Sondage indirect appliqué aux populations asymétriques

Pierre Lavallée
Sébastien Labelle-Blanchet

Statistique Canada

Journées de Méthodologie Statistique, Paris
janvier 2012

1. Introduction

Dans les enquêtes économiques, la plupart des variables sont fortement asymétriques.

Deux niveaux :

- Entreprise
- Établissement

Les entreprises considérées comme étant des grappes d'établ.

Niveau de l'établissement :

- Stratification pour tirer l'échantillon.
- Permet de contrôler la représentativité géographique/industrielle/de taille.
- Difficile si la stratification est au niveau de l'entreprise.

En plus des statistiques au niveau de l'établ., nous voulons souvent obtenir des statistiques au niveau de l'entreprise.

⇒ Sélection de l'échantillon au niveau de l'établissement et élargir l'échantillon à l'entreprise complète une fois que l'échantillon des établissements est obtenu.

Base de sondage : Ensemble des établissements

Population cible : Ensemble des entreprises (grappes d'établ.)

⇒ *Sondage indirect*

Sondage indirect:

Choisir un échantillon s^A dans la base de sondage U^A
pour produire une estimation pour la population cible U^B
en utilisant les liens existants entre les deux populations.

Poids d'estimation sans biais obtenus au moyen de la
Méthode généralisée de partage des poids (MGPP).

Problème possible avec l'application de la MGPP pour les enquêtes économiques :

GRANDES VARIANCES dans les estimations produites

Problème dû à l'asymétrie de la population.

Possibilité de faire des ajustements aux poids d'estimation pour réduire la variance, tout en gardant la méthode sans biais.

2. Sondage indirect et la MGPP

Plan de sondage le plus fréquemment utilisé pour les enquêtes économiques : Sondage aléatoire simple sans remise (SASSR)

Population U^A de M^A établ. stratifiée en H strates, où la strate h contient M_h^A établissements.

Sélectionner un échantillon s_h^A de m_h^A établ. en utilisant le SASSR.

La population cible U^B contient N^B entreprises, où l'entreprise i contient ces M_i^B établissements de U^A.

Liens entre U^A et U^B identifiés par la variable indicatrice $I_{j,i}$.

Parce que chaque établissement peut seulement appartenir à une seule entreprise, les liens entre U^A et U^B sont surjectifs (*plusieurs à un, ou un à un*).

Processus de sondage indirect:

1. Pour chaque établissement j choisi dans s^A, nous identifions l'entreprise correspondante i de U^B.
2. Pour chaque entreprise i identifiée, nous supposons que nous pouvons dresser la liste de tous les M_i^B établ. de cette entreprise.
3. Nous sondons **tous les** M_i^B établ. et mesurons y_{ij}.

Échantillon s^B de n^B entreprises ($m^B = \sum_{i=1}^{n^B} M_i^B$ établ.)

Estimation du total $Y = \sum_{i=1}^{N^B} \sum_{j=1}^{M_i^B} y_{ij} = \sum_{i=1}^{N^B} Y_i$ pour U^B.

En utilisant la MGPP:

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

où

$$Z_{hj} = \sum_{i=1}^{N^B} \frac{Y_i}{L_i^B} l_{j,i} .$$

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

$$Z_{hj} = \frac{Y_i}{M_i^B} = \bar{Y}_i$$

pour $j \in i$.

$\hat{Y}$ est sans biais et

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

2.1 Utilisation des liens pondérés

Il est possible de remplacer la variable indicatrice $I_{j,i}$ par une variable quantitative $\theta_{j,i}$ représentant l'importance que nous voulons apporter au lien $I_{j,i}$.

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

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

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

$\hat{Y}_\theta$ demeure sans biais.

2.2 Utilisation des liens pondérés optimaux

Deville et Lavallée (2006) : valeurs de $\theta_{j,i}$ telles que la variance de l'estimateur $\hat{Y}_\theta$ soit (presque) minimale.

La solution n'est pas simple à écrire, et elle dépend souvent de la variable d'intérêt y.

Optimalité faible:

- Minimiser la variance de $\hat{Y}_\theta$ 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 la valeur de y.
- Relativement facile à calculer.

3. Le problème des populations asymétriques : un petit exemple

Population cible U^B de $N^B = 3$ entreprises de tailles :

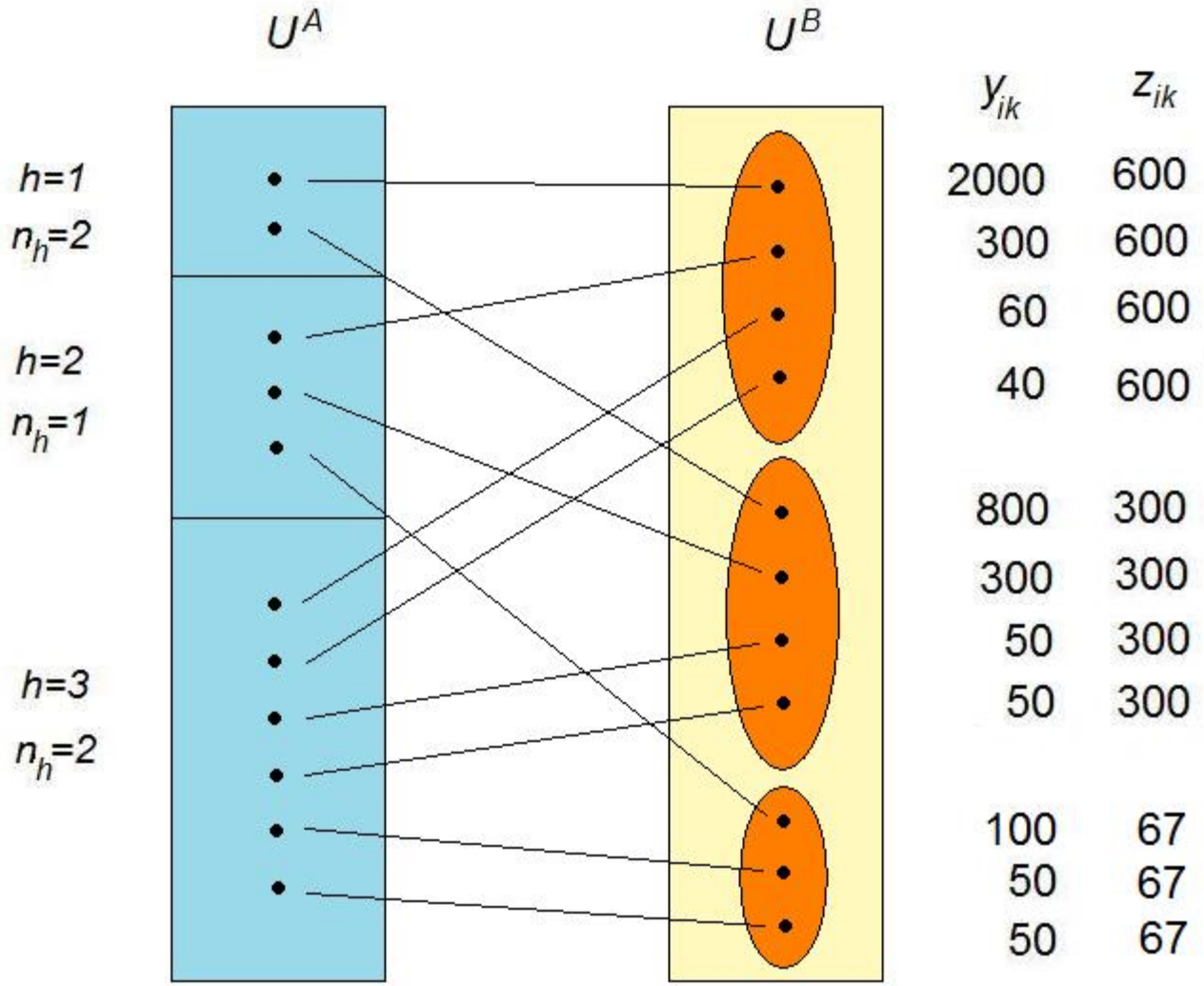
$$M_1^B = 4,\ M_2^B = 4,\ M_3^B = 3$$

Revenu y des $M^B = 11$ établ. : population asymétrique

Base de sondage U^A contenant les $M^A = 11$ établ.

Stratification :

- h=1: établ. avec $y \geq 750$; $f_1 = m_1^A / M_1^A = 1$
- h=2: établ. avec $100 \leq y < 750$; $f_2 = m_2^A / M_2^A = 1/3$
- h=3: établ. avec $y < 100$; $f_3 = m_3^A / M_3^A = 2/6 = 1/3$



y_{ik}	z_{ik}
2000	600
300	600
60	600
40	600
800	300
300	300
50	300
50	300
100	67
50	67
50	67

Calcul de la variance de $\hat{Y}$ (selon le sondage indirect).

Aussi, calcul des estimations de Y en supposant que l'on utilise un SASSR stratifié :

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

Résultats :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	**80 480!**

4. Méthodes proposées pour réduire la variance dans les estimations

Méthodes proposées fondées principalement sur l'emploi des liens pondérés.

Trois ensembles de méthodes :

1. Liens pondérés $\theta_{j,i}$ proportionnels à une certaine mesure de la taille des établissements.
2. Solutions optimales de Deville et Lavallée (2006), sous diverses hypothèses.
3. Utilisation des probabilités de sélection exactes.

4.1 Méthodes fondées sur l'emploi des liens pondérés

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

Comme la stratification est habituellement effectuée selon la taille, on attribue des poids élevés aux liens des grands établ.

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

où

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

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

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	**439 111**

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

Variable x_j disponible pour tous les établ. $j \in U^A$.

Attribue des poids élevés aux liens des grands établissements.

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

où

$$Z_{hj}^x = \frac{Y_i}{X_i} x_j$$

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

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	439 111
$V(\hat{Y}_{x})$	**686 540**

Méthode 3 : $\theta_{j,i}$ proportionnel à y_j

Nota : La méthode est applicable en pratique, à cause de la correspondance surjective entre U^A et U^B.

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

$$\Rightarrow \quad \hat{Y}_y = \hat{Y}_{\text{classique}}$$

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	**80 480**
$V(\hat{Y}_\pi)$	439 111
$V(\hat{Y}_x)$	686 540
$V(\hat{Y}_y)$	**80 480**

4.2 Méthode utilisant des liens pondérés faiblement optimaux

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

$$\hat{Y}_{\text{w-opt,SAS}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{\text{w-opt,SAS}}$$

où

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

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

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	439 111
$V(\hat{Y}_{x})$	686 540
$V(\hat{Y}_{y})$	80 480
$V(\hat{Y}_{\text{w-opt,SAS}})$	**23 111**

Explication : Pour deux des trois entreprises de U^B, il y a un établ. provenant de la strate à tirage complet h=1, ce qui place la valeur complète de Y_i sur l'établ. j_0 dont la contribution à la variance est 0.

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

Sondage de Poisson (ou sondage de Bernoulli stratifié) :

- Plan de sondage très simple.
- Plan relativement proche du SASSR stratifié.

Supposer que l'on procède à un sondage de Poisson peut être une approche raisonnable pour calculer les liens pondérés faiblement optimaux.

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

où

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

pour $j \in h$ et $j \in i$, où $\tau_i = \sum_{j=1}^{M_i^B} \frac{\pi_j^A}{(1-\pi_j^A)}$.

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	439 111
$V(\hat{Y}_{x})$	686 540
$V(\hat{Y}_{y})$	80 480
$V(\hat{Y}_{\text{f-opt,SAS}})$	23 111
$V(\hat{Y}_{\text{f-opt,SP}})$	**22 857**

Explication : Même que $V(\hat{Y}_{\text{f-opt,SAS}})$. On place la valeur complète de Y_i sur l'établ. j_0 dont la contribution à la variance est 0.

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

Un *groupe d'établ.* $j*$ comprend tous les établ. 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.

La logique de l'utilisation de groupes d'établ. est de n'avoir qu'une seule unité appartenant à une entreprise donnée par strate.

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

où

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

pour $j \in h$ et $j^* \in h$, où $\tau_i^* = \sum_{j^*=1}^{M_i^{B*}} \frac{\pi_{j*}^{A*}}{(1-\pi_{j*}^{A*})}$.

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	439 111
$V(\hat{Y}_{x})$	686 540
$V(\hat{Y}_{y})$	80 480
$V(\hat{Y}_{\text{f-opt,SAS}})$	23 111
$V(\hat{Y}_{\text{f-opt,SP}})$	22 857
$V(\hat{Y}_{\text{f-opt,grp}})$	**23 000**

Explication : Même que $V(\hat{Y}_{\text{f-opt,SAS}})$ et $V(\hat{Y}_{\text{f-opt,SP}})$.

4.3 Calcul des probabilités de sélection exactes

Méthode 7 : utilisation d'un établissement désigné

On aimerait n'avoir qu'une seule unité appartenant à une entreprise donnée par strate.

On peut choisir un seul établ. qui représentera l'entreprise complète.

Un choix naturel pour l'établissement désigné est celui ayant la plus grande valeur pour une variable donnée x.

En choisissant un seul établ. 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. $M^{A+} = N^B$.

Nous sélectionnons un échantillon s^{A+} de m^{A+} établ. désignés sous SASSR en utilisant des fractions de sondage égales aux fractions d'origine.

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

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	439 111
$V(\hat{Y}_{x})$	686 540
$V(\hat{Y}_{y})$	80 480
$V(\hat{Y}_{\text{f-opt,SAS}})$	23 111
$V(\hat{Y}_{\text{f-opt,SP}})$	22 857
$V(\hat{Y}_{\text{f-opt,grp}})$	23 000
$V(\hat{Y}_{+})$	**1 820 000**

Explication : Un des établissements désignés se trouve dans une strate à tirage partiel, si bien que la distribution à l'intérieur de cette strate devient encore plus asymétrique.

Méthode 8 : Utilisation des probabilités de sélection des entreprises

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

où π_i^B est la probabilité de sélection de l'entreprise $i \in U^B$.

π_i^B : Probabilité de sélectionner n'importe lequel des M_i^B établ. de l'entreprise i.

Correspond à la Rao-Blackwellisation de l'estimateur $\hat{Y}$.
Les π_i^B et, en particulier, les $\pi_{ii'}^B$ sont difficiles à calculer.

Résultats pour le petit exemple :

$V(\hat{Y})$ (par sondage indirect)	1 115 111
$V(\hat{Y}_{\text{classique}})$	80 480
$V(\hat{Y}_{\pi})$	439 111
$V(\hat{Y}_{x})$	686 540
$V(\hat{Y}_{y})$	80 480
$V(\hat{Y}_{\text{f-opt,SAS}})$	23 111
$V(\hat{Y}_{\text{f-opt,SP}})$	22 857
$V(\hat{Y}_{\text{f-opt,grp}})$	23 000
$V(\hat{Y}_{+})$	1 820 000
$V(\hat{Y}_{HT})$	**14 545**

5. Étude par simulations

Population : établissements canadiens dans les secteurs de la :

- fabrication
- commerce de détail
- restauration

Plan de sondage : SASSR
stratifié par industrie/région/catégorie de revenu

Statistiques sur les populations à l'étude

Variable d'intérêt :

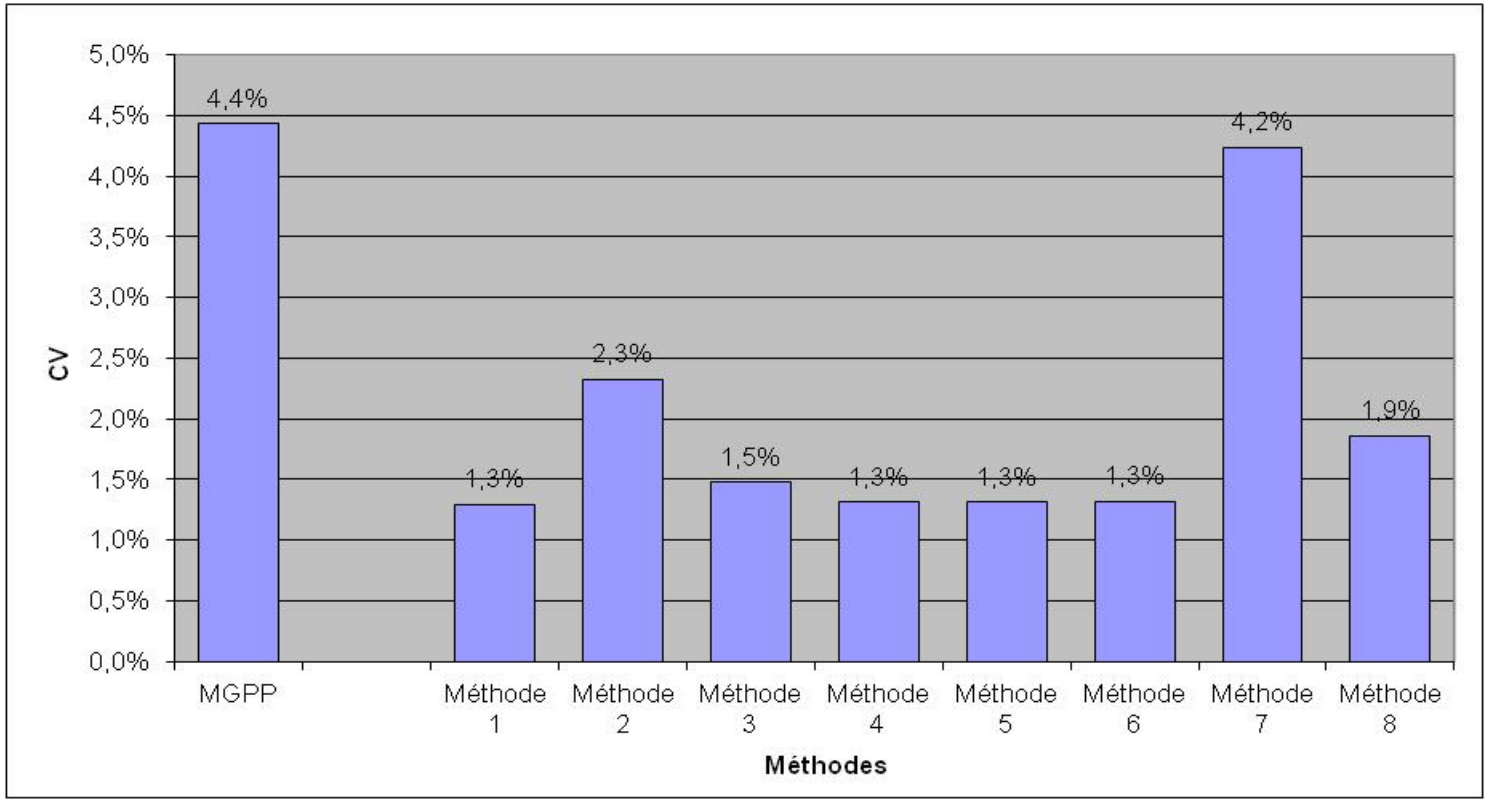
- Revenu.
- Disponible pour tous les établ. dans la population.
- Permet de calculer les vraies variances et coefficients de variation pour chaque méthode proposée.

Industrie	N^B	M^A	m^A	$\bar{Y}$	S_Y^2	**Asymétrie**
Fabrication	96 955	100 109	2 223	4 364 808	$1{,}08\text{x}10^{16}$	164
Comm. de détail	142 020	159 247	3 627	2 034 111	$3{,}29\text{x}10^{14}$	133
Restauration	107 358	113 425	2 439	561 764	$4{,}43\text{x}10^{12}$	106

Populations fortement asymétriques!

Variable auxiliaire pour la méthode 2 : nombre d'employés

Résultats :

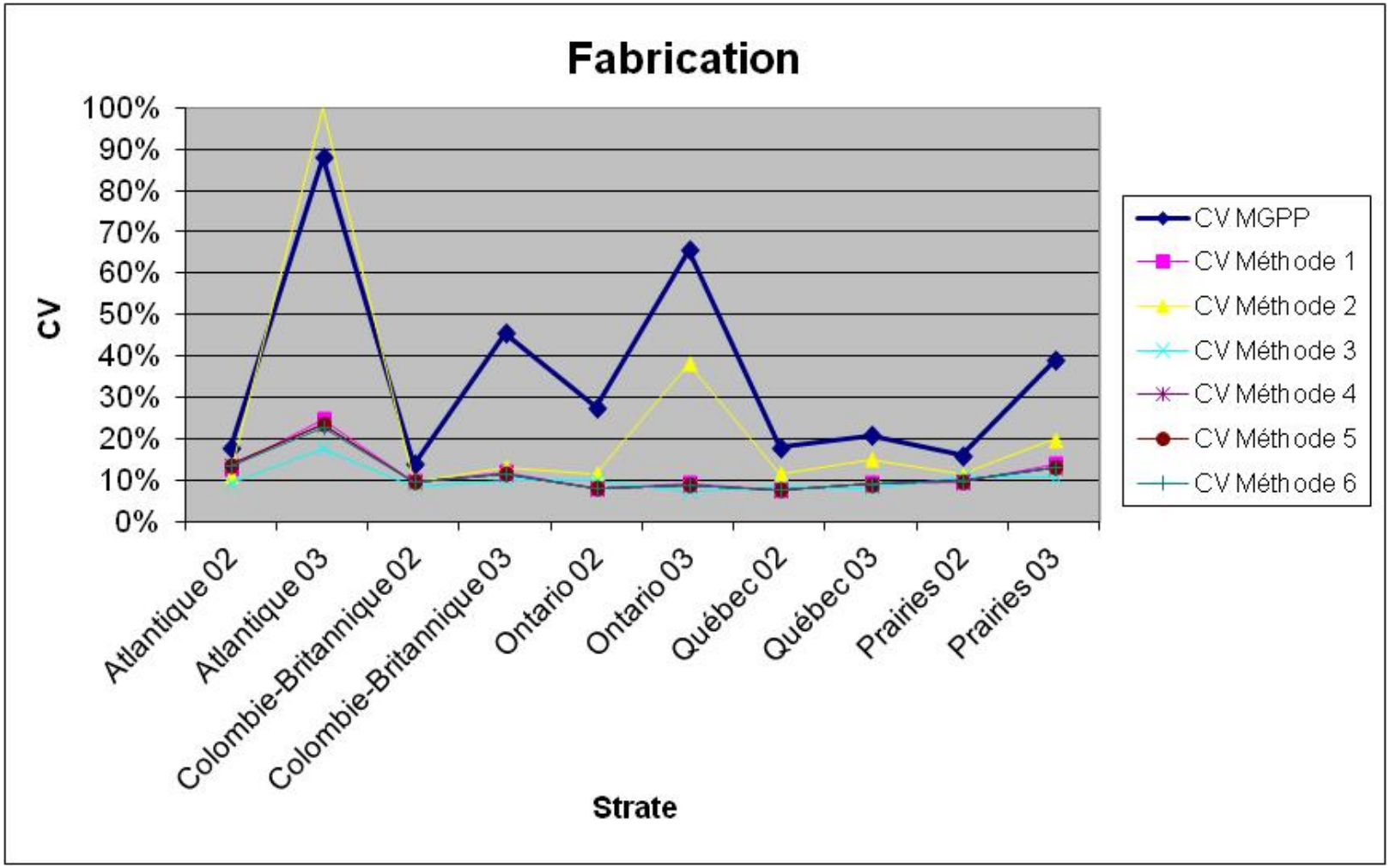


Meilleures méthodes : méthode 1 ($\theta_{j,i}$ proportionnel à π_j^A) et

méthodes 4, 5 et 6 (avec liens pondérés faiblement optimaux)

Moins bonne méthode : méthode 7 (établ. désignés)

Résultats par province (secteur de la fabrication seulement) :



Résultats :

Meilleures méthodes :

- Méthode 1 ($\theta_{j,i}$ proportionnel à π_j^A) et
- Méthodes 4, 5 et 6 (utilisant les liens pondérés faiblement optimaux)

Moins bonne méthode : méthode 2 ($\theta_{j,i}$ proportionnel à la mesure de la taille x_j)

6. Conclusions

La MGPP produit parfois de grandes variances dans le contexte des populations asymétriques.

Les méthodes proposées, sauf la méthode 7 (établ. désignés), ont généralement donné lieu à une amélioration de la variance.

Méthode 1 ($\theta_{j,i}$ proportionnel à π_j^A) offre une solution simple et facile.

Les méthodes 4, 5 et 6 (utilisant les liens pondérés faiblement optimaux) devraient être utilisées.

- Produisent des résultats très similaires.
- La **méthode 5** est la plus simple à utiliser.

Pour plus de renseignements, veuillez contacter

For more information, please contact

Pierre Lavallée

+1.613.951.2892

pierre.lavallee@statcan.gc.ca

Statistics Canada

Statistique Canada