

## DE 7 À 26 ENQUÊTES – COMPLEXITÉ GRANDISSANTE DE LA GESTION DE L'ÉCHANTILLONNAGE POUR L'ENQUÊTE UNIFIÉE AUPRÈS DES ENTREPRISES

Sylvie Gauthier<sup>1</sup>

### RÉSUMÉ

L'Enquête unifiée auprès des entreprises (EUE) a été développée en 1997. Sept enquêtes en faisaient alors partie. Depuis, une vingtaine d'enquêtes supplémentaires se sont jointes à l'EUE. Chaque année, des modifications sont apportées au plan de sondage pour le rendre plus efficace et pour répondre à des besoins spécifiques : utilisation accrue des données fiscales, intégration de nouvelles enquêtes, etc. La gestion du plan de sondage d'une telle enquête, où les intervenants et groupes de travail sont nombreux, est un réel défi. Tout changement proposé doit être bien analysé pour en mesurer les répercussions sur toutes les étapes de l'enquête. La mise en oeuvre de ces modifications requiert une grande coordination et communication entre les personnes impliquées pour en assurer le succès. Du côté de l'environnement technique, le nombre grandissant d'enquête dans l'EUE a nécessité une revue des programmes informatiques afin de les rendre plus flexibles, conviviaux et rapides d'exécution. Cet article, après avoir fait un survol du plan d'échantillonnage et des changements méthodologiques importants survenus dans l'EUE récemment, se concentrera principalement sur les avenues méthodologiques des dernières années ainsi que les défis de gestion sous-jacents à une enquête complexe dont les besoins évoluent rapidement.

### 1. INTRODUCTION

L'Enquête unifiée auprès des entreprises (EUE) a été développée en 1997. Elle est le principal véhicule de production des estimations annuelles pour plusieurs secteurs d'activité économique : statistique du commerce, de l'industrie de service, du secteur du transport, de l'agriculture et du secteur manufacturier<sup>2</sup>. Le système de comptabilité nationale détermine à partir de ces estimations les comptes sur les revenus, sur les dépenses ainsi que sur les entrées et sorties qui serviront à mesurer la production et la consommation totales de biens et services dans les provinces et territoires (Smith, 1998). En sept ans, le nombre d'enquêtes intégrées à l'EUE a triplé. Ce nombre croissant de secteurs, ainsi que l'initiative d'utiliser davantage de données fiscales en remplacement de données d'enquête, a engendré de nouveaux défis. Cet article, après avoir fait un survol du plan d'échantillonnage et des changements importants survenus dans l'EUE depuis sa création, se concentrera principalement sur les avenues considérées du point méthodologique ainsi que les défis de gestion sous-jacents à une enquête complexe dont les besoins évoluent rapidement.

La section 2 décrit brièvement le plan d'échantillonnage de l'EUE. La section 3 met l'emphase sur les changements apportés au cours des dernières années, principalement ceux reliés à l'utilisation des données fiscales. La section 4 discute de la gestion matricielle de l'EUE et brosse un tableau des répercussions de toutes les ramifications du plan sur la gestion du projet.

### 2. LE PLAN D'ÉCHANTILLONNAGE DE L'EUE

Le plan d'échantillonnage de l'EUE de Statistique Canada a été développé pour satisfaire les objectifs suivants: obtenir des données provinciales plus détaillées pour divers domaines économiques, uniformiser les méthodes, recourir davantage aux données fiscales et assurer une gestion proactive du fardeau de réponse.

---

<sup>1</sup> Sylvie Gauthier, Statistique Canada, RHC11-J Tunney's Pasture, Ottawa, Ontario, KIA 0T6 ([sylvie.gauthier@statcan.ca](mailto:sylvie.gauthier@statcan.ca))

<sup>2</sup> Le secteur manufacturier est traité différemment des autres secteurs de l'EUE. La méthodologie présentée ici s'applique principalement aux autres secteurs.

La base de sondage utilisée est le Registre des entreprises (RE), le système de base de données de Statistique Canada. Le RE contient toutes les entreprises non-incorporées (T1) et incorporées (T2) connues opérant au Canada. Sur le RE, la structure de chaque entreprise est hiérarchisée en fonction des besoins statistiques et comporte quatre niveaux qui sont dans l'ordre l'entreprise, la compagnie, l'établissement et l'emplacement. Une entreprise se compose d'une ou plusieurs compagnies. Une compagnie se compose d'un ou plusieurs établissements et ainsi de suite. Pour plusieurs entreprises canadiennes, les 4 entités coïncident, c'est ce qu'on appelle une structure simple. Chaque mois, l'Agence du Revenu du Canada fait parvenir à Statistique Canada les données fiscales associées à ces entreprises. En 2002, une initiative a été lancée pour augmenter davantage l'utilisation de ces données pour les unités simples. Pour cette raison, le plan d'échantillonnage des récentes années tient compte de la disponibilité des données fiscales.

La population de l'EUE est stratifiée en fonction de la géographie, de l'industrie et de la taille des entreprises. L'unité d'échantillonnage est composée de l'ensemble des établissements appartenant à la même entreprise au sein de la même cellule. La cellule est définie comme le groupement de tous les établissements ayant des activités dans la même province ou territoire, et au même niveau d'agrégation du Système de Classification des Industries de l'Amérique du Nord (SCIAN).

Dans chaque cellule, un seuil d'exclusion basé sur le revenu de l'unité d'échantillonnage délimite la portion de la population qui est éligible à être enquêtée. Pour les entreprises sous le seuil d'exclusion, des données fiscales sont utilisées pour créer les estimations afin de réduire le fardeau de réponse des petites entreprises. Ces seuils d'exclusion sont calculés de façon à assurer qu'au maximum 10% du revenu total de la cellule soit estimé par les données fiscales. Étant donné le caractère asymétrique des populations, cette méthodologie permet d'exclure beaucoup de petites unités. Cette partie sous les seuils est communément appelée strate à tirage nul (TN).

La taille de l'échantillon de chaque cellule est ensuite définie en utilisant l'algorithme de Lavallée-Hidiroglou (voir les détails dans Simard, Girard, Parent & Smith, 2001). Cet algorithme définit les bornes de strates en fonction d'une précision visée tout en minimisant la taille de l'échantillon. Trois strates sont créées par cellule: une strate à tirage complet (TC) et deux strates à tirage partiel (TP). La répartition de l'échantillon à travers les strates de la cellule se fait ensuite de façon proportionnelle à la racine carrée du revenu total entre les trois strates. Cette taille initiale obtenue est ensuite gonflée pour tenir compte des imperfections de la base de sondage et de la non-réponse. Un échantillon aléatoire simple est ensuite sélectionné dans chaque strate.

Ces grandes lignes de la méthodologie sont communes à toutes les 26 enquêtes. Toutefois, il existe une diversité au niveau de l'application de cette méthodologie. Le niveau de stratification, le type d'échantillon (recensement ou enquête), le type de rotation (chevauchement maximal ou indépendant), le nombre de bornes utilisé pour déterminer le seuil d'exclusion, le niveau d'exclusion et les taux de sur-échantillonnage<sup>3</sup> diffèrent par enquête. Les stratégies par enquête sont variées pour mieux satisfaire les besoins en données. Toutefois, d'un point de vue opérationnel et de gestion, ces différences complexifient la situation.

### **3. MODIFICATIONS AU PLAN D'ÉCHANTILLONNAGE**

Des modifications sont apportées annuellement au plan d'échantillonnage pour diverses raisons : améliorer le processus actuel, intégrer des nouvelles enquêtes ou modifier les programmes actuels pour répondre à de nouvelles exigences, telle l'utilisation accrue des données fiscales.

#### **3.1 Améliorations au processus actuel**

Le plan initial de l'EUE a été amélioré au fil des années pour être plus efficace et répondre plus adéquatement aux besoins. Les principaux changements sont : i) l'introduction d'un module de retrait des unités inactives pour assurer la représentativité de l'échantillon pour les enquêtes ayant un chevauchement maximal des échantillons d'une année à l'autre; ii) l'utilisation d'une méthode itérative pour allouer l'échantillon en fonction d'une précision visée mais d'une taille fixe d'échantillon; et iii) l'harmonisation de la méthodologie employée pour déterminer la variable de taille des

---

<sup>3</sup> Les taux de sur-échantillonnage sont utilisés pour compenser pour la non-réponse, les unités inactives et les unités classées de façon erronée.

enquêtes sur le commerce de gros et le commerce de détail afin de favoriser la cohérence entre les résultats de l'enquête annuelle et l'enquête mensuelle.

### 3.2 Intégration de nouvelles enquêtes

Chaque année, de nouvelles enquêtes s'intègrent à l'EUE. Les programmes informatiques sont modifiés pour les inclure. Étant donné le nombre grandissant d'enquêtes, il y a eu création de programmes supplémentaires pour permettre des changements rapides à une seule enquête.

### 3.3 Impact de l'utilisation accrue des données fiscales

L'utilisation accrue des données fiscales dans l'EUE a revêtu une grande importance puisqu'elle a fait l'objet d'une initiative de rationalisation stratégique de Statistique Canada. Depuis 1997, les données fiscales étaient utilisées de manière restreinte: on produisait à partir de celles-ci des estimations pour les très petites entreprises, et on les utilisait principalement dans le processus de vérification et d'imputation ainsi que pour la validation de données d'enquêtes. Depuis 2002, les données fiscales sont utilisées en remplacement des données d'enquêtes; ceci dans l'optique de réduire le fardeau de réponse et les coûts de l'enquête. Jusqu'à ce jour, l'utilisation des données fiscales s'est donc fait au niveau des microdonnées, cette approche étant plus facile à mettre en place.

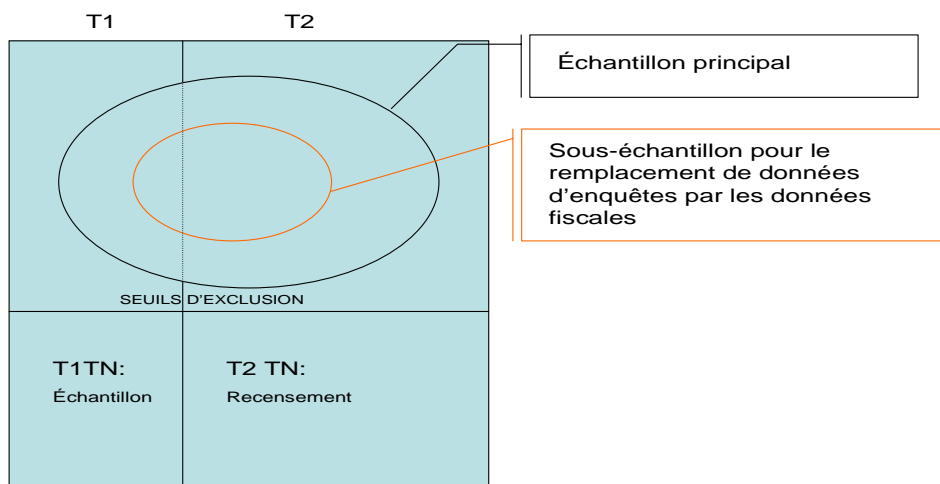
#### 3.3.1 Année de référence 2002

La première action qui a été prise pour augmenter l'utilisation des données fiscales fut de hausser les seuils d'exclusion des cellules de 5% à 10%, ceci a permis de réduire le nombre des unités enquêtées d'environ 4 000 pour atteindre une taille d'échantillon de 44 000. Mentionnons également qu'au moment de la collecte de 2002, il fut décidé pour certaines enquêtes de remplacer par des données fiscales toutes les unités simples provenant des petites strates à tirage partiel. Les données fiscales ont aussi été utilisées pour obtenir des données pour les non-répondants chroniques afin de diminuer davantage le fardeau de réponse.

#### 3.3.2 Année de référence 2003

L'objectif principal en 2003 était de réduire d'environ 5 000 le nombre d'unités simples enquêtées par rapport à 2002. Seules sept enquêtes ont été ciblées par ce procédé. Pour ce faire, une fois l'échantillon principal tiré, un sous-échantillon des unités simples était sélectionné afin que les données fiscales soient utilisées en remplacement des données d'enquête (voir Figure 1).

Figure 1 : Plan de sondage de l'EUE



Divers scénarios ont été étudiés et discutés afin de déterminer comment allouer le sous-échantillon de 5 000 unités. Dans un premier temps, il fallait décider si les entreprises simples T1 et T2 étaient éligibles au remplacement de

données fiscales. Puisque peu d'études existaient alors sur la relation conceptuelle entre les données fiscales T1 et les données enquêtées correspondantes et que le bassin des entreprises simples T2 était suffisant pour atteindre l'objectif des 5 000 unités, seuls les T2 ont été sous-échantillonnées pour la majorité des enquêtes.

Un taux de remplacement de 50% des unités simples échantillonnées a été recommandé. Ce taux assurait un nombre suffisant de données d'enquête pour permettre l'imputation des variables non-disponibles à partir des dossiers fiscaux. Un certain nombre de répondants était aussi nécessaire afin de permettre la détermination d'un taux d'imperfection de la base compte tenu du manque de rétroaction d'enquête. D'autres scénarios ont ensuite été produits pour déterminer, selon la taille de chaque échantillon, si seules les unités provenant des strates à tirage partiel devaient faire partie du sous-échantillon ou si les unités des strates à tirage complet devaient également être éligibles au remplacement. Les décisions finales ont varié par enquête, selon les besoins en données et la contribution des unités simples dans la population.

Une fois le taux de remplacement par des données fiscales déterminé, ainsi que les critères d'éligibilité finalisés, une question demeurait : comment répartir le sous-échantillon? Des études ont été effectuées pour mesurer l'impact de choisir une répartition proportionnelle versus une répartition de Neyman. L'avantage de la répartition proportionnelle est qu'elle permet d'avoir un nombre égal de données fiscales et d'enquête dans chacune des strates. Ceci est avantageux au niveau de la vérification et de l'imputation pour assurer un plus grand nombre de donneurs potentiels, surtout s'il existe des différences considérables entre les strates à tirage partielles, petites et moyennes. Pour ce qui est de la répartition de Neyman, elle choisit plus d'unités à être enquêtées dans les strates plus variables (les strates à tirage complet). Encore une fois, la répartition a varié d'une enquête à l'autre selon les besoins des clients.

### **3.3.3 Année de référence 2004**

En 2004, une méthode plus sophistiquée a été utilisée pour déterminer la répartition de l'échantillon entre les données d'enquêtes et les données fiscales. L'objectif était d'utiliser les données fiscales pour au moins 1500 unités de plus comparativement à 2003.

Pour ce faire, les variables clefs de chaque enquête<sup>4</sup> ont été utilisées pour dériver des scénarios de répartition entre les données d'enquête et les données fiscales. Les paramètres utilisés comprenaient des coûts de collecte et de traitement, des différences conceptuelles existant entre les variables sur les questionnaires et les dossiers fiscaux et de la variabilité de chacune des sources de données. Pour quantifier les différences conceptuelles, le coefficient de détermination ( $R^2$ ) a été utilisé. Il a été obtenu en modélisant les données d'enquête rapportées en fonction des données fiscales. Les données fiscales avaient préséance sur les données d'enquête quand les différences conceptuelles étaient mineures puisque leur coût est plus bas.

Après analyses et discussions, il fut décidé que la répartition finale serait basée sur la variable profit. La même variable a été utilisée pour toutes les enquêtes pour assurer une uniformité dans la méthode. Le taux de répartition a été assez similaire d'une enquête à l'autre, l'approche utilisée étant assez conservatrice. Les taux de remplacement ont été légèrement supérieurs à ceux de 2003, et encore une fois, le sous-ensemble d'unités visé variait d'une enquête à l'autre en fonction des besoins des clients.

## **4. GESTION DE L'EUE**

### **4.1 Gestion matricielle de l'EUE et comité du plan de sondage**

L'EUE est gérée selon une approche matricielle. La gestion de l'enquête relève principalement de la Division de la statistique des entreprises (DSE). La DSE planifie, coordonne et surveille les activités de l'EUE par le biais de comités interdisciplinaires. La Division de la méthodologie est responsable de donner du soutien méthodologique à tous les intervenants des différents secteurs couverts par l'EUE. C'est dans ce cadre de travail qu'a été créé le comité du plan de sondage, dont le mandat est d'identifier et de discuter des questions méthodologiques reliées à l'EUE, ainsi que de recommander, de mettre en oeuvre et de documenter le plan d'échantillonnage. La coordination des activités de méthodologie avec toutes les autres activités de l'EUE, telles la collecte, la vérification et l'imputation, est

---

<sup>4</sup> Revenu d'opération total, revenu total, dépenses du travail total, profit, dépréciation et amortissement, coûts des biens vendus.

fondamentale pour s'assurer que les répercussions sur les autres étapes de l'enquête de tout changement proposé sont analysées.

## **4.2 Gestion des changements**

En fonction des nouveaux besoins, la stratégie d'échantillonnage et les programmes utilisés pour la sélection sont revus chaque année. La sélection de l'échantillon repose sur la soumission d'une vingtaine de programmes informatiques, chacun ayant un rôle différent : définition des enquêtes, stratification, détermination de la précision, répartition, etc. La sélection des 26 échantillons se fait simultanément. La gestion des travaux d'analyse et des modifications temporaires apportés aux programmes en phase de développement est un défi en soi. La communication entre les membres de la méthodologie est un élément essentiel à la réussite de ce projet.

Le nombre grandissant d'enquêtes a eu plusieurs répercussions : i) une plus grande source d'information à gérer et à implémenter due aux différences notables entre les 26 enquêtes, ii) une augmentation du nombre d'économistes avec qui interagir, et par conséquent, une augmentation des requêtes, iii) une augmentation du temps de soumissions des programmes, iv) une création d'un plus grand nombre d'outils diagnostic pour assurer l'intégrité de l'échantillon et de son niveau de qualité et v) une nécessité de gérer judicieusement les requêtes pour y répondre dans un temps raisonnable.

Il est bon également de mentionner que le développement du plan de sondage de l'année courante se fait simultanément avec la production des estimations de l'année précédente. Ceci ajoute un niveau supplémentaire de gestion de l'information afin d'incorporer si nécessaire des corrections aux nouvelles procédures afin de tenir compte de l'évaluation du procédé de l'année précédente. Voir Nadeau (2004) et Pelletier (2004) pour une description des défis reliés à l'utilisation des données fiscales lors de l'estimation.

## **5. CONCLUSION**

Les défis des prochaines années seront d'adapter le plan d'échantillonnage actuel afin de tenir compte des changements importants à venir à Statistique Canada soit l'extension de l'utilisation des données fiscales, la refonte du Registre des entreprises, le nouveau système de classification qui entrera en vigueur en 2007, et l'intégration de nouvelles enquêtes dans l'EUE. Ainsi, l'initiative d'utiliser les données fiscales se poursuivra et évoluera en fonction des apprentissages faits les années précédentes. Des analyses sont en cours afin d'évaluer la possibilité d'utiliser les données fiscales afin de faire du remplacement direct pour les entreprises ayant une structure complexe. Des projets de recherche sur la possibilité de calibrer en utilisant les données fiscales lors de l'estimation sont aussi amorcés. Le défi résidera ainsi à trouver un équilibre acceptable entre deux souhaits : la quête d'une stabilité du plan d'échantillonnage et l'ambition de répondre aux besoins spécifiques changeants des enquêtes.

## **REMERCIEMENTS**

L'auteure tient à remercier Hélène Bérard, Isabelle Marchand, Marie Brodeur, Claude Nadeau et Eric Pelletier pour leur aide à la révision de ce document.

## **RÉFÉRENCES**

- Nadeau, C. (2004), «Challenges associated with the increased use of fiscal data for the Unified Enterprise Survey », article présenté à la rencontre annuelle du Joint Statistical Meetings, Toronto, Canada.
- Pelletier, E. (2004), «L'utilisation accrue des données fiscales dans le cadre de l'enquête unifiée sur les entreprises », article présenté à la rencontre annuelle de la Société statistique du Canada, Montréal, Canada.
- Simard, M., Girard, C., Parent, M-N., et Smith, J (2001), « Les plans d'échantillonnage des enquêtes unifiées sur les entreprises: Les premières années », rapport non publié, Ottawa, Canada: Statistique Canada.
- Smith, Philip (1998), « Trousse d'information sur le PASEP », rapport non publié, Ottawa, Canada : Statistique Canada.