

## **TRAITEMENT DES DONNÉES FISCALES POUR LE REMPACEMENT DES DONNÉES D'ENQUÊTES**

Nathalie Hamel<sup>1</sup>

### **RÉSUMÉ**

L'Agence du revenu du Canada (ARC) partage certaines données avec Statistique Canada (SC) dont celles sur les états financiers pour les entreprises incorporées (T2). Ces renseignements provenant de la déclaration financière, du bilan et de l'état des résultats des entreprises sont utilisés à SC, entre autres, pour remplacer les données d'enquêtes économiques. Les buts sont de réduire le fardeau de réponse et les coûts des enquêtes, ainsi que de potentiellement améliorer la qualité des données. Les données T2 proviennent mensuellement de l'ARC et sont disponibles sous le format de l'Index général des renseignements financiers (IGRF). SC reçoit donc un recensement des données T2 qui contiennent environ 700 variables.

Lorsqu'elles arrivent à SC, les données T2 sont soumises à des règles de contrôle dans le but de préparer une base de données nettoyée et complète pour des fins statistiques. Puisque le remplacement des données d'enquête a lieu avant la réception complète de l'univers T2, les données sont également soumises à des procédures d'imputation. La désagrégation des champs dits « génériques » aux détails, l'imputation de la variable « salaire et traitement », l'imputation par le quotient à partir de données historiques et l'imputation par donneur selon le plus proche voisin sont quelques-unes des procédures appliquées aux données. Dans cet article, on présente un survol des procédures de contrôle et d'imputation effectuées en vue de préparer les données fiscales pour le remplacement des données d'enquêtes.

### **1. INTRODUCTION**

L'Agence du revenu du Canada (ARC) partage certaines données avec Statistique Canada (SC) dont celles sur les états financiers pour les entreprises incorporées (T2). Ces renseignements, provenant de la déclaration financière, du bilan et de l'état des résultats des entreprises, sont utilisés à SC, entre autres, pour remplacer certaines données d'enquêtes économiques. Les buts sont notamment de réduire le fardeau de réponse et les coûts des enquêtes, ainsi que de potentiellement améliorer la qualité des données. Les données financières T2 proviennent mensuellement de l'ARC et sont disponibles sous le format de l'Index général des renseignements financiers (IGRF). SC reçoit, pour chaque année de référence, un recensement des données T2 qui représentent 1,3 million de sociétés et contiennent environ 700 variables.

Lorsqu'elles arrivent à SC, les données T2 sont soumises à des règles de contrôle dans le but d'obtenir une base de données nettoyée et complète (incluant certaines imputations partielles) pour des fins statistiques. Ces données reçues sont alors acheminées mensuellement, tout dépendant de la fin de l'année financière de l'entreprise, aux représentants des enquêtes économiques de SC. Plusieurs enquêtes économiques substituent les données recueillies traditionnellement auprès des répondants par des données fiscales, entre autres, les données T2. Puisque le remplacement des données d'enquête a lieu avant la réception de l'univers T2 au complet, les données non reçues sont soumises à des procédures d'imputation. Cet univers des enquêtes économiques composé des T2 reçus et imputés est alors partagé une fois par année avec les représentants des enquêtes économiques de SC.

Dans cet article, on présente un survol des procédures appliquées aux données fiscales en vue de les préparer pour le remplacement des données d'enquêtes. À la section 2, on décrit l'univers des données T2. À la section 3, on présente les règles de contrôle appliquées aux données à SC. À la section 4, on décrit le

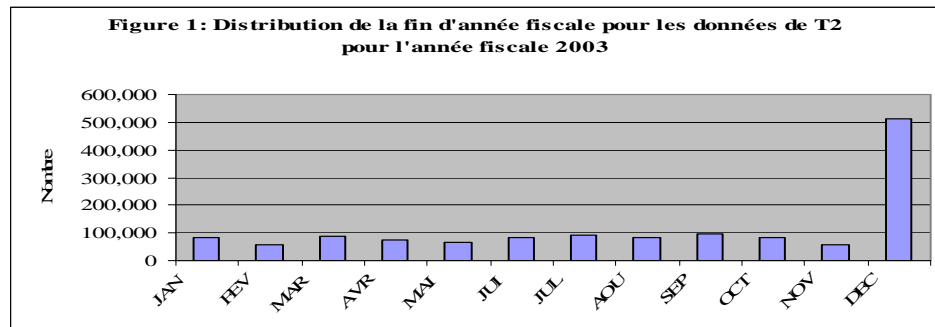
---

<sup>1</sup>Nathalie Hamel, Statistique Canada, 11<sup>e</sup> étage, section R, Immeuble R.-H.-Coats, Ottawa, Ontario, Canada, K1A 0T6 (nathalie.hamel@statcan.ca)

processus de désagrégation des champs génériques aux détails. À la section 5, on présente les procédures d'imputation. À la section 6, on décrit le plan comptable appliqué aux données T2 pour obtenir un format standard des variables de remplacement pour les enquêtes économiques. On termine finalement par une brève conclusion.

## 2. UNIVERS T2

L'univers des données T2 est apparié à celui des enquêtes économiques, représentant ainsi 1,3 million de sociétés ou enregistrements. Parmi ces enregistrements, environ 1 million parviennent à SC. La période de référence pour les données T2 est, dans la plupart des cas, annuelle. La fin de l'année fiscale peut arriver n'importe quand durant l'année calendrier. La figure 1 qui suit présente la distribution de la fin d'année fiscale pour les entités légales pour l'année fiscale 2003. Lorsque la période fiscale est terminée, les entreprises ont jusqu'à six mois pour remplir leur déclaration fiscale.



Les données T2 reçues de l'ARC sont divisées en plusieurs annexes et formulaires. On reçoit, entre autres, les états des résultats, les bilans, le formulaire 200 « Déclaration de revenus des sociétés », l'annexe 1 « Revenu/perte net aux fins de l'impôt sur le revenu » ainsi que l'annexe 5 « Calcul supplémentaire de l'impôt - sociétés ». D'autres annexes sont également reçues de l'ARC mais elles ne sont pas encore soumises aux règles de contrôle et aux processus d'imputation, et ainsi, elles ne sont pas couvertes dans ce document.

L'état des résultats et le bilan représentent l'information financière. Les états des résultats sont divisés en quatre sections, soient les revenus et dépenses, agricoles et non agricoles. Pour leur part, les bilans sont divisés en trois sections, soient les actifs, le passif et l'avoir. Ces deux formulaires possèdent environ 700 variables au total. La plupart des sociétés rapportent de 30 à 40 variables, dont seulement 8 sont obligatoires pour l'ARC, soient les totaux de section ainsi que le profit/perte net. L'information fiscale provient du formulaire 200 ainsi que des annexes 1 et 5.

## 3. RÈGLES DE CONTRÔLE

Les données T2 sont soumises à différentes règles de contrôle dans le but de préparer une base de données nettoyée avant de la partager avec les représentants des enquêtes économiques. On applique des règles déductives pour dériver au besoin certains champs manquants et s'assurer que les données d'un même enregistrement sont en équilibre. Une de ces règles s'assure, entre autres, que le revenu moins les dépenses soit égal au profit/perte net, à plus ou moins 1 000 \$. D'autres règles de ce genre sont également appliquées. Pour satisfaire le remplacement des données d'enquête par les données fiscales, de nouvelles règles ont également été développées. Les inventaires manquants des états des résultats sont imputés à partir des inventaires des bilans (les inventaires de fermeture des états des résultats si la valeur provient de l'année courante et les inventaires d'ouverture si la valeur provient de l'année précédente). Les amortissements manquants des états des résultats sont dérivés à l'aide de l'annexe 1. Certaines valeurs négatives sont transférées dans d'autres champs dans le but d'obtenir des valeurs positives et ainsi simplifier les processus ultérieurs. Les champs communs entre les différents formulaires et annexes sont comparés et ajustés au besoin pour s'assurer de la cohérence entre ceux-ci. Les valeurs aberrantes sont

détectées selon la méthode de Hidioglou-Berthelot (1986) et les plus importantes sont corrigées manuellement au besoin.

Une nouvelle règle a également été développée pour palier au manque d'information pour le salaire et traitement, qui s'avère être une variable clé pour les usagers. Comme cette variable constitue une forme de dépense pour les entreprises et qu'elle n'est pas obligatoire à l'ARC, l'information est souvent incluse dans un autre champ de dépenses plus générique (se référer à la section suivante pour de plus amples détails sur les champs génériques). Douze champs génériques ont été identifiés comme sources potentielles pouvant contenir cette valeur. Des études ont été effectuées pour trouver différentes sources de données fiables qui permettraient de déterminer la valeur ajoutée aux champs génériques et qui devrait être transférée au salaire et traitement. On a identifié cinq sources de données pouvant contenir l'information, telles que les annexes 5 et 27 « Calcul de la déduction pour des bénéficiaires de fabrication et de transformation au Canada », les états des résultats des années précédentes (une ou deux années précédentes), le fichier RC107 « fichier administratif contenant les déductions à la source des entreprises » et le fichier T4 « fichier administratif sur l'état de la rémunération payée ». La règle pour imputer la valeur au salaire et traitement est la suivante : lorsque cette variable est nulle, on identifie la valeur attribuée à cette variable selon les autres sources de données. Lorsque cette valeur entière peut être déduite d'un des douze champs génériques, on soustrait ce montant du champ générique en question et on impute ce montant à une des trois variables associées au salaire et traitement. Ces trois champs sont le montant des coûts des salaires directs, le montant des dépenses agricoles de salaires et traitements non spécifiées, ainsi que le montant des dépenses de salaires et traitements non spécifiées. Le champ générique d'où provient le montant détermine lequel des trois champs sera imputé.

#### 4. DÉSAGRÉGATION DES CHAMPS GÉNÉRIQUES AUX DÉTAILS

Un problème de manque de détails découle du fait que seulement huit champs sont obligatoires à l'ARC et que certains répondants ont tendance à utiliser davantage les champs génériques. Puisque les détails sont nécessaires pour le remplacement des données d'enquêtes, une méthodologie a donc été développée pour désagréger les données génériques aux détails. Le répondant peut divulguer ses données selon trois scénarios, soit en donnant seulement une donnée générique, seulement les détails, ou un mélange des deux. Le tableau 1 nous montre les trois scénarios possibles pour un bloc de variables.

Type de champ	Champ	Définition du champ	Scénario 1	Scénario 2	Scénario 3
Générique	8810	Frais de bureau	100		25
Détails	8811	Papeterie & fournit. de bureau		50	25
	8812	Services de bureau		25	25
	8813	Traitement des données		25	25

Le but ici est de désagréger les données génériques de manière à obtenir les détails pour toutes les variables. Les classes d'imputation sont formées à partir du niveau à 2 chiffres du code du Système de classification de l'industrie de l'Amérique du Nord (SCIAN), niveau associé aux industries, ainsi que de trois groupes de revenu (*petit* si inférieur à 5 millions de dollars, *moyen* si entre 5 et 25 millions de dollars et *grand* si supérieur à 25 millions de dollars) pour chacun des blocs de variables. Les ratios sont calculés dans chacune des classes d'imputation à partir des enregistrements reçus pour lesquels seulement les détails sont déclarés. Ces ratios sont par la suite distribués aux données contenant seulement des données génériques ou un mélange de données génériques et de détails à l'intérieur des classes d'imputation.

#### 5. IMPUTATION

Une fois par année, les données T2 sont soumises aux différents processus d'imputation dans le but de produire une base de données complète représentant l'univers des enquêtes économiques. Ainsi en septembre de l'année  $t$ , l'univers composé des données T2 de l'année fiscale  $t-1$  doit être disponible aux représentants des enquêtes. Les données non reçues avant le mois de juillet au temps  $t$  ainsi que les données reçues mais qui ont échoué les règles de contrôle déductives sont imputées. Environ 160 000

enregistrements échouent les règles de contrôle déductives annuellement et de 400 000 à 500 000 enregistrements sont imputés au total. Le taux d'imputation est relativement élevé, soit près de 40%.

Le processus de désagrégation des données génériques aux détails est d'abord appliqué aux enregistrements qui ont échoué les règles de contrôle déductives. Par la suite, les autres enregistrements qui ont échoué les règles déductives et ceux non reçus sont imputés par le quotient à partir de données historiques. Pour terminer, les enregistrements résiduels sont imputés par donneur selon le plus proche voisin. Le tableau 2 nous montre la distribution des enregistrements imputés selon chacune des méthodes considérées.

<b>Tableau 2 : Distribution des enregistrements imputés selon la méthode d'imputation</b>	
Taux d'imputation	Méthode d'imputation
1,5% à 2,3%	Par le processus de désagrégation des champs génériques aux détails
1,1%	Par le quotient pour le bilan seulement
1,7%	Par le quotient pour l'état des résultats seulement
33,5%	Par le quotient pour l'information financière au complet
Moins de 1%	Par donneur pour le Coût des marchandises vendues seulement
3,6%	Par le donneur pour l'information financière au complet

L'information financière est d'abord imputée suivie de l'information fiscale, la méthode d'imputation de la première influant sur la méthode de la seconde. L'imputation par le quotient est la méthode la plus fréquemment utilisée. La méthodologie reliée aux deux méthodes d'imputation est présentée dans les sous-sections suivantes.

### **5.1 Imputation par le quotient à partir de données historiques**

On identifie d'abord les enregistrements qui doivent être imputés en déterminant lesquels font partie de l'univers des enquêtes économiques. Ensuite, on exclue les sociétés potentiellement décédées ou inactives et les données non T2. Parmi ces enregistrements, on identifie lesquels possèdent de l'information pour l'année précédente qui a été, soit déclarée, soit imputée par le quotient, ou encore imputée par donneur où les règles de post-imputation sont les plus strictes.

On détermine les classes d'imputation à partir du code SCIAN. On débute au niveau le plus détaillé, soit le niveau à 6 chiffres. Si on obtient 50 enregistrements et plus à l'intérieur de ce groupe, le groupe est alors déterminé. Sinon, on agrège au niveau à 5 chiffres du code SCIAN. Dès que l'on obtient 50 enregistrements et plus à ce niveau, le groupe est également déterminé. Sinon, on recommence en agrégeant au niveau à 4 chiffres et ainsi de suite, jusqu'à ce que tous les enregistrements soient distribués dans une classe d'imputation.

On calcule, par la suite, les tendances en s'assurant d'exclure au préalable les valeurs aberrantes, selon la méthode de Hidiroglou-Berthelot (1986). La tendance est calculée à partir du total pour l'année  $t$  divisé par le total de l'année  $t-1$  à l'intérieur de la classe d'imputation. Une tendance est calculée pour chacune des variables utilisées, soient le revenu, les dépenses, les actifs et le passif.

Pour un enregistrement donné et pour l'information financière, on impute les données de l'année courante à partir des données de l'année précédente ajustées par la tendance calculée sur la variable et la section associée. Si nécessaire, on impute l'information fiscale manquante de la même manière mais sans tendance et finalement, on s'assure que les données sont en équilibre.

### **5.2 Imputation par donneur selon le plus proche voisin**

On identifie d'abord les enregistrements qui doivent être imputés en déterminant lesquels font partie de l'univers des enquêtes économiques. Ensuite, on exclue les sociétés potentiellement décédées ou inactives, les données non T2, ainsi que les enregistrements imputés par le quotient à partir des données historiques.

Pour l'information financière, on détermine les variables d'appariement à partir de différentes sources, soient les données T2 courantes et antérieures, les fichiers administratifs de 1998 et 1999 ainsi que les données du Registre des entreprises (RE). Les variables d'appariement pour l'imputation partielle sont le revenu, les dépenses et le coût des marchandises vendues provenant des données T2 courantes. Les variables d'appariement pour l'imputation complète sont le revenu, les actifs et le ratio revenu-dépenses à partir de toutes les sources énumérées précédemment. Puisque la variable dépenses n'est pas disponible sur le RE, elle est modélisée à partir de la variable revenu. Le même donneur que pour l'information financière est utilisé pour imputer l'information fiscale manquante. Lorsque seule l'information fiscale nécessite une imputation, les variables d'appariement sont alors le revenu, les dépenses et le profit/perte net provenant des données T2 courantes. Des tendances sont calculées selon la méthode présentée pour l'imputation par le quotient à partir des données historiques pour représenter l'année courante lorsque les données antérieures sont utilisées. Les tendances sont calculées pour le revenu et les actifs.

On détermine les classes d'imputation à partir du code SCIAN. On débute au niveau le plus détaillé, soit le niveau à 6 chiffres. On applique des règles de post-imputation très strictes au départ qu'on relâche au besoin jusqu'à ce qu'on trouve un donneur pour chaque receveur. Par exemple, pour l'imputation complète de l'information financière, le donneur doit être au même niveau à 6 chiffres du code SCIAN et dans la même région géographique; il doit avoir la même longueur et même fin de la période fiscale; le revenu et les actifs doivent être à l'intérieur de l'intervalle de 33%; et le signe du profit/perte net doit être le même. Si on ne trouve pas de donneur pour chacun des receveurs, on relâche les règles une à une jusqu'au dernier niveau, soit le donneur doit être au même niveau à 4 chiffres du code SCIAN et posséder le même signe pour le profit net.

Pour l'information financière, on impute les données du donneur au receveur. On ajuste proportionnellement les données au coût des marchandises vendues pour l'imputation partielle, ainsi qu'au revenu et actifs pour l'imputation complète. On impute l'information fiscale sans ajustement, au besoin, et finalement, on s'assure que les résultats sont en équilibre.

## **6. PLAN COMPTABLE**

Suite à l'imputation, le processus de désagrégation des valeurs génériques aux détails est appliqué aux données T2 pour s'assurer d'obtenir le niveau de détails désiré. Toutefois, un autre problème se pose dû au fait que la définition des variables entre les enquêtes économiques et les données T2 est parfois différente. Pour pallier à ce problème, un plan comptable a été développé par les comptables de SC pour standardiser les variables fiscales nécessaires au remplacement des données d'enquêtes économiques. Par exemple, le champ comptable revenu de commission est défini à partir des champs revenus de commissions et revenus de commissions de transactions immobilières des données T2. Puisque certains renseignements ne sont pas toujours disponibles au niveau des données T2 pour établir le plan comptable qui réponde à tous les besoins de chaque enquête économique, certaines variables des états des résultats n'ont pas encore été considérées.

## **7. CONCLUSION**

Les données T2 reçues de l'ARC sont utilisées pour le remplacement des données d'enquêtes économiques pour, entre autres, réduire les coûts et le fardeau de réponse ainsi que de possiblement améliorer la qualité des données. Elles sont reçues mensuellement et sont ensuite soumises à divers processus dans le but de nettoyer les données et de représenter l'univers des enquêtes économiques. Les données sont donc soumises, entre autres, aux règles de contrôle, à l'imputation, à la désagrégation des données génériques aux détails, ainsi qu'au plan comptable. Malgré certaines différences notées entre les données T2 et les enquêtes économiques, l'utilisation des données T2 pour le remplacement des données d'enquêtes économiques s'avère être une avenue très prometteuse. Certaines études sont présentement en cours pour évaluer la qualité de différents processus, notamment la désagrégation des champs génériques aux détails. La méthodologie reliée à ce processus pourrait être modifiée au besoin. Finalement, le plan comptable s'étendra également à d'autres variables de l'état des résultats ainsi que possiblement, au bilan.

## **RÉFÉRENCES**

Hidiroglou, M. A. et Berthelot, J.-M. (1986), "Contrôle statistique et imputation dans les enquêtes-entreprises périodiques", *Techniques d'enquête*, juin 1986, Vol 12, No. 1, p. 79-89.