L'UTILISATION DES DONNÉES FISCALES : L'EXPÉRIENCE DES ENTREPRISES NON INCORPORÉES

Javier Oyarzun[1] et Richard Laroche[2]

RÉSUMÉ

Statistique Canada accorde de plus en plus d'importance à l'utilisation des données administratives dans ses différents programmes d'enquêtes. Par exemple, les données fiscales provenant des entreprises non incorporées (T1) sont utilisées afin de produire des estimations pour différents secteurs économiques. Toutefois, une telle base de données demande un certain traitement statistique, tel que la détection de valeurs aberrantes, l'imputation de variables manquantes et la classification à l'aide de modèles statistiques. Cet article décrit les aspects essentiels dans la résolution de problèmes méthodologiques afin de produire des estimations pour la population des entreprises non incorporées.

MOTS CLÉS : Données administratives; données fiscales; entreprises non incorporées; imputation; sociétés de personnes; T1; valeurs aberrantes.

ABSTRACT

Statistics Canada has begun relying more and more on the use of administrative data for its various survey programs. T1 fiscal data provided by unincorporated enterprises are an example of the kinds of data used at Statistics Canada in order to produce estimates for different economic sectors. At the same time, that sort of database requires a certain degree of statistical processing, such as outlier detection, imputation of missing variables and classification using statistical models. This paper will describe the essential aspects involved in resolving methodological problems related to estimate production for the unincorporated enterprise population.

KEY WORDS: Administrative data, Fiscal data, Imputation, Outliers, Partnerships, T1, Unincorporated enterprises.

1. INTRODUCTION

Statistique Canada accorde de plus en plus d'importance à l'utilisation des données administratives dans ses différents programmes d'enquêtes. Un des principaux avantages que procurent les données administratives est la réduction du fardeau de réponse auprès des répondants aux enquêtes. C'est la raison pour laquelle celles-ci sont utilisées dans le programme d'estimation des entreprises non incorporées, plus communément appelées T1. Tous les Canadiens ayant perçu un revenu sur lequel un impôt doit être payé est tenu de remplir un formulaire d'impôt appelé T1 à la fin de l'année durant laquelle ce revenu a été perçu. Si un revenu de profession, d'entreprise, de commission, de location, d'agriculture ou de pêche est déclaré sur le formulaire T1 pour une année donnée, l'individu est alors considéré comme une entreprise non incorporée.

L'Agence du Revenu du Canada (ARC) partage avec Statistique Canada ses fichiers de données fiscales. Deux de ces fichiers sont relatifs aux données T1 et sont fournis sur une base annuelle. Le premier, le fichier des documents T1 ayant fait l'objet d'une cotisation (ARF), contient l'ensemble des individus ayant déclaré un montant différent de zéro pour au moins un des six types de revenus listés plus haut. Les revenus bruts et nets pour chacune des six sources de revenus sont les variables principales apparaissant sur ce fichier. En 2006, on comptait un peu plus de 3,7 millions d'individus ayant rapporté des revenus d'entreprises non incorporées au pays. Le deuxième, le fichier des données T1 transmises

[1]Javier Oyarzun, Statistique Canada, 100, promenade du Pré Tunney, Ottawa, Canada, K1A 0T6, javier.oyarzun@statcan.gc.ca
[2]Richard Laroche, Statistique Canada, 100, promenade du Pré Tunney, Ottawa, Canada, K1A 0T6, richard.laroche@statcan.gc.ca

électroniquement (E-File), est un sous-ensemble du ARF. Il contient toutes les entreprises non incorporées ayant fourni leurs données électroniquement à l'ARC. Dans ce fichier, plus d'une centaine de variables sont disponibles. En 2006, environ 75% des répondants ont rapporté leurs informations de façon électronique. L'utilisation de ces fichiers administratifs pose toutefois de nombreux défis. Le tableau 1.1 illustre bien les défis à relever, que ce soit par l'absence du code industriel (en gris foncé), par la présence de valeurs aberrantes (en gris) ou encore par l'ambiguïté des sociétés de personnes (en gris pâle). Cet article explique comment nous avons surmonté ces défis pour produire des estimations robustes et précises. À la section 2, nous discutons de l'imputation du code d'activité industrielle lorsque celui-ci est manquant. La section 3 est consacrée à l'importance de la détection des valeurs aberrantes. À la section 4, il est question des sociétés de personnes et de la surestimation que celles-ci occasionnent si elles ne sont pas correctement identifiées. La dernière section porte sur la méthodologie d'estimation mise en place pour tenir compte du fait que certaines variables ne sont disponibles que pour les répondants électroniques.

Tableau 1.1 : Exemple du fichier T1 fourni par l'ARC à Statistique Canada

Présence de valeur aberrrante

Présence d'une société de personnes

Numéro d'assurance social	Nom	Adresse	Code de la province	Code industriel	Revenu Brut	Revenu net	Source principale de revenu	Type
111 111 111	X	131 St-Laurent	10	531111	12 051	5 000	Location	E-File
222 222 222	X	131 St-Laurent	10	531111	12 051	1 000	Location	E-File
333 333 333	A	22 Maisoneuve	11	485310	29 000	10 000	Entreprise	P-File
444 444 444	B	111 Dalhousie	13	722110	60 000	500	Entreprise	E-File
555 555 555	C	1555 1st Ave	48	000000	1 200 000 000	100	Agriculture	P-File
666 666 666	D	222 George	59	541110	25 000	2 000	Professionnel	E-File
777 777 777	E	2333 York	61	311111	120 000	30 000	Entreprise	E-File
888 888 888	F	2222 5th Ave	13	000000	1 025	100	Location	P-File
999 999 999	G	231 Sherbrooke	12	111140	10 000	-1 000	Agriculture	E-File
000 000 001	H	521 Clark	46	000063	50 000	20 000	Commission	E-File
000 000 002	L	5678 Elizabeth	47	000000	1 000	300	Entreprise	P-File

Absence du code industriel

2. IMPUTATION DU CODE D'ACTIVITÉ INDUSTRIELLE

2.1 Description de la méthode

Une variable importante du fichier ARF est le code SCIAN tel que rapporté par le déclarant. Le SCIAN est un système de classification des industries ayant une structure hiérarchique dont le code permet d'identifier l'activité industrielle à différents niveaux, soit du secteur général d'activité (par le code à deux chiffres) jusqu'à niveau très détaillé (code à six chiffres). Lorsque Statistique Canada reçoit le fichier ARF, environ 30% des codes SCIAN (au code à 6 chiffres) sont manquants ou erronés. Puisque des estimations pour tous les croisements SCIAN x PROVINCE sont requises, un code SCIAN valide est nécessaire pour chaque unité afin de ne pas créer de sous-estimation. Si celui-ci est manquant ou erroné, il sera partiellement ou complètement imputé à partir des autres informations rapportées par le répondant ou en se servant d'autres sources de données. Le processus d'imputation se divise en deux étapes. Nous tentons d'abord de trouver un SCIAN valide (au moins au niveau 2) à l'aide d'informations existantes provenant d'autres sources, puis nous imputons un SCIAN au niveau 6 valide de manière probabiliste.

Ainsi, il arrive fréquemment qu'un individu possède plusieurs entreprises non incorporées dans différentes industries. Dans ce cas, l'individu peut rapporter autant de déclarations qu'il a d'entreprises, chacune apparaissant comme une entrée différente dans le fichier E-File et un code SCIAN valide est généralement associé à chaque déclaration. Lorsque cela survient, le code SCIAN du ARF est manquant. Dans cette situation, le code SCIAN du ARF est imputé par le code SCIAN de la déclaration financière ayant le revenu brut total le plus élevé. Un peu plus de 50% des SCIAN manquants sont imputés de cette manière.

Lorsqu'un code SCIAN valide ne peut être trouvé sur le fichier E-File, le fichier ARF de l'année précédente est utilisé. Si aucun SCIAN valide n'est trouvé à l'aide de cette source, le fichier ARF datant de deux ans est alors utilisé. Environ 2% des SCIAN manquants sont imputés avec cette méthode.

En troisième lieu, le registre des entreprises (RE) de Statistique Canada est utilisé (le RE est une base de sondage centrale contenant de l'information sur l'ensemble des entreprises du pays). Seules les entreprises non incorporées ayant des employés ou collectant la taxe sur les produits et services (TPS) apparaissent sur le RE. Le RE ne couvre donc qu'une faible partie des entreprises non incorporées. Environ 7% des unités du ARF sont imputés avec un code SCIAN provenant du RE.

Indirectement, le type de revenu déclaré sur le ARF peut donner une bonne indication de l'activité industrielle de l'entreprise. Par exemple, plus de 94% des T1 sur le ARF ayant comme source de revenu principale la location ont SCIAN='531111'. Il est donc raisonnable de croire que 94% des unités du ARF dont le SCIAN est manquant et ayant déclaré des revenus de location comme activité principale ont en fait SCIAN='531111'. La méthode d'imputation utilisant une distribution multinomiale consiste donc à attribuer un nombre aléatoire entre 0,0000 et 100,0000 aux unités du ARF ayant comme source de revenu principale la location et dont le SCIAN est inconnu. Nous créons un tableau de fréquence cumulée (voir tableau 2.1) montrant la distribution des SCIAN pour les unités ayant déclaré comme source principale un revenu de location.

Tableau 2.1 : Distribution des unités dont la source principale de revenu est la location

SCIAN	Fréquence	Pourcentage	Fréquence cumulative	Pourcentage cumulatif
111110	404	0,047	404	0,0470%
…	…	…	…	…
441110	4	0,0005	21 536	2,5188%
441210	**7**	**0,0008**	**21 543**	**2,5196%**
441220	13	0,0015	21 556	2,5211%
485310	4	0,0005	21 560	2,5216%
531111	**804 880**	**94,1359**	**826 440**	**96,6575%**
…	…	…	…	…
811111	3	0,0004	855 019	100%

Par exemple, si pour une unité dont le SCIAN doit être imputé nous choisissons un nombre aléatoire entre 2,5188 et 2,5196, le SCIAN sera imputé à '441210'. De même, si le nombre aléatoire tiré se situe entre 2,5216 et 96,6575, le SCIAN sera imputé à '531111'. Il est facile de voir que la distribution des SCIAN imputés sera la même que celle des SCIAN non imputés; les résultats de l'imputation suivent donc une loi de distribution multinomiale. Nous répétons le processus pour les 5 autres types de revenu.

2.2 Impact de l'imputation du code industriel

Le tableau 2.2 illustre les dangers d'ignorer les entreprises non incorporées sans code industriel valide ou de créer une catégorie « code industriel inconnu » : basé sur les données de 2006, nous perdrions 1 019 293 observations (27,89%) et aurions une sous-estimation de 14,21% du revenu brut total et de 18,48% du revenu net total.

Tableau 2.2 : Estimations avec et sans imputation du code industriel (2006)

	Sans imputation[1]	Total (avec imputation)[1]	Différence (observations imputées)	Différence relative
Nombre d'observations	2 634 987	3 654 280	1 019 293	27,89%
Revenu brut total	221 769 118 250$	258 511 591 889$	36 742 473 639$	14,21%
Revenu net total	39 007 558 898$	47 852 831 361$	8 845 272 463$	18,48%

1. *Sans la présence de valeurs aberrantes (voir section 3).*

3. DÉTECTION DES VALEURS ABERRANTES

3.1 Description de la méthode

Les valeurs aberrantes sont des valeurs erronées ou incohérentes compte tenu des autres données. Un bon exemple d'une telle valeur est 999 999 999$, qui apparaît à une cinquantaine de reprises dans le champ « revenu brut total » sur le fichier ARF. Il arrive aussi que certaines sociétés en commandite, entreprises incorporées (T2) ou autres types d'entreprises se retrouvent dans le fichier des entreprises non incorporées. Différentes méthodes existent pour détecter ces valeurs.

La première technique utilisée mesure la différence observée dans les valeurs rapportées d'une année à l'autre pour un même déclarant T1 (en utilisant la méthode d'Hidiroglou-Berthelot (1986)). La deuxième technique consiste à étudier les différences entre les observations d'une même industrie pour une même année en utilisant la méthode Sigma Gap (Ingram et Davidson, 1983).

3.2 Impact de la détection des valeurs aberrantes

Le tableau 3.2 présente les dangers de l'utilisation des données administratives sans préalablement faire la détection de valeurs aberrantes. L'estimation du revenu brut des entreprises non incorporées sans le traitement des valeurs aberrantes est 8,4 fois plus élevée que ce qu'il est en réalité.

Tableau 3.2 : Estimations avec et sans valeurs aberrantes (2006)

	Avec valeurs aberrantes[1]	Sans valeurs aberrantes[1]	Différence	Différence relative
Nombre d'observations	3 670 182	3 654 280	15 902	0,44%
Revenu brut total	2 165 613 100 000$	258 511 591 889$	1 907 101 508 111$	737,72%
Revenu net total	49 247 823 938$	47 852 831 361$	1 394 992 577$	2,92%

1. Avec l'imputation du code SCIAN (voir section 2).

4. IDENTIFICATION DES SOCIÉTÉS DE PERSONNES

4.1 Description de la méthode

Nous sommes en présence d'une société de personnes lorsque deux individus ou plus sont associés à une même entreprise non incorporée. Au niveau fiscal, tous les partenaires doivent déclarer les montants relatifs à la société de personnes et non pas les montants relatifs à la part possédée dans la société de personnes (à l'exception du revenu net). Par exemple, si un couple possède un logement qui rapporte des revenus de location de 5 000$ annuellement, chacun des deux membres du couple doit déclarer 5 000$ de revenu brut à l'ARC. Seul le revenu net doit être rapporté proportionnellement à la participation de chacun dans l'entreprise. Pour éviter de surestimer les revenus et les dépenses des entreprises non incorporées, il est donc important que ces sociétés de personnes soient bien identifiées. Or, il n'existe pas de variable unique permettant de tous bien les identifier. Il faut donc utiliser plusieurs variables provenant de différents fichiers administratifs pour faire un bon travail d'identification.

Certaines variables du ARF nous permettent, par déduction, d'identifier des sociétés de personnes (à l'aide des revenus historiques, nom de famille, code industriel, adresse, etc.). Les trois fichiers suivants sont aussi utilisés :

- **E-File** : Certains répondants ont clairement identifié leur(s) partenaire(s).
- **T5013** : Ce fichier fournit les détails de toutes les sociétés de personnes de 6 membres et plus.
- **TCOP**: Ce fichier donne la concordance entre les numéros d'assurance sociale et les numéros d'entreprise.

4.2: Impact de l'identification des sociétés de personnes

Les sociétés de personnes peuvent causer une surestimation des revenus et dépenses si elles ne sont pas correctement identifiées. Le tableau 4.2 démontre l'impact de ce processus. L'absence d'identification des sociétés de personnes produit 598 243 dédoublements d'observations (16,37%) et conduit à une surestimation du revenu brut total de 74,2 milliards de dollars pour les entreprises non incorporées.

Tableau 4.2 : Estimations avec et sans le processus d'identification des sociétés de personnes (2006)

	Sans processus d'identification des sociétés de personnes[1]	Avec le processus d'identification des sociétés de personnes[1]	Différence	Différence relative
Nombre d'observations	3 654 280	3 056 037	598 243	16,37 %
Revenu brut total	258 511 591 889$	184 321 912 409$	74 189 679 480$	28,70 %
Revenu net total	47 852 831 361$	47 852 831 361$	0$	0 %

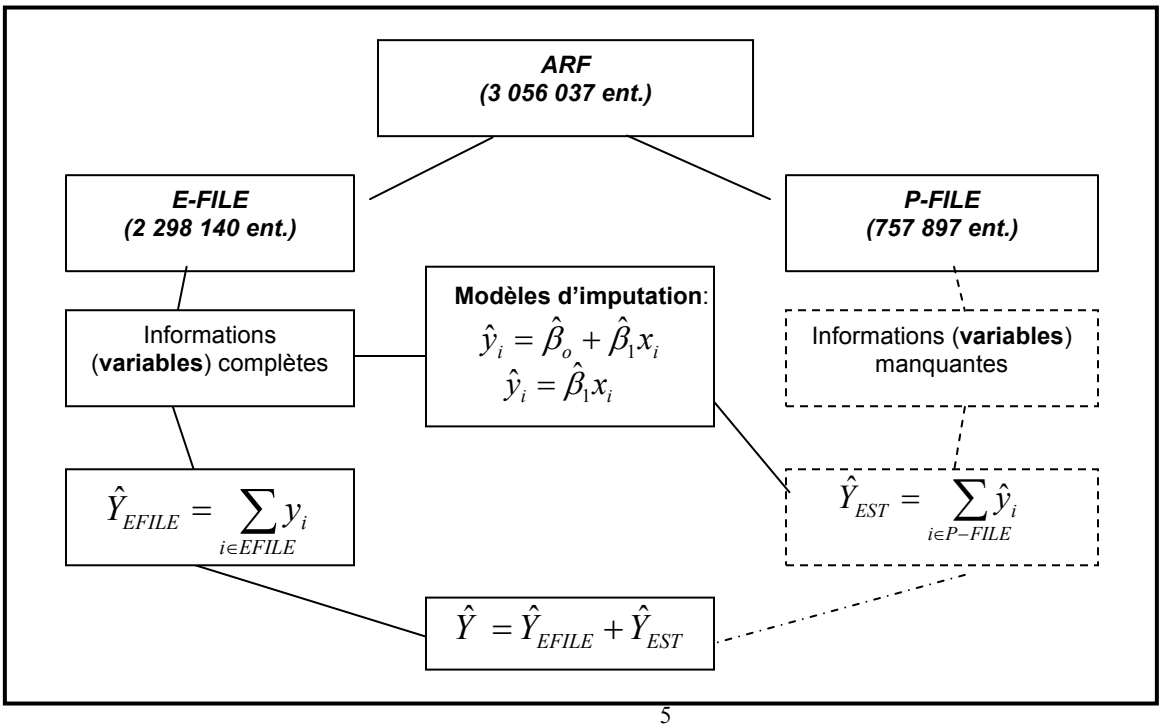
1. Avec l'imputation du code SCIAN et sans la présence de valeurs aberrantes.

5. PROCESSUS D'ESTIMATION

5.1 Méthodologie utilisée pour l'année de référence 2006

Une fois les données administratives traitées, il est possible de produire des estimations. Le principal utilisateur de ces estimations, produites pour tous les croisements de secteurs industriels et de province, est le système de comptabilité nationale de Statistique Canada. Des estimations sont faites pour une soixantaine de variables de revenus, dépenses, capital, etc. Ces variables sont disponibles directement sur le fichier E-File. Toutefois, le fichier E-File contient seulement 75 % de la population. Les estimations pour ces variables sont dérivées à l'aide de modèles statistiques. Ces modèles sont construits à l'aide des déclarants électroniques (toutes les variables sont disponibles pour ceux-ci), puis sont appliqués aux déclarants papiers (voir figure 5.1).

Figure 5.1 : Diagramme du processus d'estimation pour les entreprises non incorporées



5.2 Modèles d'imputation

Deux modèles différents ont été utilisées pour l'imputation : l'imputation par la régression linéaire simple et l'imputation par le ratio. Les détails sont donnés dans ce qui suit.

5.2.1 Modèle d'imputation par régression linéaire

Pour les variables corrélées avec le revenu brut total, le modèle choisi est le suivant:

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i \qquad (5.2.1.1)$$

où β_0, β_1 et ε_i sont les paramètres habituels d'un modèle de régression et x_i est la variable « revenu brut total » disponible pour toutes les unités du ARF. Nous cherchons donc à estimer $Y = \sum_{i \in U} y_i$ par l'approche prédictive.

Soit $\hat{Y}$ un estimateur de *Y*. On a alors

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \in U-s} \hat{y}_i \qquad (5.2.1.2)$$

avec $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$. Un estimateur de la variance de $\hat{Y}$ sous le modèle de régression linéaire s'obtient comme suit (Särndal, Swensson et Wretman, 1992):

$$V(\hat{Y}) = \frac{(N-n)}{f} \hat{\sigma}^2 \left[1 + \frac{(\bar{x}_s - \bar{x}_U)^2}{(1-f)\sum_{i \in s} (x_i - \bar{x}_s)^2 / n} \right] \qquad (5.2.1.3)$$

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

5.2.2 Modèle d'imputation par ratio

Pour les variables qui ne sont pas corrélées avec le revenu brut total, un modèle par le ratio a été préféré. Noter que ce modèle n'est qu'en fait une forme simplifiée du modèle de régression utilisée ci-haut où l'ordonnée à l'origine est absente du modèle. Sous forme mathématique, ce modèle s'écrit de la façon suivante :

$$y_i = \beta_1 x_i + \varepsilon_i \qquad (5.2.2.1)$$

où β_1 et ε_i sont les paramètres habituels d'un modèle par ratio et x_i est la variable « revenu brut total » disponible pour toutes les unités du ARF. Nous cherchons donc à estimer $Y = \sum_{i \in U} y_i$ par l'approche prédictive.

Soit $\hat{Y}$ un estimateur de Y. On a alors

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{i \in U-s} \hat{y}_i \qquad (5.2.2.2)$$

avec $\hat{y}_i = \hat{\beta}_1 x_i$. Un estimateur de la variance de $\hat{Y}$ sous le modèle par ratio s'obtient comme suit (Särndal, Swensson et Wretman, 1992) :

$$V(\hat{Y}) = N^2 \hat{\sigma}^2 \left(\frac{\bar{x}_u}{\bar{x}_s} \right)^2 \frac{1-f}{n} \sum_{i \in s} (y_i - \hat{y}_i)^2 / (n-1) \qquad (5.2.2.3)$$

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

5.3 Nouvelle méthodologie d'estimation et d'imputation

Une nouvelle méthodologie sera mise en place pour l'année de référence 2007 puisque de moins en moins de variables sont manquantes. Cela est dû au fait que Statistique Canada reçoit un nouveau fichier de l'Agence du revenu du Canada, le fichier « barcode ». Le fichier « barcode » contient les données des individus ayant utilisé un programme informatique pour remplir leur déclaration fiscale et qui l'ont envoyée par la poste. De plus, la proportion de répondants électroniques croit d'année en année. Ainsi, la création d'un fichier de micro-données complet a été établi. La méthode d'imputation du plus proche voisin est utilisée pour trouver un donneur aux répondants papiers pour lesquels les détails des revenus et des dépenses seront imputés. En 2007, les répondants papiers ne représentent plus que 15% du total de la population T1. Une variance due à l'imputation sera également calculée. Cette nouvelle méthodologie sera décrite en détail dans un article ultérieur.

6. CONCLUSION

Par l'entremise de l'article, nous avons discuté à travers chaque section d'un aspect de la méthodologie employée dans le processus d'estimation des entreprises non incorporées. L'imputation du code industriel, la détection des valeurs aberrantes, l'identification de sociétés de personnes et l'emploi de modèles d'imputation sont toutes des étapes importantes dans la production d'estimations pour les entreprises non incorporées. Ces étapes sont essentielles compte tenu que les données administratives ne sont pas exemptes d'erreurs non dues à l'échantillonnage. L'utilisation de données administratives procurent l'avantage d'être sans coût et d'éliminer le fardeau de réponse, mais leur utilisation ne doit se faire sans la garantie d'une bonne qualité du produit final. Les méthodes de traitement et d'estimation discutées dans cet article répondent à cette exigence et permettent ainsi à Statistique Canada de mieux répondre aux besoins en statistiques pour la population des entreprises non incorporées.

REMERCIEMENTS

Les auteurs remercient Mathieu Thomassin et François Brisebois pour leurs commentaires judicieux.

RÉFÉRENCES

Hidiroglou, M.A., et Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology,* **12**, 73-83.

Ingram, S. et Davidson, G. (1983). Methods used in designing the National Farm Survey, *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 220-225.

Särndal C, Swensson B. et Wretman, J. (1992). Model Assisted Survey Sampling, New York, Springer-Verlag.