

## Une application de l'échantillonnage équilibré: le plan de sondage des entreprises non incorporées.

Steve Fecteau et Wisner Jocelyn<sup>1</sup>

### RÉSUMÉ

**Mots clés :** échantillon équilibré, estimation, plan de sondage, représentativité

Les entreprises individuelles ou non incorporées remplissent leurs déclarations de revenus soit sur papier ou par voie électronique (Internet, disquette, etc.). Les déclarations sur papier sont coûteuses à transcrire et sont vraisemblablement entachées de plus d'erreurs. Dans cette perspective, on cherche depuis quelques années à utiliser au maximum les données acheminées par voie électronique dans le processus de production des données couvrant ces entreprises.

Maintenant que les données électroniques représentent la moitié des effectifs de la population, on commence à explorer des moyens d'utiliser uniquement celles-ci pour la production d'estimations couvrant toute la population. Dans cet article, nous nous proposons d'appliquer l'échantillonnage équilibré au fichier électronique afin d'en tirer un échantillon représentatif de toute la population et de s'en servir par la suite pour produire des estimations couvrant l'univers des entreprises non incorporées.

### 1. INTRODUCTION

Pour bien mesurer l'économie canadienne, il est impératif d'obtenir des données sur les entreprises. Ces entreprises sont incorporées ou non incorporées. Les plus grandes sont généralement incorporées, mais il existe quelques secteurs de l'économie où les entreprises non incorporées sont sinon dominantes, ou à tout le moins significatives. Par ailleurs, il y a nettement plus d'entreprises non incorporées que d'entreprises incorporées.

Les entreprises non incorporées au Canada remplissent leurs déclarations fiscales soit sur papier ou par voie électronique (Internet, disquette, etc.). Les déclarations transmises sur papier doivent être saisies manuellement afin d'obtenir un format électronique en vue d'exploitation à des fins statistiques. Un peu plus de la moitié des entreprises non incorporées choisissent de transmettre leurs déclarations électroniquement. C'est cet important volume de données électroniques disponible et les coûts de saisie importants pour les déclarants sur papier qui nous incitent à explorer des méthodes permettant l'utilisation maximale des données acheminées par voie électronique dans le processus de production des statistiques couvrant ces entreprises.

L'article est organisé ainsi : à la section 2, nous décrivons brièvement l'univers des entreprises non incorporées. Nous exposons et justifions l'utilisation de l'échantillonnage équilibré et des méthodes d'estimation que nous entendons utiliser à la section 3. Nous présentons quelques résultats et analyses à la section 4 et finalement, quelques embryons de recommandations en guise de conclusion sont exposés à la section 5.

### 2. ENTREPRISES NON INCORPORÉES

Une entreprise non incorporée est définie comme étant une entreprise ayant rempli une déclaration de revenus des particuliers (formulaire T1 de l'Agence du revenu du Canada) et ayant également déclaré des revenus provenant d'une ou de plusieurs des activités commerciales suivantes : revenus d'entreprises, de profession libérale, d'agriculture, de pêche, de location, de commissions.

---

<sup>1</sup> Steve Fecteau, méthodologiste, Division des méthodes d'enquêtes auprès des entreprises, 120 avenue Parkdale (RHC-11M), Ottawa, Ontario, Canada, K1A 0T6 ([steve.fecteau@statcan.ca](mailto:steve.fecteau@statcan.ca)). Wisner Jocelyn, méthodologiste principal, Division des méthodes d'enquêtes auprès des entreprises, 120 avenue Parkdale (RHC-11M), Ottawa, Ontario, Canada, K1A 0T6 ([wisner.jocelyn@statcan.ca](mailto:wisner.jocelyn@statcan.ca)).

## 2.1 Caractéristiques des entreprises non incorporées

Il y a environ 3,5 millions d'entreprises non incorporées au Canada dont un peu plus de la moitié (52%) sont des déclarants électroniques. Les entreprises non incorporées génèrent environ 5% du revenu brut total au Canada et près de 18% du revenu net. Le tableau 1 plus bas fournit le revenu moyen des entreprises selon le type de déclaration produit (électronique ou sur papier). Nous constatons que les moyennes des revenus bruts et nets sont significativement plus grandes pour les déclarants sur papier. Bien qu'il n'y ait pas eu d'étude spécifique pour savoir pourquoi les plus grandes entreprises semblent rapporter en général sur papier, il semble bien que l'obligation de faire attester sa déclaration par un professionnel de la fiscalité et de l'authentifier par une signature soit une des raisons qui poussent les plus grandes entreprises à envoyer leurs déclarations sur papier.

**Tableau 1. Moyenne des revenus par type de déclaration**

Déclarant	Moyenne Revenu Brut	Moyenne Revenu net
Électronique	261819	11423
Papier	694587	13048

## 2.2 Le problème

L'Agence de revenu du Canada fournit à Statistique Canada un fichier contenant tous les déclarants électroniques. Ce fichier contient toutes les variables pertinentes que nous cherchons à mesurer. Essentiellement, nous souhaitons produire des estimations pour les variables de dépenses (salaires, les paiements d'intérêts, etc.) ainsi que pour les variables de revenus (revenu brut, le revenu net, etc.). Nous disposons également de totaux de revenus pour la population complète des entreprises non incorporées. Le problème consiste donc à produire des estimations pour les variables de revenus et de dépenses en se servant uniquement du fichier des répondants électroniques et des totaux de revenus disponibles sur le fichier de données externe.

À la section précédente, nous avons constaté que le revenu moyen des déclarants électroniques est passablement différent de celui des déclarants sur papier, il serait donc hasardeux de conduire l'inférence statistique directement sur cet échantillon. Néanmoins, notre démarche consiste à tirer le plus grand sous-échantillon possible des répondants électroniques qui « représentera » l'univers complet des entreprises non incorporées. À cette fin, nous utilisons l'approche de l'échantillonnage équilibré.

## 3. ÉCHANTILLONNAGE ÉQUILIBRÉ

Bien que les plans de sondage équilibrés ne soient pas couramment utilisés, la méthode de l'échantillonnage équilibré elle, est connue depuis longtemps. Elle peut être abordée sous le prisme d'un plan de sondage où sa mise en œuvre implique la dérivation de poids de sélection des unités ou encore comme un simple moyen permettant d'arriver à un échantillon et de conduire ensuite l'inférence sans égard au plan utilisé. Bien que nous ayons considéré l'approche par le plan, nous adoptons ici le point de vue modéliste de Royall et Cumberland (1981a et 1981b). Nous utilisons ainsi la définition donnée dans Valliant, Dorfman et Royall (2000).

Soit  $\bar{x}_s^{(j)} = \sum_{i \in s} x_i^j / n$  et  $\bar{x}_U^{(j)} = \sum_{i \in U} x_i^j / N$  où  $s$  représente l'échantillon et  $j$  l'ordre du moment

( $j=1,2,\dots,J$ ). Nous dénotons par  $N$  et  $n$  respectivement la taille de la population et de l'échantillon. Si pour un échantillon  $s$  donné on a :

$$\bar{x}_s^{(j)} = \bar{x}_U^{(j)} \quad (1)$$

alors nous dirons que  $s$  est un échantillon équilibré d'ordre  $j$ .

### 3.1 Application aux données sur les entreprises non incorporées

Comme nous l'avons vu à la section précédente, nous disposons donc de la portion des déclarants électroniques, et à partir de cette sous-population, nous utilisons plusieurs algorithmes d'échantillonnage équilibré pour obtenir un échantillon équilibré d'ordre 2 ou plus. Formellement, soit  $x_i$  le revenu net provenant de l'unité  $i$  de la population  $E$  des déclarants électroniques, nous chercherons le plus grand échantillon de  $E$  tel que

$$\left| \frac{\bar{x}_U^{(j)} - \bar{x}_E^{(j)}}{\bar{x}_U^{(j)}} \right| \leq \varepsilon \quad (2)$$

où  $\bar{x}_U^{(j)}$  et  $\bar{x}_E^{(j)}$  représentent respectivement le moment de la population et de l'échantillon équilibré provenant de  $E$ , et  $\varepsilon$  représente l'erreur relative tolérée. Par exemple,  $\varepsilon \in [0.05 \text{ à } 0.10]$ .

### 3.2 Algorithmes de sélection

Tillé (2001) décrit deux méthodes intéressantes que nous avons pu adapter à notre problème. La méthode de Deville, Grobras et Roth et la méthode du cube. Nous avons également adapté un algorithme de tirage aléatoire basé sur l'échantillonnage « rejective » tel que décrit dans Hajék (1981). Finalement, nous avons également considéré une méthode que nous avons baptisé « méthode des quadrants » qui consiste à diviser la population en  $2^j$  axes. On forme le  $j$ -vecteur des moments de la population. Ensuite en partant de  $E$ , on désélectionne une unité à la fois et on compare le  $j$ -vecteur des moments de l'échantillon résultant à ceux de la population. On utilise le critère d'arrêt mentionné à l'équation (2) plus haut.

### 3.3 Estimation

Nous cherchons donc à produire des estimations pour les variables d'intérêt de dépenses et de revenus.

Soit  $s$ , l'échantillon équilibré provenant de la population des répondants électroniques  $E$ ,  $x$ , une variable auxiliaire (ex. Revenu brut),  $y$ , une des variables d'intérêt,  $n$  la taille de l'échantillon et  $N$ , la taille de la population  $U$ .

Nous postulons le modèle suivant, qui a l'avantage de produire des estimations du total sans biais si l'échantillon est équilibré (ce qui est notre cas) :

$$\text{Modèle 1 : } y_i = \beta_0 + \varepsilon_i \quad (3)$$

où  $\beta_0$  et  $\varepsilon_i$  sont les paramètres habituels d'un modèle de régression. Nous cherchons donc à estimer :

$Y = \sum_{i \in U} y_i$  par l'approche prédictive. Soit  $\hat{Y}$  un estimateur de  $Y$  alors,

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \in U-s} \hat{y}_i \quad (4)$$

avec  $\hat{y}_i = \hat{\beta}_0$ . Un estimateur de la variance de  $\hat{Y}$  sous le modèle se déduit aisément comme:

$$v(\hat{Y}) = \frac{(N-n)}{f} \hat{\sigma}^2 \quad (5)$$

avec  $f=n/N$  et  $\hat{\sigma}^2 = \sum_{i \in s} (y_i - \bar{y})^2 / (n-1)$ .

Une analyse de régression préalable nous a montré que le modèle suivant était acceptable pour la plupart des variables :

$$\text{Modèle 2 : } y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (6)$$

Pour ce modèle, l'estimateur du total est similaire à (4) avec  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ . L'estimateur de variance du modèle s'obtient comme :

$$v(\hat{Y}) = \frac{(N-n)}{f} \hat{\sigma}^2 \left[ 1 + \frac{(\bar{x}_s - \bar{x}_U)^2}{(1-f) \sum_{i \in s} (x_i - \bar{x})^2 / n} \right] \quad (7)$$

où  $\hat{\sigma}^2 = \sum_{i \in s} (y_i - \hat{y}_i)^2 / (n-2)$ .

Pour éviter les problèmes potentiels qui peuvent survenir si le modèle n'est pas adéquat, on préconise l'estimateur de variance robuste  $v_h(\hat{Y})$  suggéré par Royall et Cumberland (1981). Cet estimateur est une modification légère de (7) pour protection contre le choix du modèle.

#### 4. ANALYSE DES RÉSULTATS

Pour tester la validité de notre stratégie, nous avons utilisé les données de l'année 2003 (plan de sondage stratifié) en considérant les sous-populations des comptables, des chauffeurs de taxis et limousines, des spécialistes de laboratoires, des spécialistes de la traduction et les commerçants de gros (grossistes). Nous avons conservé uniquement une partie de chaque sous-population. Nous avons utilisé les paramètres du plan de sondage courant (plan stratifié, avec échantillon contenant des déclarants électroniques et sur papier). Ensuite, les estimations ponctuelles furent reproduites en utilisant également les 2 modèles donnés par les équations (3) et (6) ainsi que les estimateurs de variance donnés par les équations (5) et (7). Nous avons également incorporé l'estimateur de variance  $v_h(\hat{Y})$  mentionné à la section précédente. Le tableau 2 plus bas montre l'erreur relative *err* pour le plan de sondage courant, le modèle 1 donné par l'équation (3) et le modèle 2 donné par l'équation (6) :

$$err = \left| \frac{Y_U - \hat{Y}}{Y_U} \right| * 100 \quad (8)$$

Remarquons au passage, que l'erreur relative est significativement plus petite pour les deux modèles, comparativement à celle du plan. Il ne semble pas y avoir de gain important à faire au niveau de l'estimateur ponctuel entre le modèle simple (modèle 1) et le modèle de régression un peu plus avancé (modèle 2).

**Tableau 2. Erreur relative entre le revenu brut total et son estimation (%)**

Enquête	Plan de sondage	Modèle 1	Modèle 2
Laboratoires	35.68	9.39	6.85
Traduction	25.46	3.14	3.56
Comptabilité et tenue de livres	71.70	13.93	15.50
Taxis et Limousines	55.76	4.25	6.42
Grossistes	28.16	17.71	16.66

Le tableau 3 ci-dessous présente les estimations ponctuelles pour les salaires rapportés. Malheureusement, cette variable n'étant pas rapporté au niveau de la population, nous ne pouvons donc pas calculer l'erreur relative, cependant, les mêmes différences observées entre le plan de sondage et les deux modèles semblent persister. Encore une fois, il ne semble pas y avoir de différence importante entre les résultats des deux modèles.

**Tableau 3. Estimations pour les salaires et autres bénéfices**

Enquête	Plan de sondage	Modèle 1	Modèle 2
Laboratoires	\$5,098,667.56	\$2,486,034.58	\$2,536,247.28
Traduction	\$3,587,679.63	\$1,429,777.13	\$1,427,098.99
Comptabilité et tenue de livres	\$54,002,288.20	\$230,980,746.42	\$236,814,994.06
Taxis et Limousines	\$82,879,561.03	\$25,178,393.71	\$24,317,938.92
Grossistes	\$96,800,694.82	\$98,434,351.51	\$99,345,901.78

Il semble bien que les estimations provenant des modèles soient plus proches des paramètres de la population que l'on désire mesurer. Nous allons maintenant considérer la précision de ces estimateurs en regardant les estimations des coefficients de variation *cv*:

$$cv(\hat{Y}) = \frac{\sqrt{v(\hat{Y})}}{\hat{Y}} \quad (9)$$

Nous ne pouvons comparer directement les variances, puisque la variance selon le modèle indique la précision sur l'échantillon sélectionné et uniquement cet échantillon tandis que la variance selon le plan est une tentative d'estimer la variabilité entre les échantillons possibles. Cependant, le coefficient de variation fournit une manière simple de comparer la précision des estimateurs ponctuels.

**Tableau 4. Coefficients de variation pour le revenu brut**

Enquête	Plan de sondage	Modèle 1	Modèle 2	C.V. vh (%)
	C.V. (%)	C.V. (%)	C.V. régression (%)	
Laboratoires	16.97	6.40	3.12	3.15
Traduction	5.24	2.34	1.56	1.56
Comptabilité et tenue de livres	5.54	7.61	7.34	7.29
Taxis et Limousines	6.94	2.88	1.99	2.02
Grossistes	6.37	2.63	2.45	2.58

Le tableau 4 nous montre les coefficients de variation selon le modèle ou le plan et également pour l'estimateur robuste de la variance. Nous constatons que les coefficients de variation sont nettement plus petits pour les modèles que pour le plan. Ceci s'explique sans doute par le volume important de données disponibles pour bâtir le modèle (4 fois plus d'unités au moins dans les échantillons équilibrés pour les modèles que dans l'échantillon du plan). Par ailleurs, les mêmes conclusions s'appliquent pour les coefficients de variation de la variable salaire.

## 5. CONCLUSION

La grande quantité de données disponible à travers le fichier électronique permet de construire des modèles simples et robustes pour estimer les paramètres de totaux et leur variabilité, pour l'ensemble des variables d'intérêts. Nous n'avons pas montré de diagnostic sur la validité des modèles, mais la plupart des tests effectués sur les modèles confirment leur validité pour la majorité des variables d'intérêts. En effet, rien dans l'analyse de données entreprise en amont de notre étude sur l'échantillonnage équilibré ne semble indiquer de défaillance majeure des modèles postulés. D'ailleurs, l'intuition même nous permet de croire que la relation existant entre le revenu et les principales variables de dépenses pour de petites entreprises peut être assez bien approximée par une relation linéaire voire même par une constante dans certains cas.

Il reste bien sûr à comparer ce que qu'il aurait été obtenu en empruntant le chemin d'un plan de sondage particulier utilisant également l'ensemble des répondants électronique accompagné d'une stratégie d'estimation utilisant toute l'information auxiliaire disponible. Nous le ferons à un stade ultérieur de notre étude.

## 6. BIBLIOGRAPHIE

Hájek, J. (1981), *Sampling in a finite population*, New York, Marcel Dekker.

Royall, R.M., Cumberland, W.G., (1981a), *An empirical study of the ratio estimator and estimators of its variance*, Journal of the American Statistical Association, 76, 66-77.

Royall, R.M., Cumberland, W.G., (1981b), *The finite Population Linear Regression Estimator and estimators of its variance – An Empirical Study*, Journal of the American Statistical Association, 76, 924-930.

Tillé, Y. (2001), *Théorie des sondages*, Paris, Dunod.

Valliant, R., Dorfman, A.H., Royall, R.M., (2000), *Finite population sampling and inference*, New York, Wiley.