

No 11-522-XIF au catalogue

**La série des symposiums internationaux
de Statistique Canada - Recueil**

**Symposium 2005 : Défis
méthodologiques reliés aux
besoins futurs d'information**



2005



Statistique
Canada

Statistics
Canada

Canada

Combiner des données administratives et des données d'enquête dans l'Enquête mensuelle sur les industries manufacturières

Steve Thomas, Krista Cook¹

RÉSUMÉ

L'utilisation de sources de données administratives par les organismes statistiques est devenue très populaire car elle permet de réduire considérablement le fardeau de réponse et les coûts. Un progrès récent à Statistique Canada a été l'accès à des données mensuelles calendrialisées sur le revenu disponible grâce à une entente conclue avec l'Agence du revenu du Canada (ARC). On a montré qu'il existe une forte association entre cette information et les renseignements sur le revenu recueillis par Statistique Canada au moyen de l'Enquête mensuelle sur les industries manufacturières (EMIM). Le présent article donne un bref aperçu de cette enquête et de l'information fournie par l'ARC et s'intéresse surtout à décrire comment les données disponibles ont été intégrées dans le processus d'enquête.

MOTS CLÉS : données administratives; calendrialisation; fardeau de réponse.

1. INTRODUCTION

L'utilisation de sources de données administratives par les organismes statistiques est devenue très populaire à mesure que le fardeau de réponse et le coût lié à la collecte des données sont devenus des questions problématiques. Le recours à des sources de données administratives permet de réduire considérablement les coûts et d'alléger le fardeau de réponse, tout en améliorant simultanément la qualité des données si la source utilisée est fiable. Les coûts de collecte sont habituellement les coûts les plus importants associés aux enquêtes. Le fardeau de réponse a toujours été un problème, mais grâce à l'utilisation de données administratives, le répondant n'aura plus que celui de continuer à fournir des données à la source administrative. Outre les économies et l'allègement du fardeau de réponse, les données administratives peuvent être de meilleure qualité que les données d'enquête et, si elles sont utilisées correctement, peuvent effectivement améliorer les estimations produites d'après les données d'enquête. Pour un exposé plus détaillé des questions liées à l'utilisation des données administratives, consulter l'article de Brackstone (1987).

L'utilisation de données administratives peut améliorer la qualité des données si la source choisie est fiable. Cependant, les avantages de l'utilisation de données administratives peuvent avoir pour prix le sacrifice de la qualité des données si celles utilisées ne correspondent pas à celles recueillies par enquête. L'unité que représentent les données administratives ne correspond pas nécessairement à l'unité statistique d'intérêt, la période qu'elles couvrent ne correspond pas toujours à la période de référence de l'enquête, la variable d'intérêt ne correspond pas nécessairement à la variable provenant de la source administrative ou encore le calendrier de production des données peut ne pas correspondre aux besoins de l'enquête. Un autre problème que posent les données administratives tient au fait qu'il est fort difficile de les valider. Elles proviennent souvent « telles quelles » de la source administrative et il incombe généralement à l'organisme statistique de procéder à leur vérification et à leur imputation.

Après un examen minutieux des données administratives et de leur source, l'organisme statistique doit mettre au point des méthodes exploitant pleinement le potentiel de ces données, sans perdre de vue leurs inconvénients éventuels. Ce défi est celui auquel a dû faire face le programme de l'Enquête mensuelle sur les industries manufacturières (EMIM) de Statistique Canada lors de l'utilisation des données de la taxe sur les produits et services (TPS) provenant de l'Agence du revenu du Canada (ARC). Au printemps de 2003, l'EMIM a été choisie comme candidate possible pour la substitution de données fiscales aux données d'enquête. Après des travaux approfondis de mise à l'essai, de développement et de mise en œuvre, le programme de l'EMIM est devenu le premier à publier des

¹ Steve Thomas, Statistique Canada, immeuble R.-H.-Coats, 16^e étage, Pré Tunney, Ottawa (Ontario) K1A 0T6; Krista Cook, Statistique Canada, immeuble R.-H.-Coats, 11^e étage, Pré Tunney, Ottawa (Ontario) K1A 0T6.

estimations mensuelles calculées en utilisant des données de la TPS comme source de données administratives à l'automne de 2004. Le présent article décrit l'enquête, la taxe sur les produits et services et certains défis qu'il a fallu relever lors de l'application des procédures en vue d'utiliser les données sur la TPS.

2. LA TAXE SUR LES PRODUITS ET SERVICES

La taxe sur les produits et services (TPS), qui a été introduite en 1991, est perçue sur la plupart des biens et services vendus au Canada. En général, le montant de la TPS correspond à 7 % des ventes de biens et de services, mais il y a des exceptions. Pour trois provinces atlantiques (Nouveau-Brunswick, Nouvelle-Écosse et Terre-Neuve-et-Labrador), la taxe est harmonisée avec la taxe de vente provinciale pour donner un taux de taxation global de 15 %. En outre, certains produits et services, comme une partie des produits de l'industrie des aliments, ne sont pas taxés. La taxe est perçue par l'Agence du revenu du Canada (ARC), sauf au Québec, où elle est perçue par le gouvernement provincial, puis versée à l'ARC.

Toute entreprise dont les recettes annuelles sont supérieures à 30 000 \$ doit produire une déclaration de TPS pour le gouvernement. La fréquence de la déclaration dépend du revenu annuel de l'entreprise. Celles dont le revenu annuel est compris entre 30 000 \$ et 500 000 \$ produisent une déclaration annuelle; celles dont le revenu est compris entre 500 000 \$ et 6 millions de dollars produisent une déclaration trimestrielle et celles dont le revenu est supérieur à 6 millions de dollars doivent produire une déclaration mensuelle. Les déclarants annuels ont jusqu'à trois mois après la date de fin de la période de déclaration pour produire leur déclaration, tandis que les déclarants trimestriels et mensuels doivent produire celle-ci dans les 30 jours après la date de fin de la période.

Toutes les transactions sont traitées au moyen d'un numéro d'entreprise (NE) à 15 chiffres attribué à l'entreprise par l'ARC. Les neuf premiers chiffres de ce numéro identifient une entreprise, une partie d'une entreprise ou un groupe d'entreprises selon la structure de l'entreprise et ses préférences en matière de déclaration, tandis que les six autres représentent les comptes de TPS multiples que peut posséder l'entreprise. Lors de chaque versement, ou transaction, l'entreprise fournit le chiffre de ventes total et les autres revenus, les crédits de taxes sur les intrants et la TPS perçue sous le NE à 15 chiffres pour la période de taxation. Durant l'année civile 2004, l'ARC a reçu environ 8,4 millions de transactions provenant de 2,5 millions d'entreprises. En termes de nombre d'entreprises, la plupart de ces transactions étaient trimestrielles, mais en termes de vente, la plupart étaient mensuelles (voir le tableau 1). Pour plus de renseignements sur la TPS et l'ARC, le lecteur est invité à visiter le site Web de l'ARC à www.cra_arc.gc.ca.

Tableau 1 : Transactions concernant la TPS – 2004

Fréquence	Nombre d'entreprises	Nombre de transactions	Revenu
Mensuellement	8,1 %	24,4 %	80,5 %
Trimestriellement	59,4 %	66,0 %	16,7 %
Annuellement	32,5 %	9,6 %	2,8 %

3. TRAITEMENT DES DONNÉES SUR LA TPS À STATISTIQUE CANADA

Les données sur la TPS sont transmises à l'ARC qui compile un fichier de données administratives comprenant la période couverte (date de début et de fin), le chiffre de ventes et autre revenu, ainsi que d'autres renseignements. L'information contenue dans ce fichier ne correspond pas nécessairement à celle dont a besoin Statistique Canada. Premièrement, il se peut que les données sur la TPS ne soient pas encore disponibles pour un grand nombre d'unités (déclarants tardifs, disparition d'entreprises, etc.). Deuxièmement, les données sur la TPS pourraient ne pas être déclarées uniformément : elles peuvent l'être mensuellement, trimestriellement ou annuellement. Troisièmement, la période de déclaration peut différer selon que la déclaration est produite mensuellement, trimestriellement ou annuellement et, enfin, elle peut changer au cours du temps pour chaque entreprise qui fait des versements de TPS à l'ARC. Quelle que soit la période de déclaration, la fréquence de déclaration ou l'existence de données manquantes, les données disponibles sont fournies (telles quelles) par l'ARC à Statistique Canada dans les sept semaines après la

fin de chaque mois de référence. Autrement dit, toutes les transactions doivent encore être vérifiées, imputées et calendrialisées.

La Division des données fiscales (DDF) de Statistique Canada est responsable, chaque mois, du traitement du fichier de données sur la TPS. Ce traitement comprend la vérification sous forme de contrôles d'intervalle et de détection des valeurs extrêmes, l'imputation pour les déclarations tardives, les valeurs incohérentes et les valeurs extrêmes critiques, ainsi que la calendrialisation des données sur le revenu. L'opération de calendrialisation consiste à transformer toutes les données sur la TPS, y compris les données non mensuelles (trimestrielles, annuelles, période de 13 mois) de sorte qu'elles correspondent aux données du mois civil requises pour les enquêtes-entreprises. Pour plus de renseignements sur la calendrialisation, consulter les travaux de Quenneville, Cholette et Hidiroglou réalisés en 2003. Une fois le traitement achevé, la DDF fournit un fichier au programme d'enquête dans les quatre jours après la réception du fichier de l'ARC.

4. L'ENQUÊTE MENSUELLE SUR LES INDUSTRIES MANUFACTURIÈRES

L'Enquête mensuelle sur les industries manufacturières (EMIM) est une enquête par sondage réalisée par Statistique Canada depuis plus de 50 ans. Elle est conçue pour recueillir des renseignements sur les livraisons, les stocks et les commandes auprès des établissements manufacturiers au Canada. Cette information est utilisée dans le calcul du produit intérieur brut (PIB), qui est un indicateur économique important. Le Système de comptabilité nationale est le principal utilisateur des données recueillies. Industrie Canada, le ministère des Finances et les entreprises manufacturières elles-mêmes utilisent aussi les estimations.

4.1 Contexte de l'EMIM

4.1.1 Aperçu

L'EMIM est conçue pour recueillir des données auprès des entreprises au niveau de l'établissement statistique, où les caractéristiques des livraisons (produits de propre fabrication – PPF), des stocks (matières premières – MP, produits en cours – PEC, produits finis – PF) et des commandes (commandes en carnet – CC) sont recueillies sur une base mensuelle. Dans le contexte de l'EMIM, un établissement manufacturier correspond à tout établissement correspondant à un code du secteur de la fabrication du Système de classification des industries de l'Amérique du Nord (SCIAN).

4.1.2 Base de sondage

Le Registre des entreprises (RE) de Statistique Canada est utilisé comme base de sondage pour l'EMIM. Le RE contient des renseignements sur environ 100 000 établissements manufacturiers et est mis à jour mensuellement afin de tenir compte des naissances et des décès d'entreprises et de tout autre changement subi par les établissements. Afin d'alléger le fardeau de réponse et de réduire les coûts de collecte, il est nécessaire de créer des strates à tirage nul, où les unités les plus petites qui constituent les 2 % inférieur de la taille totale (fondée sur les recettes provenant des livraisons) pour chaque province sont exclues de la population échantillonnée. La base de sondage est ainsi réduite à 35 000 établissements seulement.

4.1.3 Échantillonnage

Un échantillon d'environ 12 000 unités est tiré à partir de la population d'échantillonnage. Celle-ci est stratifiée selon l'industrie, la province et la taille. Les strates de taille sont construites suivant la procédure de Lavallée-Hidiroglou (Lavallée et Hidiroglou, 1988) où les limites de taille sont créées de façon à atteindre un coefficient de variation (c.v.) cible de 3,5 % pour la variable de taille au niveau de l'industrie (code du SCIAN à 4, 5 ou 6 chiffres) par province. L'échantillon est réparti selon une répartition-puissance avec $p=0,5$. Il reste le même de mois en mois, sauf en ce qui concerne les nouveaux établissements, qui sont échantillonnés avec la même probabilité que les unités figurant dans la population originale. Environ 7 000 unités à tirage complet sont incluses dans l'échantillon.

4.1.4 Imputation

Les valeurs manquantes, qui peuvent être le résultat de la non-réponse totale ou partielle, sont remplacées par une méthode d'imputation historique où un facteur de tendance est appliqué aux valeurs du mois précédent. Ce facteur de tendance est déterminé d'après un groupe d'établissements répondants ayant le même code du SCIAN dans la même province. Si les valeurs du mois précédent ne sont pas disponibles pour l'imputation (p. ex., nouveaux établissements), on utilise l'imputation par le ratio, qui est fondée sur le revenu brut d'entreprise (RBE) inscrit dans le RE.

4.1.5 Estimation

Les estimations sont calculées au moyen de l'estimateur d'Horvitz-Thompson où chaque réponse observée est multipliée par le poids d'échantillonnage et les valeurs sont totalisées sur les domaines. Dans le cas de l'EMIM, les domaines sont la province et l'industrie, qui est basée sur l'agrégation des codes du SCIAN à 6 chiffres au niveau à 3, 4, 5 ou 6 chiffres. Avant la publication, toutes les estimations sont étalonnées d'après les valeurs de l'Enquête annuelle des manufactures (EAM) les plus récentes. Le facteur d'étalonnage est, fondamentalement, la valeur de l'EAM divisée par la somme des mois de l'EMIM pour la même année de référence. Après application du facteur d'étalonnage, le niveau de l'estimation est dicté par l'EAM, tandis que la tendance de l'estimation est tirée de l'EMIM.

5. EMIM ET LA TPS

5.1 Aperçu

L'objectif, pour l'EMIM, est d'utiliser les données sur la TPS afin d'alléger le fardeau de réponse et de diminuer le coût de la collecte des données en réduisant au minimum l'effet sur la qualité des données. Plusieurs obstacles posés par les données sur la TPS ont dû être surmontés avant de pouvoir utiliser efficacement ces dernières dans l'EMIM. En premier lieu, les données sur la TPS ne correspondent pas à l'unité d'intérêt de l'enquête. La TPS est recueillie au moyen d'un numéro d'entreprise (NE), qui équivaut habituellement à l'entreprise statistique, alors que l'EMIM est conçue pour recueillir des données au niveau de l'établissement statistique. En deuxième lieu, les données sur la TPS ne sont pas toujours disponibles au moment où le programme de l'EMIM en a besoin. Enfin, les données sur la TPS ne contiennent des renseignements que sur les ventes et les autres revenus, et aucune information sur les stocks et les commandes.

5.2 Numéro d'établissement c. numéro d'entreprise

Statistique Canada utilise dans son Registre des entreprises (RE) une structure statistique représentative des entreprises au Canada. La structure hiérarchique comprend quatre niveaux avec les entreprises statistiques au plus haut niveau, suivies par les sociétés, les établissements et enfin les emplacements. Comme nous l'avons mentionné précédemment, l'établissement statistique est l'unité utilisée pour l'EMIM, puisque c'est à ce niveau que les entreprises peuvent fournir les informations financières requises. La structure statistique peut être simple, si l'entreprise possède un établissement et un emplacement, ou complexe, quand l'entreprise possède plusieurs sociétés, établissements ou emplacements. Les données sur la TPS sont obtenues au niveau du NE à 15 chiffres, où les neuf premiers chiffres correspondent approximativement à l'entreprise statistique. Les six autres représentent les comptes multiples que peut posséder l'entreprise. Cela signifie que cette dernière peut produire une déclaration plusieurs fois pour la même entité et que, étant donné la structure du RE, plusieurs établissements peuvent produire une déclaration sous le même NE à neuf chiffres.

5.2.1 Définition de l'établissement simple

Étant donné que le NE à neuf chiffres est relié à l'entreprise et que la collecte des données de l'EMIM est faite au niveau de l'établissement, il semble logique que l'on ne remplace les données d'enquête que pour les établissements faisant partie d'une structure statistique simple. Essayer de remplacer les établissements dont la structure est

complexe signifierait que les données sur l'entreprise doivent être réparties entre ces établissements. Dans le cas de l'EMIM, un établissement simple est défini comme toute unité incluse dans la partie non intégrée (PNI) du Registre des entreprises, ainsi que toute unité incluse dans la partie intégrée (PI) qui ne possède aucun des indicateurs de multiplicité associés à l'entreprise. Les indicateurs de multiplicité sont les suivants :

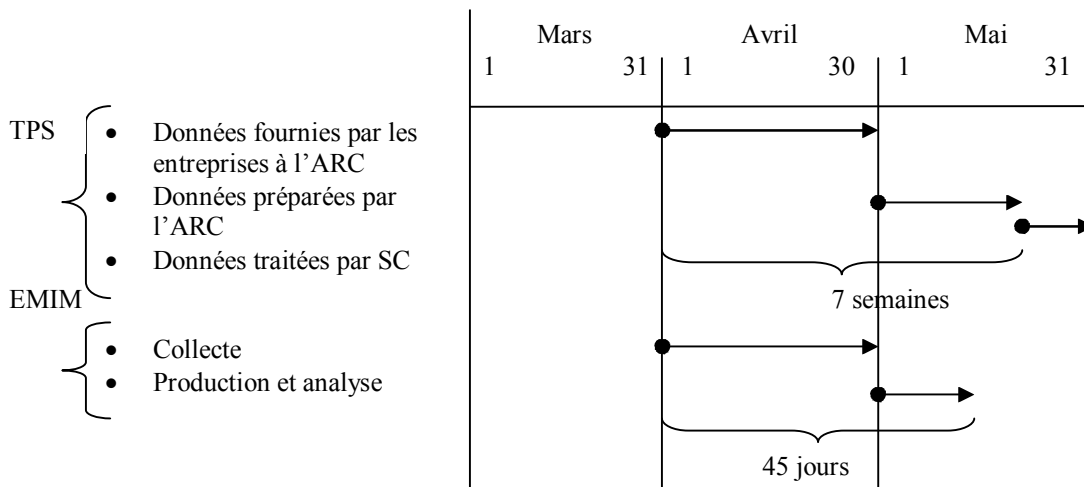
1. Emplacements multiples
2. Établissements multiples
3. Provinces multiples
4. Activités multiples
5. Coentreprises
6. Sociétés en nom collectif
7. Entreprises individuelles combinées

Cette définition comporte des exceptions. Si une structure est constituée d'un seul établissement avec emplacements multiples et que tous les emplacements sont des fabricants établis dans la même province, alors l'utilisation des valeurs de la TPS pour l'établissement ne devrait pas poser de problème, puisque celui-ci représente tous les emplacements. Si une unité est définie comme étant simple par l'EMIM, mais qu'elle déclare sa TPS en utilisant plusieurs comptes (c'est-à-dire multiples comptes à 16 chiffres pour le même NE à neuf chiffres), alors l'unité ne pourra pas être utilisée pour le remplacement. L'étude de l'agrégation des comptes multiples semble indiquer que la corrélation entre les NE à neuf chiffres agrégés et les livraisons n'est pas aussi bonne que celle observée pour les NE uniques. Par conséquent, aucun effort n'a été déployé en vue d'utiliser l'information agrégée dans le processus d'estimation.

5.3 Actualité

La figure 1 illustre l'actualité des données de l'EMIM et de celles sur la TPS. Le programme de l'EMIM produit son premier ensemble d'estimations un mois après la fin du mois de référence et l'ensemble final est habituellement publié 45 jours après le mois de référence (mois m). Après comparaison de ce calendrier à celui des données sur la TPS, le programme de l'EMIM a décidé d'utiliser les données sur la TPS recueillies deux mois avant le mois de référence ($m-2$). Il a également estimé que les données provenant du mois $m-2$ seraient de meilleure qualité que celles provenant du mois $m-1$, puisque les imputations devraient être moins nombreuses. En outre, les modèles d'imputation pour la TPS devraient être plus robustes, étant donné que le nombre de répondants sur lequel est fondé ce modèle est plus élevé. La DDF reçoit maintenant le fichier de données sur la TPS produit par l'ARC sept semaines après le mois de référence et a besoin de deux jours supplémentaires pour traiter les données. Cela signifie que les programmes d'enquête qui souhaitent utiliser la TPS comme source de données peuvent maintenant utiliser celles du mois précédent ($m-1$) au lieu du mois $m-2$, ce qui pourrait être l'un des futurs objectifs pour l'EMIM.

Figure 1 : Calendrier actuel de la disponibilité des données sur la TPS et des données de l'EMIM – Mois de référence - mars



5.4 Disponibilité des données

Les données sur la TPS fournies par la DDF aux divisions spécialisées chargées des enquêtes contiennent le chiffre de ventes et autres revenus, la TPS perçue et les crédits de taxes sur les intrants, qui sont toutes des variables reliées aux ventes. Aucun renseignement n'est disponible sur les stocks ni les commandes, et les corrélations entre les inventaires/commandes et les autres variables se sont avérées faibles. Une deuxième étude a eu pour but d'examiner les corrélations entre les stocks et le crédit de taxes sur les intrants, l'idée étant que le crédit est appliqué aux produits utilisés pour la fabrication d'autres produits. Malheureusement, les corrélations étaient fort semblables à celles calculées d'après les revenus associés à la TPS. Par conséquent, il a été décidé qu'il ne convenait pas de modéliser les caractéristiques des stocks et des commandes d'après des variables auxiliaires de la TPS.

En dernier recours, nous avons étudié la possibilité de modéliser les variables de stocks/commandes en utilisant le processus d'imputation actuel. Dans cette situation, les valeurs mensuelles courantes sont modélisées d'après les valeurs du mois précédent et une tendance. La tendance est calculée en se fondant sur les valeurs des réponses provenant du groupe d'imputation et appliquée à la valeur déclarée ou imputée le mois précédent au niveau des microdonnées. Cette approche était justifiée par le fait que les tendances observées pour les stocks et les commandes sont fondamentalement uniformes et que des mesures supplémentaires étaient prises pour s'assurer que les unités de remplacement ne représentent pas une proportion importante de l'estimation globale.

6. PROCESSUS D'ENQUÊTE

6.1 Identification initiale de l'échantillon de déclarants de la TPS

La première étape du processus d'enquête consistait à identifier les unités disponibles pour le remplacement des données en commençant par identifier les établissements simples. La même définition que celle décrite à la section 5.2.1 a été utilisée. Notons qu'elle demande un lien entre les données d'enquête et les données sur la TPS. Nous avons mentionné à la section 5.3 un problème de synchronisation obligeant à utiliser les données sur la TPS du mois $m-2$ pour l'estimation. Toutefois, à cause de ce problème, pour l'identification des unités à remplacer durant l'échantillonnage, nous avons dû établir un lien avec des données encore plus anciennes ($m-3$) en espérant que les unités identifiées demeureraient simples au moment de l'estimation.

6.2 Seuils de taille

Nous avons pensé au départ que les unités simples étaient toujours les établissements les plus petits qui contribuaient peu au total global des estimations. Cependant, nous avons constaté que certains grands établissements

ont une structure statistique simple. Afin de nous assurer que les établissements les plus grands ne soient pas remplacés, nous avons adopté une règle 75/50/50. Si l'établissement comptait parmi les plus grands représentant 75 % des livraisons, 50 % du total des stocks ou 50 % des commandes en carnets, il ne s'agissait pas d'un candidat au remplacement des données. Nous nous sommes fondés sur des déclarations pour une année afin d'éviter tout effet des fluctuations mensuelles ou saisonnières. Les données recueillies pour les mois $m-2$ à $m-14$ ont été utilisées, puisque celles pour le mois actuel et le mois antérieur n'étaient pas disponibles. Étant donné que les unités comprises dans l'échantillon depuis moins d'un an étaient aussi des candidates au remplacement, une correction a été faite pour ces dernières afin d'obtenir un total annuel « projeté » aux fins de comparaison. Le total annuel projeté est calculé comme suit :

$$\text{Total annuel projeté} = (\text{total annuel}) * 12 / (\text{nombre de mois dans l'échantillon})$$

Aux fins de nos travaux, nous avons considéré qu'une unité était un contributeur important si elle se situait au dessous des seuils 75/50/50 au niveau du groupe SCIAN (c.-à-d. SCIAN à quatre chiffres pour les transports et SCIAN à trois chiffres pour le reste du secteur de la fabrication) par province pour chacune des trois caractéristiques principales. Les seuils de taille d'exclusion 75/50/50 ont été utilisés pour créer un fichier de seuils pouvant servir pour établir les limites pour l'introduction de nouvelles unités dans la partie de remplacement des données (S2) du registre dans l'avenir.

Après avoir déterminé quelles unités étaient simples et avoir appliqué les seuils, certaines unités supplémentaires ont été exclues afin de mettre le point final à la liste d'unités disponibles pour le remplacement des données. Les agents du domaine de spécialisation ont fourni une liste des unités importantes pour lesquelles les données ne doivent pas être remplacées; les unités faisant partie d'une déclaration combinée ont été éliminées de la liste d'admissibilité, puisqu'elles continuent de faire partie de l'entité de collecte; en outre, dans le cadre de la sélection initiale, nous avons également éliminé les unités considérées comme mortes/hors du champ d'observation dans la table de gestion des décès de l'EMIM, mais qui étaient encore actives selon le RE. Il s'agissait d'unités mortes d'après l'EMIM, mais pour lesquelles l'information rétroactive provenant de l'enquête n'avait pas encore été traitée dans le RE. Les unités éliminées n'ont pas été considérées comme admissibles pour le remplacement des données ou pour l'introduction dans le modèle.

6.3 Échantillonnage pour le remplacement des données

L'étape suivante consistait à sélectionner un échantillon à 50 % parmi les unités identifiées comme étant admissibles. Il s'agissait d'un échantillon aléatoire simple des unités admissibles au niveau des codes à trois et à quatre chiffres du SCIAN par province. Les unités sélectionnées, ainsi que tous les cas de refus simples ont constitué l'échantillon pour le remplacement, qui a été nommé S2. Pour les unités simples ayant refusé de participer, on a estimé que leur remplacement par des renseignements provenant du fichier de la TPS, plutôt que par les techniques d'imputation existantes, produiraient des données plus fiables. Les unités simples qui n'ont pas été sélectionnées demeurent dans l'échantillon et continuent de faire l'objet d'une collecte mensuelle de données. Ces unités, nommées S1, sont utilisées dans la modélisation des unités S2. La subdivision de l'échantillon en sous-échantillons S1 et S2 demeure constante de mois en mois, sauf dans les situations où une unité n'est plus simple ou perd son lien avec les données sur la TPS.

6.4 Estimation mensuelle

Une fois que les données d'enquête nécessaires ont été recueillies et que les fichiers TPS appropriés sont disponibles, les estimations peuvent être calculées. Il convient de souligner que, dans la description qui suit, une partie de l'échantillon de remplacement S2 ne possède pas de lien avec la TPS et est dénotée $S2_{\overline{GST}}$. Il s'agit de la proportion d'unités qui ont été identifiées comme des candidates pour le remplacement durant l'échantillonnage, mais pour lesquelles, au moment de l'estimation (jusqu'à deux mois plus tard), aucun lien ne peut être établi. Si nous représentons maintenant par y les recettes provenant des livraisons de l'EMIM pour le mois m , et par x , la valeur la plus à jour des ventes du fichier TPS (pour le mois $m-2$), alors les estimations des totaux peuvent être créées comme suit.

$$\tilde{Y} = \sum_s w_k y_k^*$$

$$\text{où } y_k^* = \begin{cases} y_k & \text{if } k \notin S2 \\ \left(\frac{\hat{Y}_{S1_{GST}}}{\hat{X}_{S1_{GST}}} \right) \times x_k & \text{if } k \in S2_{GST} \text{ et active} \\ \text{Imputation historique} & \text{if } k \in S2_{\overline{GST}} \text{ et active} \\ 0 & \text{if } k \in S2 \text{ et morte.} \end{cases}$$

Pour être certain de créer un modèle convenable, les valeurs extrêmes sont éliminées du groupe modèle (S1 avec TPS) en utilisant la même routine multivariée de détection des valeurs extrêmes que celle appliquée dans le processus de production ordinaire de l'EMIM (Franklin et Brodeur, 1997). Il s'agit d'observations pour lesquelles le revenu d'après les données sur la TPS ne concorde pas avec les valeurs des livraisons déclarées. Les observations pour lesquelles la valeur des livraisons est grande et celle des re

po

7.2 Stocks et commandes

Pour les caractéristiques des stocks et des commandes, l'effet du remplacement est plus important que celui observé pour les livraisons. Ce résultat n'est pas étonnant, puisqu'une plus grande proportion des estimations ont pu être calculées d'après des données sur la TPS. Au niveau national pour l'ensemble du secteur de la fabrication, 16,7 %, 16,1 %, 15,8 % et 18,3 % des estimations calculées pour les matières premières, les produits en cours, les produits finis et les commandes en carnet, respectivement, ont été représentées par des valeurs modélisées en utilisant les données sur la TPS. Aucune analyse des coefficients de variation n'est disponible à l'heure actuelle.

8. QUESTIONS À EXAMINER DANS L'AVENIR

Pour l'instant, l'inclusion des données sur la TPS dans le plan de sondage de l'EMIM donne d'assez bons résultats. Dans l'avenir, si l'on dispose du temps et des ressources nécessaires, plusieurs idées devraient être explorées. En premier lieu, le plan de sondage actuel ne permet aucun renouvellement des unités qui sont modélisées d'après les données sur la TPS. Une option consiste à faire entrer et sortir les unités de la partie S2 mensuellement. L'autre option consiste à procéder à un renouvellement complet chaque année ou durant la restructuration, qui a lieu tous les trois ans. Le renouvellement permettrait de tenir à jour les renseignements du profil de base, comme le code du SCIAN, la taille et un profil caractéristique.

La deuxième question qui devrait être envisagée dans l'avenir est l'accroissement du nombre d'unités remplacées par l'utilisation de données sur la TPS. Le processus de production courant a montré que le remplacement de 50 % des établissements admissibles donne de bons résultats, avec peu d'incidence sur les estimations. Pourrait-on porter cette proportion à 60 % ou à 70 % afin d'alléger davantage le fardeau de réponse et d'accroître les économies?

Les données de l'EMIM devraient-elles être étalonnées d'après les revenus totaux selon la TPS? Le fait est que l'effet de l'utilisation des données sur la TPS sur la qualité ne pourra pas être évalué tant que ne sera pas intégrée dans le calcul de la variance celle due à la modélisation pour l'imputation. L'étalonnage rehausserait sans doute la qualité des données à condition que nous utilisons la bonne population, la bonne variable, l'étalonnage par rapport aux bons totaux et les bons groupes modèles. L'étude de ces méthodes pourrait être envisagée durant tout remaniement futur de l'enquête.

RÉFÉRENCES

- Brackstone, G. J. (1987), "Utilisation des dossiers administratifs à des fins statistiques", *Techniques d'enquête*, 13, pp. 35-51.
- Dubreuil, G., Hidirolou, M.A., Pierre, L. (2003), "Use of Administrative Data in Modeling of the Monthly Survey Data", *Recueil de la section des méthodes d'enquête, Société Statistique du Canada*.
- Franklin, S. and Brodeur, M. (1997), "A Practical Application of a Robust Multivariate Outlier Detection Method", *1997 Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 186.
- Hurtubise D. (2004), "Variance due à l'imputation à l'aide d'un modèle par le ratio", *Le Bulletin d'imputation, Vol. 4 no 2, Direction de méthodologie, Statistique Canada*.
- Lavallée, P. et Hidirolou, M. (1988), "Sur la stratification de populations asymétriques", *Techniques d'enquête*, 14, pp. 35-45.
- Quenneville, B., Cholette P., Hidirolou M.A. (2003), "Estimating calendar month values from data with various reporting frequencies", *Contributed paper presented at the Annual Meeting of the Joint Statistical Meeting held in San-Francisco, California*.

Yung, W., Cook, K., Thomas, S. (2004), "Use of GST Data by the Monthly Survey of Manufacturing", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2004, à paraître.