

L'UTILISATION ACCRUE DES DONNÉES FISCALES DANS LE CADRE DE L'ENQUÊTE UNIFIÉE SUR LES ENTREPRISES

Eric Pelletier ¹

RÉSUMÉ

En 2002, un projet a été mis de l'avant ayant pour objectif d'utiliser davantage les données fiscales dans le but de réduire le fardeau de réponse des entreprises, de réduire le coût des enquêtes et d'améliorer ou maintenir la qualité des données. Ainsi, pour l'année de référence 2003, l'Enquête unifiée sur les entreprises (EUE) utilisera des données fiscales comme données de remplacement pour environ 50% des établissements simples échantillonnés. Cet article traitera des difficultés et des défis reliés à l'estimation impliquant des données d'enquête et des données fiscales.

MOTS CLÉS : Données fiscales; enquêtes entreprises; estimation; modélisation.

ABSTRACT

In 2002, a project was launched with the objective to use more tax data in order to reduce the response burden of businesses, reduce the cost of survey programs and to improve or maintain the data quality. For reference year 2003, the Unified Enterprise Survey (UES) will use tax data as replacement data for about 50% of the sampled simple establishments. The paper will describe the difficulties and issues at the estimation stage when using survey data and tax data.

KEY WORDS: Business Surveys, Estimation, Tax Data, Modelling.

1. INTRODUCTION

Suite au Projet d'amélioration des statistiques économiques provinciales (PASEP) mis en place à Statistique Canada en 1997 et décrit dans Laniel et Royce (1998), plusieurs enquêtes économiques annuelles ont été intégrées en une seule enquête soit l'Enquête unifiée sur les entreprises (EUE). Plusieurs modifications sont survenues à l'EUE au cours des années mais, pour l'année de référence 2003 (AR2003), un important changement a été mis en place, soit le Projet de remplacement des données d'enquête par des données fiscales. Ce projet a pour but d'utiliser davantage les données fiscales provenant de l'Agence du revenu du Canada, qui suite à une entente, sont à la disposition de Statistique Canada.

Cet article débute avec une brève description de l'EUE, incluant les processus d'échantillonnage et d'estimation (sections 2 à 4). Par la suite, les défis reliés au Projet de remplacement des données d'enquête par des données fiscales sont décrits (section 5). Finalement, les développements futurs reliés à l'EUE, dont ceux du Projet de remplacement des données d'enquête par des données fiscales sont discutés (section 6).

2. L'ENQUÊTE UNIFIÉE SUR LES ENTREPRISES (EUE)

L'EUE regroupe plusieurs enquêtes annuelles dont les différents processus d'enquêtes, de la sélection de l'échantillon à la diffusion des données, sont intégrés afin de favoriser l'harmonisation des méthodes et des concepts et la réduction des coûts. Lors de sa création en 1997, l'EUE couvrait 7 enquêtes distinctes. Pour l'AR2003, l'EUE contient 18 enquêtes et pour l'AR2004, 3 autres enquêtes s'ajouteront. Depuis sa création, le but premier de l'EUE est de produire des estimations fiables au niveau des provinces et des industries canadiennes. Les principaux objectifs de l'enquête pour l'AR2003 sont :

¹ Eric Pelletier (eric.pelletier@statcan.ca), Division des méthodes d'enquêtes auprès des entreprises, 11^e étage, édifice R.-H.-Coats, Statistique Canada, Ottawa, Ontario, K1A 0T6,

- La production des estimations pour les variables financières et non financières pour tous les secteurs industriels couverts par l'EUE. Les principales variables financières sont le total des revenus, le total des dépenses et le coût des biens vendus;
- Une utilisation accrue, au-dessus des seuils d'exclusion², des données fiscales provenant de l'Agence du revenu du Canada;
- Une réduction du fardeau de réponse des entreprises;
- Une réduction des coûts reliés aux enquêtes;
- Une amélioration potentielle ou un maintien de la qualité des données.

Plus précisément, au niveau de l'estimation, les objectifs secondaires sont :

- La production d'un fichier complet de micro-données de toutes les unités sélectionnées dans l'échantillon;
- La production des estimations pour tous les domaines d'intérêt pour les variables financières et non financières.

3. LE PROCESSUS D'ÉCHANTILLONNAGE DE L'EUE

3.1 Registre des entreprises de Statistique Canada

Le Registre des entreprises de Statistique Canada est utilisé comme base de sondage pour la sélection de l'échantillon. Le registre est constitué de toutes les entreprises non incorporées (communément appelées les T1), d'entreprises incorporées (communément appelées les T2) et d'autres types d'entreprises (institution publiques, etc.) opérant au Canada. Chaque entreprise est classifiée selon un code d'activité industriel de 6 chiffres. Ces codes proviennent du Système de classification des industries de l'Amérique du Nord (SCIAN). Sur le registre, les entreprises sont représentées selon une structure statistique qui se décompose en quatre niveaux hiérarchiques : l'entreprise, la compagnie, l'établissement et l'emplacement.

3.2 L'unité d'échantillonnage

L'unité d'échantillonnage est définie comme l'ensemble des établissements, à l'intérieur d'une même entreprise, qui sont dans une même province et un même groupe industriel. L'unité d'échantillonnage correspond donc à une grappe d'établissements. Par exemple, si les établissements A et B sont dans la même province, ont le même code SCIAN et appartiennent à la même entreprise, ces deux établissements feront partie de la même unité d'échantillonnage. Les entreprises peuvent être classifiées en deux catégories : simples ou complexes. Les entreprises simples sont des entreprises où l'établissement et l'entreprise représentent la même entité. Si une entreprise n'est pas constituée de la sorte, il s'agit donc d'une entreprise complexe. Par exemple, une entreprise reliée à deux établissements est complexe.

3.3 Stratification de la population

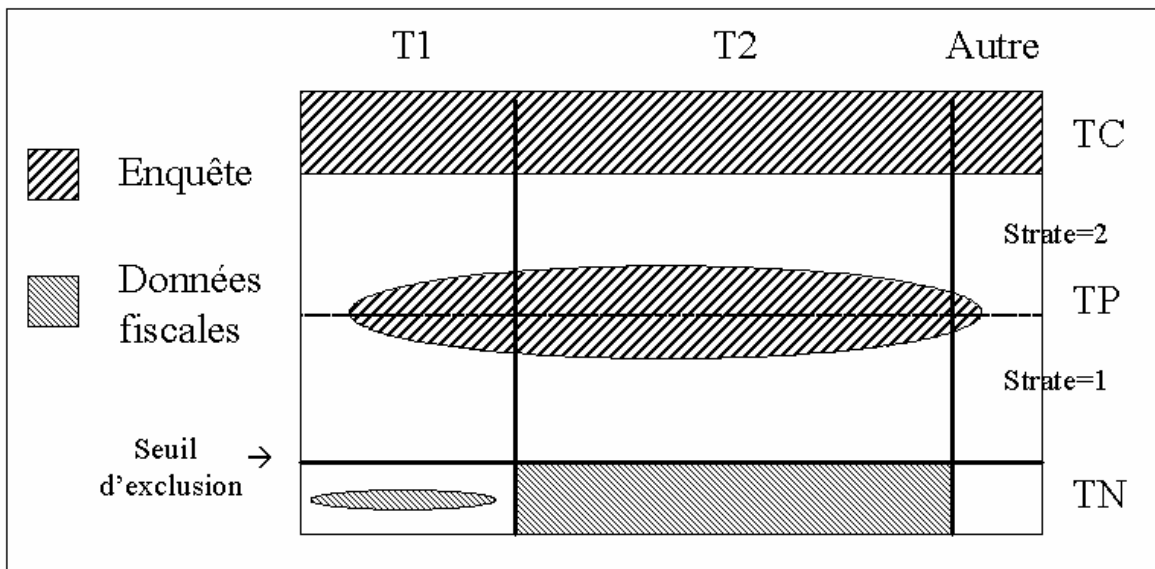
La stratification de la population s'effectue à trois niveaux : au niveau des provinces, au niveau des industries (codes SCIAN) et selon la taille définie en terme du revenu de l'unité d'échantillonnage. Les deux premiers niveaux forment une cellule. À l'intérieur de chaque cellule, quatre strates d'unités sont ensuite créées, basées sur le revenu de l'unité d'échantillonnage :

- Une strate d'unités à tirage complet (TC);
- Deux strates d'unités à tirage partiel (TP) : une strate de petites unités (strate 1) et une de moyennes unités (strate 2);
- Une strate d'unités à tirage nul (TN).

Pour ce qui est de la strate d'unités à tirage nul, aucun questionnaire n'est envoyé à ces unités. Ce sont des unités qui se trouvent sous les seuils d'exclusion et pour ces unités, des données fiscales sont utilisées afin de produire des estimations. La figure 1 présente le plan de sondage de l'EUE.

Figure 1 : Plan de sondage de l'EUE (avant le Projet de remplacement des données d'enquête par des données fiscales)

² Le concept des seuils d'exclusion sera défini à la section 3.3.



4. LE PROCESSUS D'ESTIMATION DE L'EUE

Le but du processus d'estimation est de reproduire, à l'aide des données recueillies par le biais de l'enquête et des données fiscales, le vrai portrait de la population pour l'année de référence. L'estimation totale pour chacune des variables d'intérêt est la somme des estimations provenant de la portion enquêtée (c'est-à-dire au-dessus des seuils d'exclusion, voir section 3.3) et des estimations pour la portion de la population obtenues à partir des données fiscales (i.e. pour les unités sous les seuils d'exclusion).

L'estimateur de Horvitz-Thompson est utilisé pour la portion enquêtée. Préalablement à l'estimation, un module de détection de valeurs aberrantes est appliqué. Par conséquent, si une unité est considérée aberrante, son poids est réduit à 1 et le poids restant³ de l'unité est redistribué à travers les autres unités de sa strate. Par la suite, les estimations sont produites en utilisant les poids ajustés.

Pour ce qui est de la portion des unités à tirage nul, un processus différent est mis en place pour les unités T2 et pour les unités T1 (voir figure 1). Un recensement des données fiscales est effectué pour les unités T2. Pour les unités T1, contrairement aux unités T2, un échantillon est utilisé étant donné que le coût associé à l'extraction de l'information fiscale pour les unités T1 est beaucoup plus élevé. Pour les unités à tirage nul autres que les unités T1 et T2 (c.-à-d. la portion « Autre » dans la figure 1), aucune estimation n'est produite étant donné que cette portion est très minime et négligeable.

Par la suite, les estimations sont produites pour différents domaines : au niveau des provinces et/ou des industries, de la catégorie d'unités (unités T1 ou T2), etc. Les variances échantillonnales, pour la portion enquête et pour la portion représentée par des unités T1 à tirage nul, sont calculées. À noter que pour ce qui est des unités T2 à tirage nul, étant donné qu'il s'agit d'un recensement des unités T2, la variance échantillonnale est nulle. Finalement, le coefficient de variation (CV) national est obtenu.

5. PROJET DE REMPLACEMENT PAR DES DONNÉES FISCALES : LES DÉFIS

Un projet de l'envergure du Projet de remplacement des données d'enquête par des données fiscales comporte plusieurs défis. Les principaux défis ont été regroupés en quatre catégories :

1. Les défis liés aux concepts et aux définitions;
2. Les défis liés à la base de sondage;
3. Les défis liés à l'échantillonnage;
4. Les défis liés à l'imputation et à l'estimation.

³ Le poids restant correspond à la différence entre le poids original de l'unité et son nouveau poids, suite à l'application du module de détection de valeurs aberrantes. Le nouveau poids d'une valeur aberrante est fixé à 1.

Nous passerons donc en revue les principaux défis dans les sections subséquentes.

5.1 Les défis reliés aux concepts et aux définitions

Lorsque le projet a été mis en oeuvre, une des premières étapes a été de comparer les concepts entre les données d'enquête et les données fiscales. Étant donné que les définitions des variables ne sont pas nécessairement les mêmes, les analystes et les comptables ont eu à effectuer beaucoup de travail dans le but d'établir des équivalences entre les données d'enquête et les données fiscales. Une fois les équivalences complétées, une analyse utilisant les deux sources de données a été réalisée. Dans un premier temps, l'analyse a porté sur les unités T2. Ces dernières représentent la majorité des unités simples de l'EUE.

Ainsi, une comparaison entre les données d'enquête et les données fiscales a été réalisée pour les unités T2 simples. Le tableau 1 présente les résultats pour l'Enquête sur les services de restauration pour l'AR2002 pour les principales variables financières. On remarque que pour la majorité d'entre elles, les différences relatives entre les deux sources de données sont faibles. Cependant, pour la variable « Dépréciation et amortissement », la différence relative semble plus élevée. Dans ce cas-ci, il s'agit d'un exemple où, malgré le travail des analystes, les définitions du côté de l'enquête et du côté des données fiscales n'ont pu être réconciliées.

Tableau 1 : Différences relatives entre les données d'enquête et les données fiscales pour l'Enquête sur les services de restauration pour les unités T2 simples de l'AR2002

Variables financières	Total : Données d'enquête (en millions)	Total : Données fiscales (en millions)	Différence relative
Total des revenus d'exploitation	3 360	3 383	0,67%
Total des revenus	3 367	3 396	0,85%
Salaires et traitements	903	953	5,57%
Total – rémunération du travail	997	983	-1,49%
Dépréciation et amortissement	85	95	11,92%
Total des dépenses	3 182	3 325	4,49%
Coût des biens vendus	1 129	1 237	5,78%

5.2 Les défis reliés à la base de sondage

Une des difficultés engendrées par l'utilisation accrue des données fiscales est le manque de rétroaction provenant de l'enquête pour mettre à jour la base de sondage. Par exemple, il y a la possibilité de faire de la sur-couverture si aucun ajustement n'est apporté aux données fiscales. En effet, comme il n'y a pas de pré-contact⁴ pour ces unités, on ne pourra pas savoir si l'unité est inactive ou hors du champ de l'enquête.

Pour tenter de corriger cette situation, différentes options ont été étudiées. Une première idée a été de calculer un taux d'unités inactives et hors du champ de l'enquête de l'année précédente (AR2002) et de l'appliquer à l'année courante (AR2003). Les différentes analyses effectuées ont toutefois démontré qu'il n'y a pas de stabilité dans le temps en ce qui a trait au taux d'unités inactives et hors du champ de l'enquête. Par conséquent, l'utilisation d'un taux obtenu de l'année précédente a été rejetée.

L'option retenue a été de calculer un taux d'unités inactives et hors du champ de l'enquête provenant des unités simples enquêtées de l'année courante et de l'appliquer aux unités utilisant des données fiscales comme données de remplacement⁵.

Avant de prendre la décision finale, on a simulé l'application de cette méthode en utilisant les données de l'AR2002 et on a comparé les estimations obtenues aux estimations originalement produites pour l'AR2002. En résumé, la méthode utilisée pour estimer le taux d'unités inactives et hors du champs de l'enquête a permis d'obtenir des estimations très proches des estimations attendues pour les variables financières clés. Pour ce qui est des variables financières autres que

⁴ Le pré-contact consiste à communiquer avec l'entreprise dans le but de mettre à jour uniquement l'information administrative sur la base de sondage comme le code d'activité de l'entreprise, si l'unité est active ou non, etc.

⁵ À l'exception des enquêtes où 100% des unités simples sont remplacées par des données fiscales.

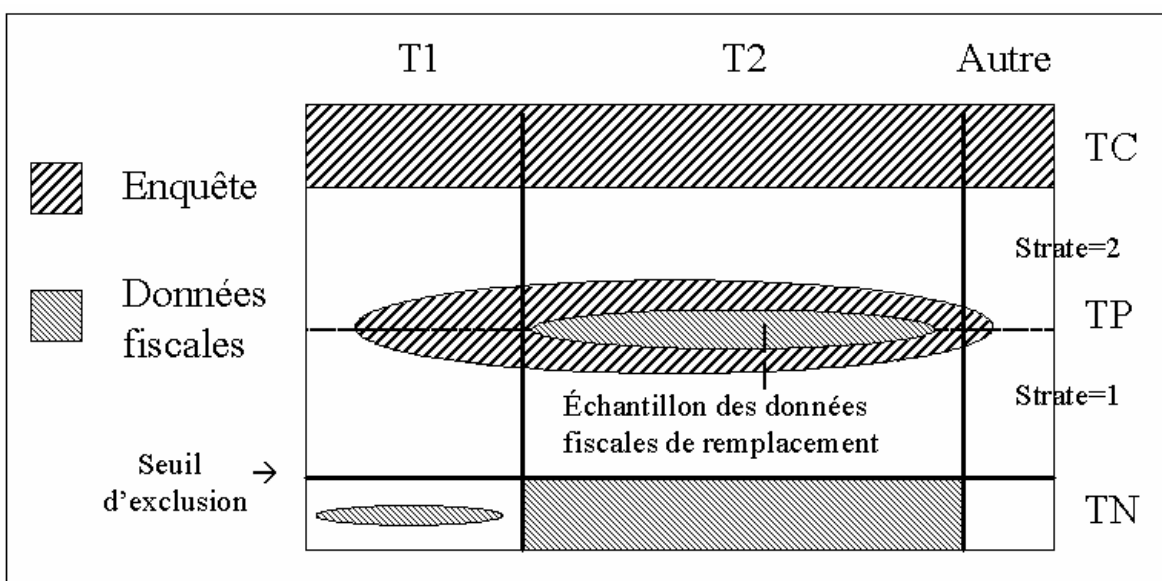
les variables financières clés, les résultats sont moins satisfaisants. On a attribué ces différences au fait qu'il existe une plus grande variabilité au sein de la population pour ces variables. Les taux d'unités inactives et hors du champ de l'enquête obtenus lors de la simulation pour l'AR2002 oscillent entre 12% et 20% pour la majorité des enquêtes.

5.3 Les défis reliés à l'échantillonnage

Le principal défi lors du processus d'échantillonnage est la façon de sélectionner le sous-échantillon d'établissements de la partie enquête pour lesquels des données fiscales seront utilisées. Pour l'AR2003, sept enquêtes ont participé au Projet de remplacement des données d'enquête par des données fiscales. Diverses méthodologies et différents critères ont été utilisés pour la sélection du sous-échantillon. Pour la majorité des enquêtes, seules les unités T2 simples étaient éligibles pour la sélection du sous-échantillon. Le taux de remplacement était de 50% ou 100%. Ce taux était appliqué soit sur l'échantillon des unités éligibles ou soit sur les unités éligibles faisant partie des strates à tirage partiel. Finalement, la répartition du sous-échantillon se faisait, selon les enquêtes, à l'aide de la répartition proportionnelle ou selon la répartition de Neyman.

La figure 2 représente une enquête qui a sélectionné un sous-échantillon d'unités T2 simples et ce dans les strates à tirage partiel uniquement. Comme le montre la légende, ce sous-échantillon utilisera des données fiscales provenant de l'Agence du revenu du Canada, comme pour les unités sous les seuils d'exclusion.

Figure 2 : Plan de sondage de l'EUE incluant l'échantillon pour le remplacement par des données fiscales



Il est intéressant de comparer les tailles des différents sous-échantillons et leur contribution en terme de revenu⁶ par rapport à l'échantillon principal. Le tableau 2 présente ces différents résultats pour les sept enquêtes du Projet de remplacement des données d'enquête par des données fiscales. On notera que le pourcentage des unités pour lesquelles on planifie utiliser des données fiscales n'est pas nécessairement 50% ou 100% étant donné que ce pourcentage était appliqué sur les unités simples éligibles et non sur la totalité de l'échantillon.

Tableau 2 : Tailles d'échantillon pour les sept enquêtes faisant partie du Projet de remplacement des données d'enquête par des données fiscales (portion au-dessus des seuils d'exclusion)

Enquête	Taille de l'échantillon	# d'unités utilisant des données fiscales	% des unités de l'échantillon	% du revenu pour les unités de l'échantillon
---------	-------------------------	---	-------------------------------	--

⁶ Revenu provenant du Registre des entreprises utilisé lors de l'échantillonnage.

	pour la partie enquête	de remplacement dans l'échantillon	utilisant des données fiscales	utilisant des données fiscales
Services de restauration	3 368	2 392	71,0%	23,7%
Location et gestion de biens immobiliers	4 105	1 286	31,3%	6,3%
Conseils en gestion	3 574	1 290	36,1%	16,9%
Services de réparations et d'entretiens	4 342	3 636	83,7%	37,9%
Design spécialisé	1 219	357	29,3%	17,4%
Commerce de détail en magasin	11 116	2 878	25,9%	3,9%
Commerce de gros	9 584	1 134	11,8%	5,7%

En analysant ces résultats, on remarque que pour cinq des sept enquêtes, le pourcentage des unités pour lesquelles on planifie utiliser des données fiscales oscille entre 11,8% et 36,1%. Pour ce qui est des deux autres enquêtes, les taux sont de 71,0% et 83,7%. L'explication est la suivante : ces deux enquêtes (Services de restauration et Services de réparations et d'entretiens) ont une stratégie qui consiste à utiliser 100% des unités simples pour le sous-échantillon tandis que pour les cinq autres enquêtes, on utilise une stratégie qui consiste à ne sélectionner que 50% des unités simples pour le sous-échantillon.

Toutefois, pour l'enquête de Services de restauration, même si 71,0% des unités sont sélectionnées pour utiliser des données fiscales de remplacement, cela ne représente que 23,7% du revenu total de l'échantillon pour cette enquête. Ceci est dû au fait que seules les unités simples sont éligibles au remplacement par des données fiscales et que les unités simples sont, dans la plupart des cas, beaucoup plus petites en terme de revenu que les unités complexes. Par exemple, l'enquête sur le Commerce de gros couvre peu d'unités simples et beaucoup d'unités complexes. Ainsi, même si on planifie d'utiliser des données fiscales pour 50% des unités simples, cela ne représente qu'un faible pourcentage d'unités par rapport à l'échantillon total (11,8%) et un pourcentage encore plus faible en terme de revenu (5,7%).

5.4 Les défis reliés à l'imputation et à l'estimation

Comme décrit dans les sections précédentes, les estimations pour la partie enquête pour l'AR2003 seront produites à l'aide d'un mélange de données d'enquête et de données fiscales. Étant donné le peu de développements dans la littérature réalisés sur la production d'estimation combinant à la fois des données d'enquête et des données fiscales, les défis sont d'autant plus grands. Un de ces défis est la production d'estimation pour les variables non financières, variables pour lesquelles les données fiscales ne sont pas disponibles. Un exemple est la variable « Coûts des boissons alcoolisées » pour l'enquête sur les Services de restauration. C'est une variable non financière propre à cette enquête et cette variable n'est pas disponible par les données fiscales.

Comme les données d'enquête ne sont pas disponibles pour les unités du sous-échantillon utilisant des données fiscales, trois options ont été considérées pour le traitement des variables financières et non financières de ces unités :

1. Repondération pour les unités simples ne faisant pas partie du sous-échantillon;
2. Utilisation directe des données fiscales pour la majorité des variables financières et imputation massive pour les autres variables financières et les variables non financières;
3. Modélisation des données fiscales pour une ou plusieurs des variables financières et/ou non financières.

5.4.1 Repondération

La première option envisagée a été la repondération mais cette option a été rapidement écartée pour plusieurs raisons. La principale raison est qu'elle ne permet pas une utilisation directe des données fiscales. La repondération simule en fait une imputation par la moyenne au niveau où elle est calculée. La repondération consiste simplement à modifier les poids des unités qui ne font pas partie du sous-échantillon pour compenser l'absence de données (que ce soit des données d'enquête

ou des données fiscales) des unités du sous-échantillon. De plus, cette méthode introduisait un nouveau poids aux unités échantillonnées et demandait certains ajustements pour respecter la cohérence entre les différentes variables. À ceci s'ajoutait la difficulté d'implanter cette option en production qui, malgré sa simplicité, nécessitait beaucoup de changements au processus actuel.

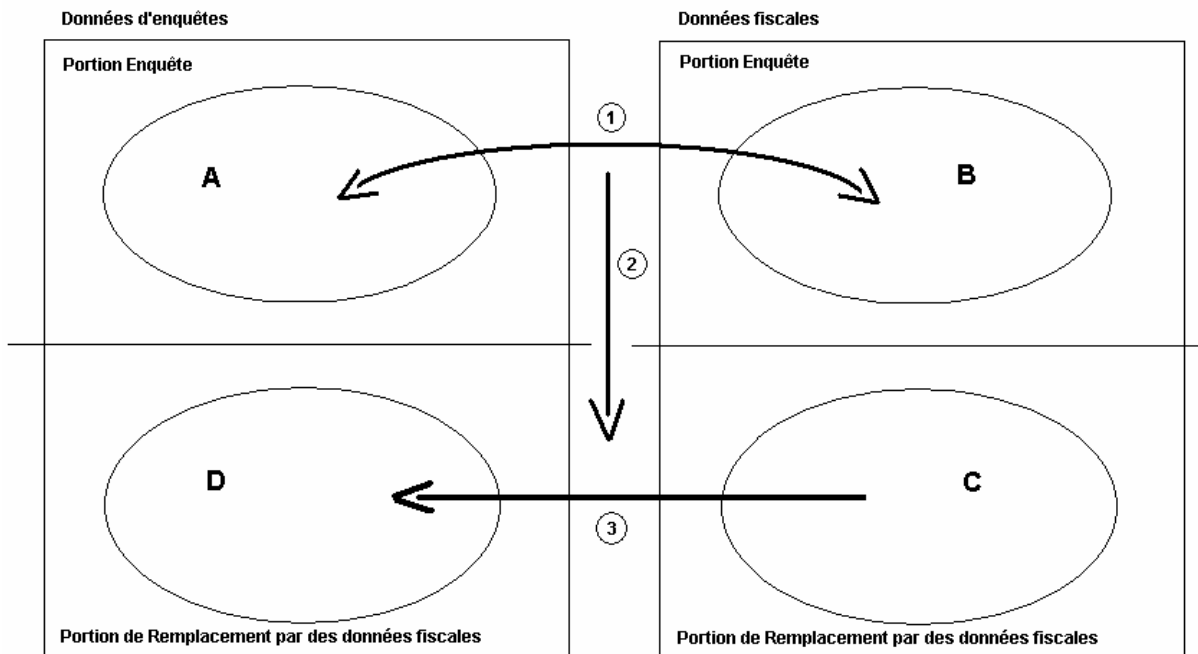
5.4.2 Utilisation directe des données fiscales

La deuxième option est une utilisation directe des données fiscales afin de remplacer les données d'enquête et ce, sans aucune modélisation. Une simulation a donc été effectuée à l'aide des données de l'AR2002. Pour les besoins de l'étude, un sous-échantillon d'essai de 50% d'unités simples a été sélectionné. Ainsi, les données fiscales déclarées ou imputées historiquement ont été utilisées directement pour remplacer les données d'enquête. De nouvelles estimations ont été produites et comparées avec les estimations obtenues en production. En résumé, des différences relatives acceptables entre les nouvelles estimations produites et les estimations obtenues en production pour les variables financières clés ont été obtenues. Cependant, des différences plus élevées pour les autres variables financières ont été observées.

5.4.3 Modélisation des données fiscales

La troisième option considérée a été d'appliquer un modèle aux données fiscales afin de dériver les variables financières et non financières se retrouvant sur le questionnaire. L'idée derrière la modélisation est d'utiliser les unités où les données d'enquête et les données fiscales sont disponibles pour développer différents modèles de régression pour les principales variables financières clés. Ceci est possible étant donné que les données fiscales des unités T2 sont disponibles et accessibles pour toutes les enquêtes, autant au-dessus qu'en dessous des seuils d'exclusion. La modélisation est présentée graphiquement à la figure 3.

Figure 3 : Relation entre les données d'enquête et les données fiscales pour la modélisation



Comme pour la deuxième option (utilisation directe des données fiscales), on a utilisé le même sous-échantillon d'essai. Utilisant des données de l'AR2002, plusieurs modèles de régression établissant une relation entre les parties A et B ont été étudiés. Les modèles de régression multiples semblaient moins performants que les modèles de régression simples. Ce sont finalement des modèles de régression simples qui ont été appliqués. Par exemple, pour la variable du « Total des revenus » (C2098), le modèle suivant a été utilisé :

$$C2098_A = \beta_1 C2098_B$$

où $C2098_A$ représente la donnée d'enquête de la partie **A** et $C2098_B$ représente la donnée fiscale de la partie **B**. Comme ce modèle est obtenu à l'aide des parties **A** et **B**, les données d'enquête et les données fiscales sont donc disponibles (la partie **A** correspond aux unités simples ne faisant pas partie du « sous-échantillon » présenté à la section 5.3 et la partie **B** aux données fiscales correspondantes). Par la suite, le coefficient β_1 est appliqué aux données de la portion de Remplacement par des données fiscales (partie **C** de la figure 3) pour obtenir les valeurs prédites pour la partie **D** de la figure 3 où aucune donnée d'enquête pour l'année de référence n'est disponible (la partie **D** correspondant aux unités simples faisant partie du « sous-échantillon » présenté à la section 5.3). Ainsi, les valeurs prédites pour la variable C2098 sont obtenues de la façon suivante :

$$C2098_D = \beta_1 C2098_C$$

où $C2098_D$ représente la valeur prédite pour les unités de la partie **D**. À noter que des modèles de régression similaires ont été utilisés pour les autres variables financières clés.

Une fois les valeurs prédites obtenues suite à la modélisation, les estimations pour les variables financières clés ont été reproduites et comparées avec les estimations obtenues en production pour l'AR2002. Ces comparaisons ont été effectuées au niveau national, au niveau des provinces et au niveau des codes SCIAN. En résumé, pour ce qui est des variables financières clés, les résultats obtenus sont acceptables. Les différences relatives entre les estimations obtenues en utilisant les modèles de régression et celles obtenues en production oscillent entre 0% et +/-10%. Mais, pour ce qui est des variables financières autres que les variables financières clés, les résultats ne sont pas aussi convaincants. Les différences relatives étant, pour certaines variables, assez élevées entre les « nouvelles » estimations obtenues suite à la modélisation et celles obtenues en production.

On doit également mentionner que les données extrêmes (données d'enquête ou données fiscales) étaient enlevées avant la détermination du coefficient de régression, ainsi que tout enregistrement pour lequel la donnée d'enquête ou la donnée fiscale était nulle.

Une contrainte non-négligeable de l'exercice de modélisation est le temps requis pour développer le nombre de modèles nécessaires pour couvrir toutes les variables et ce, pour chacune des sept enquêtes impliquées dans le projet. Il faut donc développer les modèles en utilisant les données de l'année précédente ou de l'année courante et ceci, avant l'étape de la vérification et de l'imputation.

5.4.4 Stratégie adoptée pour l'AR2003

Suite aux différentes études et en collaboration avec les analystes et de tous les différents intervenants, l'option d'utiliser les données fiscales directement, sans aucune modélisation a été choisie. Des études supplémentaires ont été réalisées par la suite dans le but « d'ajuster » certaines variables financières clés provenant des données fiscales dans le but d'améliorer la qualité des estimations. Les résultats ont démontré que les données fiscales n'avaient pas besoin d'être ajustées. Par conséquent, les données fiscales déclarées et imputées historiquement ont été utilisées directement en remplacement aux données d'enquête pour les variables financières clés pour lesquelles les équivalences sont jugées de bonne qualité par les analystes. Ce sont les analystes qui déterminent si les équivalences entre les deux sources de données sont acceptables ou non. Pour ce qui est : (1) des variables financières qui n'ont pas d'équivalence entre les données d'enquête et les données fiscales, (2) des variables financières où les équivalences ne sont pas de bonne qualité ainsi que (3) pour toutes les variables non financières, de l'imputation massive sera effectuée pour toutes ces variables.

6. DÉVELOPPEMENTS FUTURS

Les sources administratives de données fiscales peuvent représenter un mode de collecte alternatif. Ces données peuvent aussi être utilisées à des fins d'imputation et de modélisation. Étant donné l'utilisation croissante des données fiscales, il devient important de définir les différentes sources d'erreurs (effet de mode, imputation, modélisation) associées à leur utilisation.

Des études ont été réalisées dans le but de calculer la variance due à l'utilisation des données fiscales dans un contexte de remplacement direct. Pour ce faire, le logiciel SEVANI (Système pour l'estimation de la variance due à la non-réponse et à l'imputation) développé à Statistique Canada a été utilisé. Pour plus de détails sur l'intégration du calcul de la variance due à l'imputation, voir Nadeau (2004).

Pour l'AR2003, le projet de remplacement des données d'enquête par des données fiscales n'a touché que les unités simples et uniquement sept enquêtes de l'EUE. Pour l'AR2004, cinq enquêtes supplémentaires se sont jointes au projet. Par conséquent, des études seront entreprises dans le but de vérifier quel est l'impact sur les estimations d'une utilisation de plus en plus grande de données fiscales au cours des prochaines années.

7. CONCLUSION

En conclusion, l'intégration du Projet de remplacement des données d'enquête par des données fiscales à l'EUE s'est bien déroulée lors des premières étapes de l'enquête. Les principaux objectifs ayant été atteints sont :

- La production des estimations des variables financières et non financières;
- Une plus grande utilisation des données fiscales au-dessus des seuils d'exclusion;
- Une réduction des coûts des enquêtes;
- Une réduction du fardeau de réponse des entreprises.

Pour ce qui est du dernier objectif qui est une amélioration potentielle ou un maintien de la qualité des données, plusieurs analyses seront effectuées sur les données de l'AR2003 pour vérifier l'impact, sur les estimations, d'une plus grande utilisation des données fiscales.

REMERCIEMENTS

L'auteur tient à remercier Sylvie Gauthier pour son aide précieuse pour la production de l'article ainsi que Hélène Bérard et Claude Nadeau pour leurs commentaires pertinents. Il tient aussi à remercier Yanick Beaucage et Mike Sirois pour leur aide à la révision de ce document.

RÉFÉRENCES

- Laniel, N., Royce, D. (1998). Projet d'amélioration des statistiques économiques provinciales : objectifs et enquête-pilote. Société Statistique du Canada, Recueil 1998 de la Section des méthodes d'enquête, 59-63.
- Nadeau, C. (2004). Challenges Associated with the Increased Use of Fiscal Data for the Unified Enterprise Survey. Joint Statistical Meetings, Recueil 2004.

