sum_{h=1}^{H} M_h^{A+} \left(\frac{M_h^{A+} - m_h^{A+}}{m_h^{A+}} \right) S_{+Yh}^2 \tag{36}$$

où $S_{+Yh}^2 = \frac{1}{M_h^{A+} - 1} \sum_{i=1}^{M_h^{A+}} \left(Y_{hi} - \bar{Y}_h \right)^2$ et $\bar{Y}_h = \frac{1}{M_h^{A+}} \sum_{i=1}^{M_h^{A+}} Y_{hi}$.

Pour l'exemple de la section 2, nous obtenons $V(\hat{Y}_+) = 1\,820\,000$! Avec cette méthode, comme un établissement hérite de tous les revenus de l'entreprise, l'utilisation d'un établissement désigné est avantageuse quand ce dernier se trouve dans une strate à tirage complet. Cependant, il arrive aussi que l'établissement désigné se trouve dans une strate à tirage partiel, si bien que la distribution à l'intérieur de cette strate devient encore plus asymétrique. La totalité des revenus de l'entreprise multipliée par le poids de sondage de cette entreprise est attribuée à cette seule strate, ce qui fait augmenter considérablement la variance.

Méthode 8 : Utilisation des probabilités de sélection des entreprises

Comme il est mentionné dans [4] et [5], en utilisant le théorème de Rao-Blackwell, une statistique exhaustive peut être utilisée pour améliorer un estimateur existant en produisant un nouvel estimateur dont l'erreur quadratique moyenne est inférieure ou égale à celle de l'estimateur de départ (voir [1]). Notons que cette forme d'amélioration a été utilisée, entre autres, dans [10] dans le contexte du sondage par grappes adaptatif.

En partant de l'estimateur (1), ou (3), l'estimateur $\hat{Y}_{RB}$ obtenu en utilisant le théorème de Rao-Blackwell est donné par

$$\hat{Y}_{RB} = \sum_{i=1}^{n^B} \frac{Y_i}{L_i^B} \sum_{j=1}^{M^A} \frac{P(t_j^A = 1 | s^B)}{\pi_j^A} I_{j,i}. \tag{37}$$

où $P(t_j^A = 1 | s^B)$ est la probabilité d'avoir sélectionné l'établissement j dans U^A, sachant que les n^B entreprises de s^B ont été sondées. En général, cette probabilité n'est pas facile à calculer.

Étant donné la correspondance surjective entre U^A et U^B, $P(t_j^A = 1 | s^B)$ peut être obtenue comme suit : pour $j \in i$, nous avons

$$\begin{aligned} P(t_j^A = 1 | s^B) &= P(t_j^A = 1 | i \in s^B) \\ &= P(t_j^A = 1, i \in s^B) / P(i \in s^B) \\ &= P(t_j^A = 1) / P(i \in s^B) \\ &= \pi_j^A / \pi_i^B \end{aligned} \tag{38}$$

où π_i^B est la probabilité de sélection de l'entreprise $i \in U^B$, qui correspond à la probabilité de sélectionner n'importe lequel de ses M_i^B établissements. L'estimateur (37) se réduit alors à l'estimateur de Horvitz-Thompson suivant

$$\hat{Y}_{HT} = \sum_{i=1}^{n^B} \frac{Y_i}{\pi_i^B} \ . \tag{39}$$

Comme la base de sondage complète U^A est disponible pour la sélection des établissements, il est possible de calculer les probabilités de sélection π_i^B, mais cela peut être fastidieux.

Puisque l'estimateur (39) n'est autre qu'un estimateur de Horvitz-Thompson fondé sur la sélection des entreprises, sa variance est donnée par

$$V\text{ar}(\hat{Y}_{HT}) = \sum_{i=1}^{N^B} \sum_{i'=1}^{N^B} \frac{(\pi_{ii'}^B - \pi_i^B \pi_{i'}^B)}{\pi_i^B \pi_{i'}^B} Y_i Y_{i'} \ . \tag{40}$$

De nouveau, comme la base de sondage complète U^A est disponible pour la sélection des établissements, il est possible de calculer les probabilités de sélection conjointes $\pi_{ii'}^B$, mais celles-ci sont habituellement plus difficiles à calculer que les probabilités π_i^B.

Dans l'exemple de la section 2, nous obtenons $V(\hat{Y}_{HT}) = 14\,545$, ce qui correspond à la plus petite variance obtenues avec les méthodes proposées. Cependant, il est important de se rappeler que le calcul des probabilités de sélection π_i^B peut être difficile à effectuer, et que le calcul des probabilités de sélection conjointes $\pi_{ii'}^B$ est habituellement encore plus difficile, ce qui diminue de beaucoup l'attrait de cette méthode en pratique.

4. Simulations en utilisant des données réelles

Les simulations reflètent une enquête économique type de Statistique Canada. Nous avons choisi trois populations auprès desquelles Statistique Canada réalise fréquemment des enquêtes. Les populations d'établissements des industries de manufacture, du commerce de détail et de la restauration ont été extraites du Registre des entreprises (RE). On sait que ces populations ont une distribution asymétrique pour des variables économiques telles que le revenu, ce qui est le cas surtout pour les deux premières populations. Les simulations ont été effectuées sous SASSR stratifié selon l'industrie, la région et la catégorie de revenu. L'algorithme de Lavallée-Hidiroglou (voir [7]) a été utilisé pour créer les catégories de revenu, déterminer la taille d'échantillon et effectuer la répartition. Les établissements ont été répartis en quatre strates en fonction de leur taille : 01 – Tirage complet; 02 – Tirage partiel 1; 03 – Tirage partiel 2; 04 – Tirage nul. Un coefficient de variation de 5 % était visé

dans chaque strate par industrie et par région. Le tableau qui suit contient certaines statistiques sur la population.

Tableau 2 – Populations, tailles d'échantillon et statistiques

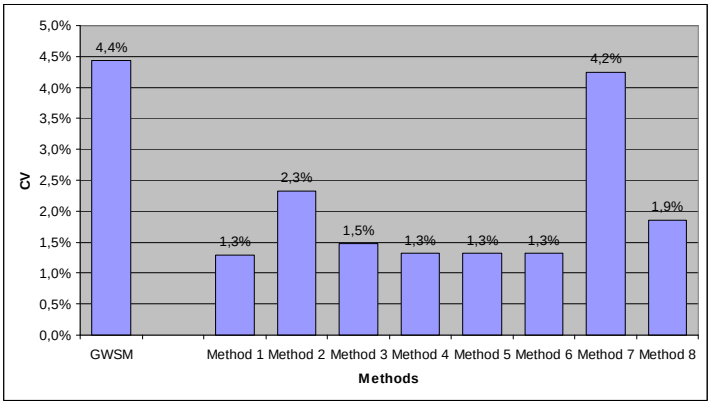
Industrie	N^B	M^A	m^A	Revenu moyen	Variance	Asymétrie
Fabrication	96 955	100 109	2 223	4 364 808	$1.08x10^{16}$	164
Commerce de détail	142 020	159 247	3 627	2 034 111	$3.29x10^{14}$	133
Restauration	107 358	113 425	2 439	561 764	$4.43*10^{12}$	106
Total	**346 333**	**372 781**	**8 289**		**---**	

La variable de revenu disponible dans le RE a été utilisée comme variable d'intérêt *y*. Pour ces simulations, puisque les valeurs de cette variable sont connues pour toutes les unités, aucune sélection d'échantillon n'était nécessaire. Les vraies variances ont été calculées d'après les données. Il convient de mentionner que, pour la méthode 2 ($\theta_{j,i}$ proportionnel à une mesure de taille x_j), le nombre d'employés a été utilisé. Pour la méthode 8, nous avons dû calculer les probabilités de sélection de toutes les entreprises. Ces probabilités ont été déterminées en exécutant une simulation Monte Carlo. Nous avons sélectionné un grand nombre d'échantillons d'établissements en utilisant le plan décrit plus haut. Pour chaque échantillon, nous avons déterminé quelle entreprise finissait par être sélectionnée. Sur ce grand nombre de répliques, nous avons pu estimer la probabilité de sélection de chaque entreprise sous ce plan de sondage à probabilités inégales. Une fois que ces probabilités ont été calculées, nous avons exécuté une simulation Monte Carlo, cette fois au niveau des entreprises en utilisant leurs probabilités de sélection respectives.

4.1. Résultats de la simulation

Pour la MGPP classique et pour toutes les méthodes présentées, nous avons calculé les estimations, les variances et les coefficients de variation (CV). Le graphique qui suit donne les CV obtenus au niveau national.

Graphique 1 – Coefficients de variation par méthode



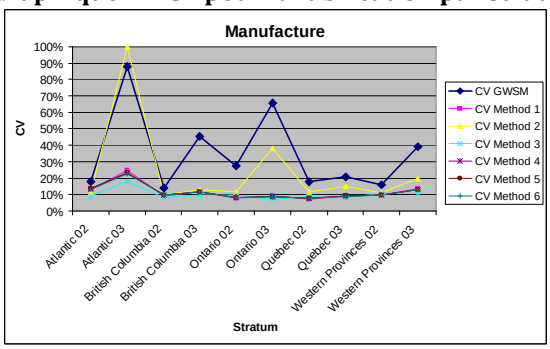
Note : « GWSM » est l'acronyme anglais de la MGPP.

Toutes les méthodes donnent une diminution de la variance, et souvent le progrès est considérable. La seule méthode ne donnant pas de diminution significative est la méthode 7 (utilisant les établissements désignés). Comme nous l'avons décrit plus haut, cette méthode consiste à sélectionner un seul établissement au sein d'une entreprise en se basant sur une variable auxiliaire et à affecter l'entreprise complète à cet établissement. Autrement dit, un établissement hérite de tous les revenus de l'entreprise. Cette approche est avantageuse quand l'établissement désigné se trouve dans une strate à tirage complet. Cependant, il arrive que l'établissement désigné se trouve dans une strate à tirage partiel, si bien que la distribution à l'intérieur de cette strate devient encore plus asymétrique. La totalité des revenus de l'entreprise multipliée par le poids de sondage de cette

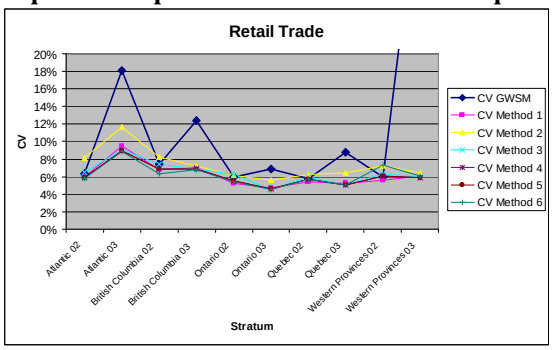
entreprise est attribuée à cette seule strate, ce qui fait augmenter considérablement la variance. Toutes les autres méthodes donnent des résultats prometteurs et nous les analyserons en détail.

Les graphiques qui suivent donnent le CV pour chaque strate à tirage partiel par industrie.

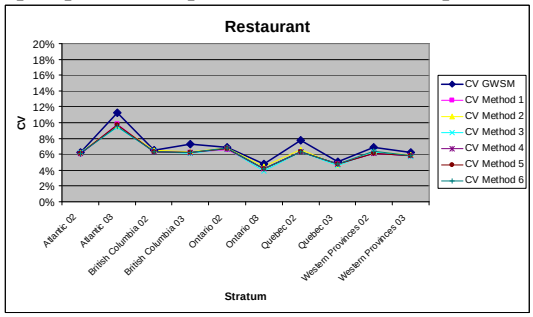
Graphique 2 – CV pour la fabrication par strate



Graphique 3 – CV pour le commerce de détail par strate



Graphique 4 – CV pour les restaurants par strate



Note 1 : « Retail Trade » correspond au commerce de détail.
Note 2 : L'échelle du CV n'est pas la même pour l'industrie de manufacture que pour les deux autres industries.
Note 3 : Les CV par strate des méthodes 7 et 8 ne sont pas présentés ici parce qu'ils ne sont pas pertinents pour la présente comparaison. En effet, la notion de strate n'est pas la même pour ces deux méthodes que pour les autres. La stratification définie par le plan de sondage original est effectuée au niveau de l'établissement. Pour les méthodes 7 et 8, l'échantillonnage a été effectué au niveau de l'entreprise et, par conséquent, une strate type pour les méthodes 1 à 6 devient un domaine pour les méthodes 7 et 8. Évidemment, les variances associées aux méthodes 7 et 8 sont beaucoup plus grandes, ce qui rend non pertinente toute comparaison avec les autres méthodes.

Les CV sont particulièrement élevés pour la MGPP classique dans certaines strates, surtout pour l'industrie de manufacture. Naturellement, nous nous y attendions, parce que cette dernière est celle pour laquelle l'asymétrie de la distribution de la variable d'intérêt était la plus prononcée parmi les trois industries. En outre, dans cette industrie, nous avons des établissements dont les revenus peuvent varier fortement au sein d'une même entreprise, et ces établissements peuvent être répartis entre plusieurs strates. En raison de ce phénomène, la variance de la MGPP classique devient très élevée.

Tous les graphiques montrent que l'utilisation de n'importe laquelle des méthodes proposées donne lieu à une réduction du CV, donc de la variance. Les CV dans les strates sont toujours (à quelques exceptions près) plus faibles, quelle que soit la méthode comparée à la MGPP classique représentée par la ligne bleu foncé avec les losanges.

4.2. Comparaison des méthodes proposées

La méthode 1 donne des résultats très prometteurs, étant donné sa simplicité. Elle produit des CV qui comptent parmi les plus faibles observés pour l'ensemble des méthodes. Elle cible réellement la source du problème de la MGPP : le besoin d'une distribution inégale des poids, proportionnelle dans une certaine mesure à la taille de la variable d'intérêt. Puisque la plupart des enquêtes économiques s'appuient sur une stratification par taille, cette méthode fonctionne très bien. Elle a également l'avantage de ne pas dépendre directement de la variable d'intérêt.

La méthode 2 utilise une variable auxiliaire (ici, le nombre d'employés) pour répartir les poids. Cette variable n'est pas bien corrélée avec la variable d'intérêt, ce qui explique principalement pourquoi cette méthode produit la réduction la plus faible de la variance. En fait, il s'agit d'une version plus faible de la méthode 3.

La méthode 3 répartit le poids proportionnellement à la variable d'intérêt *y*, c'est-à-dire le revenu de l'établissement dans l'entreprise. La méthode donne de très bons résultats aux niveaux national, ainsi que provincial. Le résultat est légèrement plus élevé que pour les méthodes 2, 4, 5 et 6, à cause de l'asymétrie prononcée de la distribution du revenu. Elle a cependant l'inconvénient de dépendre de la variable d'intérêt *y*. Autrement dit, les poids d'estimation (17) varient d'une variable d'intérêt à l'autre.

Les méthodes 4, 5 et 6 donnent des résultats fort semblables, produisant un CV compris entre 6 % et 10 % pour le commerce de détail et les restaurants, et entre 10 % et 25 % pour l'industrie de manufacture. La similarité des résultats s'explique par un élément commun aux trois méthodes, c'est-à-dire essayer d'obtenir la variance la plus faible possible. Chaque fois qu'un établissement d'une entreprise se trouve dans une strate à tirage complet, ces méthodes concentrent toutes les valeurs sur cet établissement et attribuent des liens nuls à tous les autres établissements de l'entreprise. Pour minimiser la variance, il s'agit d'un choix naturel, puisque la contribution de cette entreprise à la variance devient nulle. Comme il est fréquent qu'un établissement d'une grande entreprise se trouve dans une strate à tirage complet, la variance devient plus faible que pour n'importe quelle autre méthode. Étant donné ce phénomène, ces trois méthodes présentent le meilleur moyen de partager les poids. Les résultats ne permettent pas de déterminer laquelle est la meilleure. Toutefois, en théorie, nous nous attendrions à ce que la méthode 5 donne généralement de meilleurs résultats que les deux autres.

Comme nous l'avons vu au niveau national, la méthode 7 ne donne pas de bons résultats. Nous avons permis qu'un seul établissement désigné représente l'ensemble de l'entreprise. Les valeurs de la variable d'intérêt de tous les établissements sont totalisées au niveau de l'entreprise, puis attribuées à cet établissement désigné. Cela revient à mettre tous ses œufs dans le même panier. Si l'on examine la distribution de la variable d'intérêt par strate, elle devient encore plus asymétrique, ce qui produit une plus grande variance. Du point de vue de l'échantillonnage, une entreprise se retrouve dans une strate unique (parce qu'elle est représentée par un seul établissement) qui pourrait être une strate à tirage nul. En outre, la méthode n'est pas efficace quand il faut produire des estimations au niveau provincial ou de l'industrie. Les estimations ne peuvent pas bénéficier de la stratification des établissements, comme cela a lieu dans n'importe quelle autre méthode présentée ici. Ce fait contribue également à une plus grande variance.

La méthode 8 utilise les probabilités de sélection réelles des entreprises. Elle donne d'assez bons résultats, le CV étant de 1,9 % au niveau national, mais la réduction obtenue par la Rao-Blackwellisation de l'estimateur (1) n'est pas suffisante pour rivaliser avec la réduction obtenue en utilisant l'optimalité faible (méthodes 4, 5 et 6).

5. Conclusion

La MGPP peut être utilisée dans le contexte des enquêtes économiques, mais elle produit parfois des variances importantes, à cause de la distribution asymétrique de ce genre de populations. Afin de réduire la variance, nous avons proposé d'autres méthodes de partage des poids. Huit méthodes ont été décrites en détail. Une simulation a ensuite été exécutée pour comparer ces méthodes en se servant de données réelles. Les simulations reproduisaient une enquête économique type en utilisant un SASSR stratifié. Toutes les méthodes proposées ont donné lieu à une amélioration de la variance. Les simulations ont montré que les meilleures méthodes consistaient soit à partager les poids proportionnellement aux π_j^A (méthode 1) ou à utiliser des liens pondérés faiblement optimaux (méthodes 4, 5 et 6). Parmi ces méthodes, nous privilégions l'utilisation de liens pondérés sous sondage de Poisson (méthode 5). En théorie, cette méthode est proche de la méthode optimale, et elle est également la plus simple des méthodes 4, 5 et 6, à appliquer en pratique.

Bibliographie

[1] Cassel, C.-M., Särndal, C.-E., Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling.* John Wiley and Sons, New York, 192 pages.

[2] Deville, J.-C., Lavallée, P. (2006). Sondage indirect : Les fondements de la méthode généralisée du partage des poids. *Techniques d'enquête*, Vol. 32, No. 2.

[3] Kalton, G., Brick, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, Vol. 21, No. 1.

[4] Lavallée, P. (2002). *Le Sondage indirect, ou la Méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Belgique.

[5] Lavallée, P. (2007). *Indirect Sampling.* Springer, New York.

[6] Lavallée, P., Caron, P. (2001). Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements. *Techniques d'enquête*, Vol. 27, No. 2.

[7] Lavallée, P., Hidiroglou, M.A. (1988). Sur la stratification de la populations asymétiques. *Techniques d'enquête*, Vol. 14, No. 1.

[8] Särndal, C.-E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

[9] Statistique Canada (2010). *Le Registre des entreprises de Statistique Canada - Une brève description.* Division du Registre des entreprises, Statistique Canada, Ottawa, 9 pages, Juillet 2010.

[10] Thompson, S.K. (1990). Adaptive Cluster Sampling*. Journal of the American Statistical Association*, Vol. 85, No. 412, pp. 1050-1059.

[11] Wolter, K.M. (2007). *Introduction to Variance Estimation, 2nd edition*. Springer, New York.

Annexe

Estimations pour tous les échantillons possibles

Échantillon	$\hat{Y}$	$\hat{Y}_{classique}$
h=1: {1,2}; *h*=2: {1}; *h*=3: {1,2}	6300	4000
h=1: {1,2}; *h*=2: {1}; *h*=3: {1,3}	5400	4030
h=1: {1,2}; *h*=2: {1}; *h*=3: {1,4}	5400	4030
h=1: {1,2}; *h*=2: {1}; *h*=3: {1,5}	4700	4030
h=1: {1,2}; *h*=2: {1}; *h*=3: {1,6}	4700	4030
h=1: {1,2}; *h*=2: {1}; *h*=3: {2,3}	5400	3970
h=1: {1,2}; *h*=2: {1}; *h*=3: {2,4}	5400	3970
h=1: {1,2}; *h*=2: {1}; *h*=3: {2,5}	4700	3970
h=1: {1,2}; *h*=2: {1}; *h*=3: {2,6}	4700	3970
h=1: {1,2}; *h*=2: {1}; *h*=3: {3,4}	4500	4000

h=1: {1,2}; *h*=2: {1}; *h*=3: {3,5}	3800	4000
h=1: {1,2}; *h*=2: {1}; *h*=3: {3,6}	3800	4000
h=1: {1,2}; *h*=2: {1}; *h*=3: {4,5}	3800	4000
h=1: {1,2}; *h*=2: {1}; *h*=3: {4,6}	3800	4000
h=1: {1,2}; *h*=2: {1}; *h*=3: {5,6}	3100	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {1,2}	5400	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {1,3}	4500	4030
h=1: {1,2}; *h*=2: {2}; *h*=3: {1,4}	4500	4030
h=1: {1,2}; *h*=2: {2}; *h*=3: {1,5}	3800	4030
h=1: {1,2}; *h*=2: {2}; *h*=3: {1,6}	3800	4030
h=1: {1,2}; *h*=2: {2}; *h*=3: {2,3}	4500	3970
h=1: {1,2}; *h*=2: {2}; *h*=3: {2,4}	4500	3970
h=1: {1,2}; *h*=2: {2}; *h*=3: {2,5}	3800	3970
h=1: {1,2}; *h*=2: {2}; *h*=3: {2,6}	3800	3970
h=1: {1,2}; *h*=2: {2}; *h*=3: {3,4}	3600	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {3,5}	2900	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {3,6}	2900	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {4,5}	2900	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {4,6}	2900	4000
h=1: {1,2}; *h*=2: {2}; *h*=3: {5,6}	2200	4000
h=1: {1,2}; *h*=2: {3}; *h*=3: {1,2}	4700	3400
h=1: {1,2}; *h*=2: {3}; *h*=3: {1,3}	3800	3430
h=1: {1,2}; *h*=2: {3}; *h*=3: {1,4}	3800	3430
h=1: {1,2}; *h*=2: {3}; *h*=3: {1,5}	3100	3430
h=1: {1,2}; *h*=2: {3}; *h*=3: {1,6}	3100	3430
h=1: {1,2}; *h*=2: {3}; *h*=3: {2,3}	3800	3370
h=1: {1,2}; *h*=2: {3}; *h*=3: {2,4}	3800	3370
h=1: {1,2}; *h*=2: {3}; *h*=3: {2,5}	3100	3370
h=1: {1,2}; *h*=2: {3}; *h*=3: {2,6}	3100	3370
h=1: {1,2}; *h*=2: {3}; *h*=3: {3,4}	2900	3400
h=1: {1,2}; *h*=2: {3}; *h*=3: {3,5}	2200	3400
h=1: {1,2}; *h*=2: {3}; *h*=3: {3,6}	2200	3400
h=1: {1,2}; *h*=2: {3}; *h*=3: {4,5}	2200	3400
h=1: {1,2}; *h*=2: {3}; *h*=3: {4,6}	2200	3400
h=1: {1,2}; *h*=2: {3}; *h*=3: {5,6}	1500	3400