UTILISATION DES DONNÉES DE LA TAXE SUR LES PRODUITS ET SERVICES DANS LE REMANIEMENT DE L'ENQUÊTE MENSUELLE DU COMMERCE DE GROS ET DE DETAIL DE STATISTIQUE CANADA

M. BRODEUR et H. BÉRARD

Statistique Canada - Division des méthodes d'enquêtes auprès des entreprises

RÉSUMÉ

L'Enquête mensuelle du commerce de gros et de détail (EMCGD) est une enquête très importante menée par Statistique Canada. Elle vise à mesurer les ventes mensuelles pour divers domaines d'intérêt tels que les provinces et les groupes de commerce (regroupements de codes de classification type des industries). Les estimations mensuelles sont utilisées dans le calcul du Produit intérieur brut (PIB). L'estimation de tendance représente un indicateur économique majeur et la somme des estimations mensuelles est utilisée pour le calcul des paiements de péréquation (des transferts de fonds du gouvernement fédéral vers les provinces les plus pauvres).

L'EMCGD recourt à un plan de sondage stratifié aléatoire simple sans remise. Le dernier remaniement complet de l'enquête remonte à 1988. Plusieurs améliorations ont été apportées à l'enquête au fil des ans. Toutefois, la venue d'une nouvelle classification industrielle, le besoin de réduire les coûts d'enquêtes et l'utilisation de systèmes informatiques vétustes justifient le besoin de faire un remaniement. Au même moment, Statistique Canada a initié plusieurs projets pour utiliser davantage les données administratives provenant de la collecte de la Taxe sur les produits et services (TPS). Deux projets préparatoires au remaniement ont été amorcés. Le premier projet vise à étudier la possibilité d'utiliser les données des fichiers de TPS comme variable de stratification. Le deuxième propose de développer un estimateur par calage en employant les données de TPS comme variable auxiliaire. Un tel estimateur pourrait éventuellement remplacer l'actuel estimateur par dilatation. Cette communication présente les principales étapes du remaniement et traite des résultats préliminaires des deux projets d'utilisation de la TPS.

1. Introduction

L'Enquête mensuelle du commerce de gros et de détail (EMCGD) est une enquête très importante de Statistique Canada (SC) et son plan d'échantillonnage, conçu en 1988, a servi de modèle à plusieurs enquêtes. L'enquête fait présentement l'objet d'un remaniement. Le plan d'échantillonnage demeurera sensiblement le même. Cependant, les unités d'échantillonnage seront modifiées et le processus de renouvellement simplifié. Toutefois, un des problèmes majeurs de l'enquête est relié aux problèmes de classification des unités et le fait qu'au fil du temps, les unités se retrouvent dans de mauvaises strates. Les problèmes proviennent surtout de la variable utilisée lors de la stratification. Cet article traitera plus en détails de deux projets préparatoires au remaniement. Le premier projet vise à étudier la possibilité d'utiliser les données des fichiers de la Taxe sur les produits et services (TPS) comme variable de stratification. Il s'agit d'une taxe sur la valeur ajoutée. Le deuxième projet propose de développer un estimateur par calage en employant les données de TPS comme variable auxiliaire. Un tel estimateur pourrait éventuellement remplacer l'actuel estimateur par dilatation.

Pour bien comprendre l'impact de ces deux projets, il est important de faire une mise en contexte de l'EMCGD. La section 2 fournira une description du plan d'échantillonnage et de l'estimateur. La section 3 dressera un bref historique de l'enquête depuis le dernier remaniement de 1988, tandis que la section 4, présentera brièvement les objectifs du remaniement. La section 5 traitera du projet d'utilisation de la TPS dont, la sous- section 5.3 qui parlera de l'utilisation de la TPS comme variable de stratification. Les sous-sections suivantes parleront d'estimation par calage et de résultats.

2. Description du plan d'échantillonnage

Le dernier remaniement majeur de l'enquête remonte à 1988. L'EMCGD est une enquête mensuelle conçue pour mesurer principalement les ventes par groupe de commerce (regroupement de codes de classification type industrielle (CTI 1980)) à trois ou quatre chiffres par province et pour certaines régions métropolitaines de recensement (RMR). L'échantillon est sélectionné à partir du Registre des entreprises (RE) de Statistique Canada. Le RE contient toutes les entreprises 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 simples toutefois, l'entreprise, la compagnie, l'établissement et l'emplacement coïncident. Les entreprises dites complexes peuvent œuvrer dans plusieurs provinces et dans différents secteurs industriels et ont très souvent un revenu très élevé. La population cible de l'enquête varie pour le commerce de gros et de détail. Pour le commerce de détail, elle est définie comme étant toute compagnie comportant au moins un emplacement œuvrant dans le commerce de détail alors que pour le commerce de gros, elle est définie comme étant toute compagnie comportant au moins un établissement œuvrant dans le commerce de gros. L'unité d'échantillonnage est la compagnie et seulement les compagnies ayant des employés font partie de l'enquête.

La taille d'échantillon pour le commerce de détail est d'environ 16 000 compagnies statistiques vivantes provenant d'une population de 137 000. La taille d'échantillon pour le commerce de gros est de 7 000 compagnies statistiques vivantes provenant d'une population de 58 000. La population est stratifiée par province, territoire, certaines RMR et par groupe de commerce. Chaque combinaison de groupe de commerce et de région géographique forme une strate. Chaque strate est divisée en trois sous-strates selon la taille : une à tirage complet et deux à tirage partiel contenant respectivement les moyennes et les petites compagnies. Les strates à tirage complet englobent toutes les compagnies qui ont une structure complexe, *i.e.*, opérant dans plus d'un groupe de commerce ou région géographique, ou qui ont un revenu brut d'entreprise (RBE) supérieur à un seuil donné. Les autres compagnies sont classées dans les strates à tirage partiel selon leur RBE et les seuils des strates. En 1988, les seuils entre la strate à tirage complet et les strates à tirage partiel ont été calculés à l'aide de la méthode d'Hidiroglou (1986). Les seuils entre les deux strates à tirage partiel correspondent au niveau du RBE délimitant les unités simples et complexes sur le RE. Le coefficient de variation des ventes visé se situe à 1,5 % au niveau canadien alors qu'il est de 2,5 % au niveau provincial et de 3,5 % au niveau des groupes de commerce. La méthode de répartition de l'échantillon utilisée est celle de la racine carrée du RBE.

L'enquête emploie un plan de sondage stratifié aléatoire simple sans remise. Les compagnies d'une même strate h à tirage partiel sont réparties aléatoirement en un certain nombre P_h de grappes, appelées panels, de taille égale à une unité près. Chaque mois, toutes les compagnies d'une sélection contiguë de p_h panels forment l'échantillon. Les valeurs P_h et p_h sont des fonctions de la fraction de sondage désirée, mais aussi du nombre de mois maximal qu'une compagnie peut demeurer dans l'échantillon et du nombre minimal qu'elle doit demeurer exclue de l'échantillon une fois sortie. En effet, une rotation est appliquée chaque mois à l'échantillon en substituant un panel de l'échantillon par un nouveau panel qui était hors échantillon. Plus de détails sont disponibles dans Hidiroglou et coll. (1991). De plus, des compagnies nouvelles au commerce de détail (naissances) sont identifiées chaque mois par le RE et ajoutées systématiquement aux panels des strates auxquelles elles appartiennent. Les naissances tombant dans les panels échantillonnés font par conséquent partie de l'échantillon. Il est à noter que chaque mois les naissances sont ajoutées systématiquement aux panels de la population de sorte que l'on obtienne, dans les faits, le nombre espéré de naissances dans l'échantillon.

Les compagnies ayant cessé leurs opérations dans le commerce de détail ou de gros (cessations) apparaissent mensuellement elles aussi. Des cessations peuvent se rapporter à une compagnie statistique dont les activités ont cessé, donc retirée des affaires, ou dont les principales activités ne sont plus le commerce de détail ou de gros, donc hors du champ d'observation ou inactives. Elles sont identifiées à l'aide de sources administratives permettant la mise à jour du RE et à l'aide de mises à jour relatives aux résultats des opérations de collecte des enquêtes, y compris l'EMCGD. Les cessations sont d'abord codées comme telles dans le RE, puis dans la base de sondage de l'EMCGD. Les résultats des enquêtes sont un moyen beaucoup plus rapide d'identification des cessations puisque les sources administratives peuvent comporter un délai d'un an. Ainsi, pour avril 1999, la proportion des cessations identifiées dans la population du commerce de détail est de 10,3 % et la proportion de celles dans l'échantillon est de 18,7 %. Pour le commerce de gros, nous avons 26,8 % de cessations identifiées dans la population et 34,0 % dans l'échantillon. Ces faits ne sont pas des évidences que notre échantillon est biaisé pour autant. Nous croyons qu'il existe des unités ayant cessé leurs activités et non identifiées dans le volet non échantillonné qui sont jugées actives dans le RE. Ces unités ont une mesure de taille positive. Puisque l'EMCGD met à jour indirectement sa propre base de sondage (source qui est donc non indépendante) et puisque nous n'avons aucune façon de distinguer facilement dans le RE les cessations identifiées à l'aide de sources administratives ou de sources d'enquête, les cessations ne sont pas automatiquement éliminées de l'échantillon et de la base de sondage de l'EMCGD. Lavallée (1996) a montré l'ampleur du biais que susciterait l'utilisation, pour la mise à jour de la base de sondage, d'une source d'information non indépendante du plan de sondage de l'enquête. C'est pourquoi, une ou deux fois par année, des cessations sont éliminées de la base de sondage et de l'échantillon d'une façon non biaisée (Trépanier et coll. 1998).

Malgré l'utilisation des panels, on peut considérer à toutes fins pratiques le plan de sondage comme un plan de sondage stratifié aléatoire simple d'éléments sans remise, dont la fraction de sondage est approximativement n_h / N_h, où N_h est le nombre de compagnies dans la strate au niveau de toute la population et n_h est le nombre de compagnies dans la strate au niveau de l'échantillon. En effet, l'estimateur utilisé est un estimateur post-stratifié classique où les poids de sondage sont P_h / p_h, les post-strates sont les strates telles que définies en 1988 et les totaux de contrôle sont les tailles de population N_h des strates h. L'estimateur des ventes, $\hat{Y}$, pour un domaine d peut donc s'écrire :

$$\hat{Y}(d)=\sum_h \frac{N_h}{\hat{N}_h}\frac{P_h}{p_h}\sum_{i=1}^{n_h} y_{hi}(d)=\sum_h \frac{N_h}{\left(\frac{P_h}{p_h}n_h\right)}\frac{P_h}{p_h}\sum_{i=1}^{n_h} y_{hi}(d)=\sum_h \frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi}(d)$$

où $y_{hi}(d) = y_{hi}$, c.-à-d., les ventes de la compagnie i, si celle-ci est dans le domaine d, et $y_{hi}(d) = 0$ sinon. On retrouve donc la forme de l'estimateur par dilatation où le poids de sondage est N_h / n_h . C'est d'ailleurs à cet estimateur que nous ferons référence dans la suite de cet article. De façon plus générale, cet estimateur peut s'écrire :

où a_i représente le poids de sondage de i. Lorsqu'une compagnie fait partie de

$$\hat{Y}(d) = \sum_{i \in s} a_i y_i(d)$$

l'échantillon, un questionnaire est généré pour l'ensemble des composantes de la compagnie qui œuvre dans le même groupe de commerce. Sur ce questionnaire, on recueille les ventes mensuelles au détail effectuées dans chaque région géographique d'intérêt et dans certains cas les inventaires. Une compagnie peut recevoir plus d'un questionnaire si, en autres, elle œuvre dans plusieurs groupes de commerce. Cette procédure nous permet de produire des estimations fiables par groupes de commerce et régions géographiques. On peut donc considérer à toutes fins pratiques que le plan de sondage est un plan stratifié aléatoire simple sans remise par grappes. La grappe est cette fois-ci représentée par la compagnie statistique et les éléments de la grappe sont les différents questionnaires générés pour cette compagnie. Cette analogie sera utile plus tard lorsqu'on parlera d'estimation par l'entremise du Système généralisé d'estimation de Statistique Canada.

3. Développement de l'EMCGD depuis 1968

Après le remaniement de 1988, un certain nombre de projets furent mis en place pour augmenter la taille de l'échantillon afin d'améliorer les estimations de certains groupes de commerce problématiques. Ces projets visaient à maintenir un niveau de qualité acceptable dans l'EMCGD. Toutefois, il demeure que sans mise à jour complète de la stratification pendant toutes ces années et sans méthode d'estimation utilisant de l'information auxiliaire pour compenser, la qualité de certaines estimations de l'enquête s'est graduellement détériorée. En effet, la valeur de la strate et du domaine différant de plus en plus (p. ex. : le groupe de commerce de stratification et de domaine) et des poids de sondage élevés étant associés à certaines compagnies ayant pris de l'importance qui avaient été classées dans les strates à tirage partiel à l'époque rendaient les estimations instables dans le temps et faisaient grimper certains coefficients de variation.

Pour remédier aux problèmes, deux projets ont débuté en 1997. Le premier projet consistait à effectuer une restratification complète des compagnies de la base de sondage, à revoir certaines fractions de sondage et à sélectionner un nouvel échantillon tout en maximisant le chevauchement avec l'ancien échantillon. Le nouvel échantillon ainsi restratifié a été mis en production pour la première fois en avril 1998 après un test complet de quatre mois. Plus de détails sur ce projet sont disponibles dans Trépanier et coll. (1998).

Le second projet visait à minimiser l'inévitable bris dans la série chronologique des estimations causé par la restratification. Le problème devenait donc de trouver un moyen de raccorder les deux séries d'estimations, celle d'avant la restratification et celle d'après, tout en améliorant la qualité des estimations passées par l'utilisation à l'estimation d'informations auxiliaires comme le RBE et le nombre de compagnies. Les estimateurs étudiés étaient du type par calage (Deville et Särndal 1992). Plus de détails sont disponibles dans Bissonnette et coll. (1998).

La restratification a été très bénéfique et a permis de corriger de nombreux problèmes reliés à une mauvaise classification. Cependant, l'estimateur par calage n'a pas donné les résultats escomptés surtout à cause de la faible corrélation de la variable auxiliaire avec les ventes. Les résultats étaient intéressants au niveau agrégé mais, étaient quelquefois aberrants au niveau des groupes de commerce. Nous avons donc conservé l'estimateur par dilatation pour toutes les strates.

4. Objectifs du remaniement

Tel que mentionné auparavant, l'EMCGD entre dans une période de remaniement. La présente méthodologie de l'enquête est basée en outre sur la classification type des industries de 1980 et il est impératif que l'EMCGD soit remaniée en vertu du nouveau Système de classification des industries de l'Amérique du Nord (SCIAN). L'environnement informatique de l'EMCGD repose principalement sur l'ordinateur principal et une migration vers un environnement de micro-informatique semble inévitable pour réduire les coûts d'exploitation, favoriser l'intégration des activités d'enquêtes et accroître la flexibilité. Le remaniement ouvre la porte vers l'utilisation des produits généralisés de SC.

L'EMCGD ne couvre présentement que les compagnies ayant des employés. Un ajustement au niveau de l'estimation est fait pour contrer ce problème mais, il a tendance à faire fluctuer les estimations. Le RE possède maintenant une liste exhaustive des entreprises au Canada. Ainsi, l'EMCGD doit modifier sa couverture pour ajouter les entreprises sans employés et les détaillants hors magasins qui sont présentement sondés séparément. D'autres raisons méthodologiques justifient le remaniement de l'EMCGD. L'avènement du Projet d'amélioration des statistiques économiques provinciales (PASEP) et, en conséquence, l'intégration des enquêtes annuelles sur le commerce de gros et détail à l'Enquête unifiée des entreprises (EUE) forcent l'EMCGD à harmoniser certaines de ses caractéristiques (p. ex. unité visée, stratification) avec celles de l'EUE pour assurer la cohérence entre les deux sources de données statistiques. Dans la même veine, le remaniement de l'EMCGD permet à celle-ci de mettre en place les seuils d'exclusion Royce-Maranda permettant de réduire le fardeau de réponse des petits établissements en les exemptant de compléter un questionnaire.

Plusieurs années de renouvellement mensuel de l'échantillon du commerce de détail ont démontré que la méthodologie actuelle de renouvellement dans l'EMCGD rend difficile la gestion des unités incorrectement stratifiées selon la taille. De nouvelles approches de renouvellement de l'échantillon doivent être examinées. La méthodologie de l'EMCGD remaniée doit aussi porter une attention particulière aux conséquences de certaines méthodes sur l'ampleur des révisions aux estimations.

La disponibilité des données de la Taxe sur les produits et services (TPS) offre une occasion unique d'utiliser une source de données administratives dans le remaniement de l'EMCGD pour accroître la qualité des estimations, réduire les coûts ou diminuer le fardeau de réponse. Les données administratives peuvent être utilisées pour des totalisations directes; des estimations indirectes; comme variable de stratification; pour valider et corriger les réponses lors de la vérification et imputation et pour le remplacement de données d'enquête.

Deux projets préparatoires au remaniement ont été amorcés en 1999. Le premier projet vise à étudier la possibilités d'utiliser les données des fichiers de TPS comme variable de stratification. Le deuxième propose de développer un estimateur par calage en employant les données de TPS comme variable auxiliaire. Un tel estimateur pourrait éventuellement remplacer l'actuel estimateur par dilatation. Ces deux projets visaient essentiellement à se familiariser avec les concepts de la TPS. Il a été jugé trop prématuré de considérer la TPS comme remplacement des données d'enquête. Les sections suivantes traitent de l'utilisation des données de la TPS dans le cadre du remaniement de l'EMCGD.

5. Projet d'utilisation de la TPS

5.1 Description de la TPS

En 1991, le Gouvernement fédéral a décidé d'introduire une nouvelle taxe sur la valeur ajoutée que l'on nomme Taxes sur les produits et services (TPS). Chaque entreprise doit remettre la valeur de la TPS à l'Agence de douanes et du revenu du Canada (ADRC). Un fichier administratif contenant un ensemble de variables reliés aux ventes de TPS est maintenant accessible à Statistique Canada. Pour notre étude, nous nous intéressons à la variable Fourniture taxable. Pour cette étude, le montant de Fourniture taxable sera considéré comme la variable de ventes de TPS.

Chaque entreprise ayant 30 000$ ou plus en ventes annuelles doit s'enregistrer auprès de l'ADRC. La fréquence de remise dépend de la taille de l'entreprise. Les entreprises ayant plus de 6 millions$ en ventes annuelles doivent remettre mensuellement. Celles ayant entre 500 000$ et 6 millions$ doivent remettre à chaque trimestre alors que les autres, celles ayant moins de 500 000$ remettent annuellement. Environ 78% des détaillants et des grossistes remettent sur une base trimestrielle.

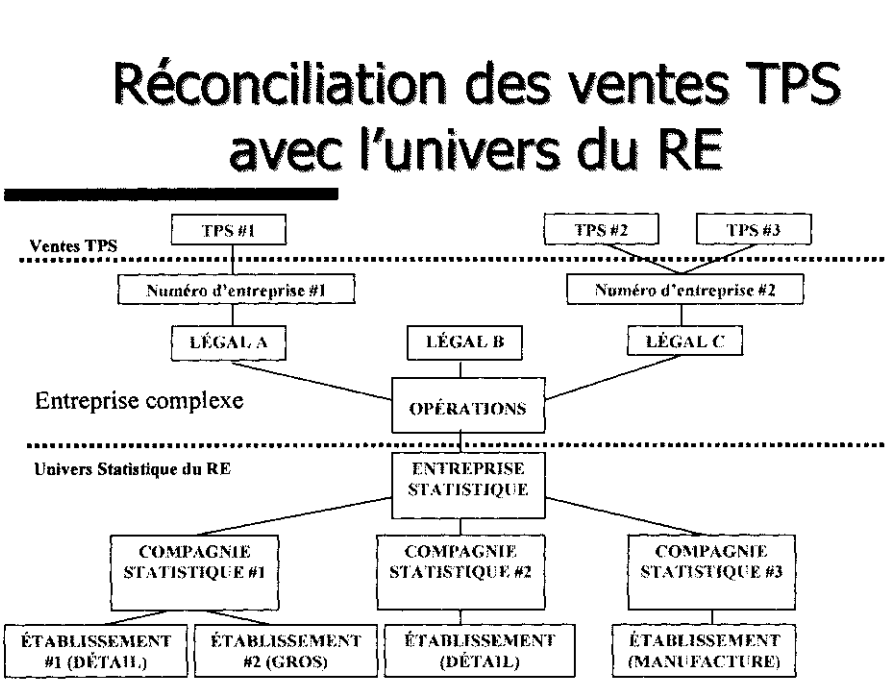
Présentement, Statistique Canada reçoit un fichier de remise de l'ADRC. Ce fichier regroupe les codes d'activité, les ventes totales de TPS et la valeur de la TPS perçue. Le fichier est reçu sur une base irrégulière. Il s'écoule 90 jours du moment où une entreprise remet sa taxe au moment où les données sont acheminées à Statistique Canada qui fait alors une vérification et imputation sommaire des données manquantes et en erreur.

La première étape de ce projet a été de créer un fichier de ventes TPS pour chaque compagnie statistique de notre population. Cette opération s'est avérée très complexe et a demandé beaucoup de travail. Au cours de cette étape, qui fut nommée « réconciliation avec le registre des entreprises », nous avons fait plusieurs découvertes importantes quant à la qualité et au contenu du fichier de données de

ventes de TPS. Citons ici quelques exemples : le fichier initial de données de ventes de TPS ne contient pas de date de début de période; les périodes de référence varient d'une compagnie à l'autre; la fréquence de remise peut varier d'une compagnie à l'autre et même d'une fois à l'autre pour une même compagnie; enfin le fichier de ventes de TPS ne tient pas compte des transactions intra-entreprises. De plus, les délais d'acquisition du fichier sont longs et le traitement est limité. Les unités de l'univers du registre des entreprises ne trouvent pas toutes un enregistrement correspondant sur le fichier de ventes de TPS et finalement, il y a des différences entre les concepts de ventes de TPS et les ventes mensuelles de l'EMCGD.

Pour cette étude, nous avons fait la somme de 12 mois de données (ventes annualisées) de avril 1998 à mars 1999. Dans environ 8% des cas, la valeur annualisée des ventes de TPS peut en réalité représenter une période de moins de 12 mois ou de plus de 12 mois. En effet, un répondant peut subitement remettre de façon annuelle alors qu'il remettait de façon trimestrielle. Comme la période visée par la remise n'est pas indiquée sur le fichier, la somme directe des 12 dernières remises peut mener à des sous-estimations ou des surestimations des ventes annuelles de TPS.

Nous avions la somme des ventes de TPS au niveau du sommet de la structure opérationnelle, donc de l'entreprise statistique qu'il fallait répartir aux niveaux statistiques inférieurs. Cette étape a demandé beaucoup de temps et d'efforts pour l'établissement des liens entre le numéro de compte de TPS et le numéro d'entreprise (NE : indicateur unique du RE). Cette tâche n'était pas triviale dans le cas d'entreprises complexes. Cette étape était nécessaire pour faire la répartition des ventes de TPS au niveau de la structure statistique. Le RBE a été utilisé pour faire la répartition des ventes de TPS au niveau de l'établissement statistique pour ensuite se rapporter au niveau de la compagnie statistique. Le diagramme suivant illustre bien le processus de répartition pour une entreprise complexe.



5.2 Variable de stratification

Dans l'EMCGD actuelle, les compagnies sont stratifiées selon le RBE. Celui-ci est estimé à tous les mois à partir de différentes sources de données fiscales. Par exemple, pour les employeurs, le RBE est estimé à partir d'une valeur annualisée du total des remises de paye. Étant donné les divergences entre le RBE et les ventes réelles, environ 17% des détaillants et 23% des grossistes sont classés dans la mauvaise strate de taille. Dans le cas des naissances, l'écart est beaucoup plus grand, seulement 36% des nouveaux détaillants et 30 % des nouveaux grossistes se retrouvent dans la strate appropriée. Les changements de taille sont surtout des changements entre les deux strates à tirage partiel et très peu d'unités passent de strate à tirage partiel à strate à tirage complet. Cependant, quelques unités mal classées peuvent avoir un effet dévastateur sur les estimations lorsque leurs ventes sont multipliés par un poids de sondage ne reflétant pas leur taille.

La différence entre le RBE et les ventes réelles s'explique par une différence de concept entre la mesure de taille utilisée, le revenu brut et la valeur collectée, les ventes. Le RBE a tendance à sous-estimer le niveau des ventes. Dans le cas des naissances, le RBE varie beaucoup les premiers mois et ce changement est

généralement à la hausse. Après quelques mois, le RBE se stabilise. Cette fluctuation est reliée en partie à l'utilisation dans le modèle d'une valeur annualisée basée parfois sur très peu de mois de données. De plus, dans bien des cas, la première remise de paye servant à calculer le premier RBE ne reflète pas la taille réelle de l'entreprise.

5.2.1 Etude de la variable de stratification

Cette étude vise à évaluer les différentes sources d'informations auxiliaires afin de mieux stratifier les unités. Les paragraphes qui suivent comparent l'efficacité lors de la stratification de l'utilisation des ventes TPS par rapport à l'utilisation du RBE.

Dans un premier temps, nous avons analysé la relation entre le niveau des ventes réelles (somme de 12 mois de avril 1998 à mars 1999) et les deux variables à l'étude. Le tableau 1 présente ces résultats. Le niveau de la TPS est très proche du niveau des ventes alors que le RBE sous-estime le niveau des ventes. Les corrélations observées sont plus élevées pour le commerce de détail que pour le commerce de gros. En considérant toutes les compagnies, les corrélations sont autour de 97% pour les deux variables dans le cas du commerce de détail. Pour le commerce de gros, les corrélations sont de 93% avec la TPS et de 91% avec le RBE.

Tableau 1
Corrélations* entre les ventes de TPS
et les ventes annuelles reportées de l'EMCGD

	Commerce de détail		Commerce de gros	
	TPS	**RBE**	**TPS**	**RBE**
Toutes les compagnies**	0.97	0.97	0.93	0.91
>10M$**	0.97	0.97	0.92	0.90
1-10M$**	0.89	0.73	0.64	0.54
<1M$ **	0.70	0.58	0.21	0.34
* corrélations avec données pondérées ** exclus les valeurs aberrantes (<1% des détaillants; 1% des grossistes)				

Le niveau de corrélation observé varie en fonction des groupes de commerce considérés. A quelques exceptions près, les corrélations entre les ventes des détaillants et les ventes de TPS pour chaque groupe de commerce demeurent autour

de 90% pour les compagnies avec des revenus annuels supérieurs à 1 million. Toutefois, pour les compagnies dont le revenu annuel est inférieur à 1 million les corrélations observées sont en général plus faibles et varient entre 27% (détaillants d'alcool, de vin et de bière) et 91% (détaillants de vêtements).

Chez les grossistes, la corrélation entre les ventes et la TPS varie entre 84% et 97% pour les compagnies aux revenus annuels supérieur à 10 millions. Les corrélations observées pour les compagnies dont le revenu annuel est entre 1 et 10 millions sont plus faibles et varient entre 48% et 82%. Dans le cas des compagnies dont le revenu annuel est inférieur à 1 million, les corrélations sont inférieures à 50% pour la très grande majorité des groupes de commerce.

Dans un deuxième temps, nous avons comparé le taux de mauvaise classification qui pourrait survenir en utilisant comme variable de stratification soit le RBE ou la TPS. Les résultats ont été validés en utilisant les ventes obtenues des répondants de l'EMCGD. En terme de nombre d'unités bien classées, l'utilisation de la TPS mène à un gain, par rapport au niveau actuel basé sur le RBE, de 6% pour le commerce de détail et de 9% pour le commerce de gros (Tableau 2).

Tableau 2

Pourcentage d'unités bien classées[1]

	RBE	**TPS**
Commerce de Détail	83%	89%
Commerce de Gros	77%	86%

[1] % d'unité assignées à la bonne strate selon les ventes réelles.

Le tableau 3 présente les résultats pour les unités les plus importantes de l'enquête. Pour le commerce de détail, le nombre d'unités mal classées est de 5.3% avec le RBE et baisse à 2.6% avec les ventes de TPS. On observe la même tendance chez les grossistes.

Tableau 3

Taux de mauvaise classification des unités importantes [1]

	RBE	TPS
Commerce de Détail	5.3%	2.6%
Commerce de Gros	7.8%	4.4%

[1] % unités assignées à une strate à tirage partiel alors quelles auraient dues être assignées à une strate à tirage complet selon leurs ventes réelles.

Toutefois, le RBE demeure plus efficace que les ventes de TPS pour classifier les nouveaux détaillants et les nouveaux grossistes (Tableau 4). La moins bonne performance de la TPS pour identifier le niveau des ventes des naissances est probablement en partie due à la méthode actuelle pour calculer les ventes annualisée (somme des remises disponibles pour les 12 derniers mois). Depuis août 2000, la période couverte pour les remises sera disponible sur le fichier TPS. Nous serons donc en mesure de produire une valeur annualisée qui sera sans doute supérieure.

Tableau 4

Taux de naissances bien classifiées selon la taille

	RBE	TPS
Commerce de Détail	36.6%	30.1%
Commerce de Gros	30.1%	29.2%

Nos analyses préliminaires nous permettent de conclure que l'utilisation des ventes TPS comme variable de stratification est très prometteuse. Le RBE a tendance à sous-estimer les ventes totales alors que le niveau des ventes TPS est en fait très proche des ventes réelles pour les compagnies avec des revenus supérieurs à 1 million. Ces compagnies représentent plus de 87% des ventes totales des détaillants et plus de 96% des ventes totales des grossistes. Dans le cas des plus petites compagnies et des naissances nous devons poursuivre nos recherches car la qualité de la relation reste faible pour les deux mesures. Nous explorons différents modèles de régression basés sur l'utilisation du RBE, des ventes TPS ou une combinaison des deux. Un des modèles des plus prometteurs est tout simplement l'utilisation d'une nouvelle mesure de taille définie comme étant le maximum de soit le RBE ou la TPS.

5.3 Estimation par calage

Les renseignements auxiliaires tirés de sources administratives permettent, à l'étape de l'estimation, d'améliorer l'efficacité de celle-ci. S'il existe une corrélation raisonnable entre les données auxiliaires et les variables d'intérêt, ces renseignements supplémentaires peuvent être intégrés au processus d'estimation grâce à plusieurs techniques d'estimation, y compris les estimateurs par calage (Deville et Särndal 1992). La stratégie du calage permet d'avoir recours à tout un choix d'estimateurs, y compris l'estimateur de régression généralisé (GREG), que nous utilisons dans notre étude.

L'estimateur GREG suppose un modèle linéaire entre la variable d'intérêt y et la variable auxiliaire x :

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k \qquad \text{où } E(\varepsilon_k)=0, \qquad Var(\varepsilon_k)=c_k\sigma^2, \quad Cov(\varepsilon_k,\varepsilon_l)=0 \quad \forall k \neq \ell \,. \quad (5.1)$$

Dans notre étude, y_k représente les ventes mensuelles de la compagnie k dans l'EMCGD. La variable auxiliaire x_k représente les ventes annuelles de biens assujettis à la TPS de la compagnie. Nous supposons que chaque résidu ε_k comporte la même structure de variance ($c_k=1$ pour chaque unité k). L'estimateur par le quotient, c'est-à-dire sans ordonnée à l'origine (β_0) et avec $c_k=x_k$, a été examiné, mais il a été exclu du reste de l'étude ultérieurement puisque β_0 différait appréciablement de 0 la plupart du temps.

5.3.1 Groupes modèles

Les groupes modèles forment une partition disjointe et exhaustive de la population à l'intérieur de laquelle il existe une bonne relation entre la variable auxiliaire et la variable d'intérêt. Les paramètres du modèle linéaire qui sous-tend le GREG sont estimés dans chaque groupe modèle. C'est pourquoi il est important qu'il y ait un nombre minimal d'unités échantillonnées dans chaque groupe modèle si l'on veut que l'estimation des paramètres de modèle soit suffisamment précise.

5.3.2 Estimateur GREG

L'estimateur GREG sert à fournir des poids « nouveaux » (w_k) les plus proches possible des poids du plan original (a_k) en fonction d'une mesure quelconque de la distance. Les w_k se laissent exprimer sous la forme $w_k=a_k g_k$. Les g_k sont calculés de façon que

$$\hat{\mathbf{X}}_p = \sum_{k\in s_p} a_k g_k \mathbf{x}_k = \sum_{k\in U_p} \mathbf{x}_k = \mathbf{X}_p \qquad (5.2)$$

où a_k est le poids du plan pour l'unité k, g_k est le facteur g pour l'unité k, $\mathbf{x}_k$ est le vecteur variable auxiliaire pour l'unité k, $\hat{\mathbf{X}}_p$ est le total estimatif du vecteur variable auxiliaire dans le groupe modèle p et $\mathbf{X}_p$ est le total de contrôle pour le groupe modèle p.

Pour calculer le facteur g, il faut connaître les valeurs de la variable auxiliaire pour les unités de l'échantillon et le total de la population de contrôle pour chacun des groupes modèles. Dans notre étude, les totaux de contrôle pour les ventes annuelles de biens assujettis à la TPS sont obtenus en prenant la somme, dans chaque groupe modèle, des valeurs x_k figurant dans les fichiers administratifs de la TPS provenant de l'ADRC. Plus précisément, le facteur g (à supposer que c_k=1) se laisse exprimer sous la forme :

$$g_k = 1 + (\sum_{k \in U_p} \mathbf{x}'_k - \sum_{k \in s_p} a_k \mathbf{x}'_k)(\sum_{k \in s_p} a_k \mathbf{x}_k \mathbf{x}'_k)^{-1}(\mathbf{x}_k) \qquad (5.3)$$

Le schéma de l'estimateur GREG fournit des estimations pour tout domaine de la population. L'expression de la variance correspond à la présentation de Särndal et coll. (1992).

5.4 Défis posés par la strtégie d'estimation de l'EMCGD

La présente section décrit l'application de l'estimateur GREG à l'EMCGD, de même que les diverses options envisagées. Il importe de noter que les unités de l'échantillon à tirage complet ont été exclues du calage. En ayant recours au calage dans l'EMCGD, notre but principal est de réduire l'influence des unités classées de façon erronée (relativement à leur taille) sur les estimations; or les unités de l'échantillon à tirage complet ne causent pas ce genre de problème puisqu'elles sont auto-représentatives.

5.4.1 Création de groupe de modèle

En plus de répondre aux critères énoncés à la section 5.3.1, les groupes modèles devraient idéalement se rapprocher le plus possible des domaines pour lesquels on

produit des estimations afin que l'on puisse éviter des estimations élevées de la variance. Par conséquent, le groupe modèle d'une unité donnée se prête à des variations d'un mois à l'autre puisque les unités peuvent changer de domaine au fil du temps. Par contre, dans le contexte d'une enquête mensuelle comportant un changement important d'un mois à l'autre, les groupes modèles devraient être aussi stables que possible. On garantira ainsi que le changement des estimations d'un mois à l'autre ne sera pas attribuable au fait que certaines unités ne relèvent pas du même groupe modèle que pour le mois précédent. Il faut tenir compte de ces deux préoccupations opposées lors de la création des groupes modèles.

Dans notre étude, les groupes modèles sont des groupements d'unités relevant d'un certain groupe de commerce et d'un certain code géographique. Ils contiennent au moins 25 unités échantillonnées actives. Chaque unité de la population doit être attribuée à un groupe modèle, et nous devons établir le total de la population de contrôle pour chaque groupe modèle. On peut obtenir des renseignements sur le groupe de commerce et le code géographique de trois sources.

Source 1 - Stratification : Chaque unité d'échantillonnage de l'EMCGD est stratifiée selon le groupe de commerce, la région géographique et la taille. Les renseignements sur cette stratification établie en décembre 1997 peuvent servir à créer nos groupes modèles. Cette source a l'avantage d'être stable au fil du temps; la stratification n'est pas mise à jour fréquemment dans l'EMCGD. L'inconvénient de cette méthode est que plus l'on s'éloigne de la dernière stratification, plus la stratification est périmée et différente des renseignements sur les plus récents domaines, ce qui risque d'entraîner une variance plus élevée dans les estimations.

Source 2 - Utilisation du plus récent univers du RE : Chaque mois, le RE crée un fichier de l'univers des populations du commerce de détail et de gros, reflétant les mises à jour provenant de sources administratives et des enquêtes. Ce fichier permet à l'EMCGD d'identifier les naissances et les cessations et de mettre à jour sa base de sondage et son échantillon. Ce sont là les plus récents renseignements sur les groupes de commerce et les régions géographiques pour ce qui est du RE. Cette information offre l'avantage de se rapprocher de l'information sur les domaines.

Source 3 - Utilisation des renseignements sur les domaines : Pour chaque compagnie statistique d'un échantillon, on obtient une combinaison de groupe de commerce et de code géographique à partir de l'information recueillie. Celle-ci est légèrement plus à jour que le plus récent univers du RE puisqu'elle intègre la collecte de données d'enquête plus récentes. Même si l'information n'est disponible que pour des unités échantillonnées, il serait possible d'utiliser les totaux de contrôle fondés sur la source 2, à supposer que celle-ci ne soit pas trop différente de la source 3.

Après avoir examiné ces trois options, nous avons décidé d'avoir recours à l'information de type stratification (source 1) pour établir les groupes modèles. Tout d'abord, la stabilité au fil du temps est une préoccupation majeure, à cause de l'ampleur du changement d'un mois à l'autre. Deuxièmement, notre but principal, pour ce qui est du calage dans l'EMCGD, est de réduire l'influence des unités classées de façon erronée relativement à leur taille. On peut y arriver même avec des groupes modèles fondés sur le groupe de commerce et le code géographique selon la stratification puisque la variable auxiliaire (total des ventes annuelles de biens assujettis à la TPS) est une mesure de la taille. Troisièmement, nous n'avons pas observé une bien grande différence dans les estimations de la variance en utilisant l'une ou l'autre des trois sources, bien qu'il soit possible que l'on observe une différence à l'avenir si la stratification n'est pas mise à jour périodiquement.

5.4.2 Traitement des unités d'échantillonnage inactives (cessations dans la base de sondage)

Pour chaque cessation que l'on identifie dans l'échantillon, la valeur de la variable d'intérêt y_k (ventes mensuelles) est de 0. Pour les x_k, toutes les unités jugées actives dans le RE (y compris les unités inactives et non identifiées comme telles) comportent habituellement une valeur x_k positive. La question est de savoir comment traiter les cessations puisque la base de sondage de l'EMCGD (voir la section 2) contient une plus forte proportion d'unités de cessations identifiées dans l'échantillon que hors échantillon. Intuitivement, on trouve raisonnable d'attribuer une valeur de 0 à la variable auxiliaire x_k pour les cessations (unités dans l'échantillon aussi bien que hors échantillon). Il en résulte cependant un problème au moment de calculer les totaux de contrôle de la variable auxiliaire.

Nous avons vu à la section 3 que, pour calculer les facteurs g, nous devons connaître le total de la population de contrôle pour la variable auxiliaire de chaque groupe modèle. Nous les calculons en prenant la somme de la valeur de la variable auxiliaire x_k dans chaque groupe modèle, y compris les cessations identifiées et non identifiées dans le RE. Si nous attribuons x_k=0 aux cessations identifiées (unités tant de l'échantillon que hors échantillon), ce total de contrôle est alors lui-même touché par le fait qu'il existerait proportionnellement plus de x_k=0 dans l'échantillon qu'à l'extérieur de l'échantillon. L'application du calage à ce type de population sans porter davantage attention aux cessations identifiées entraînerait une surestimation systématique des ventes mensuelles. Comme nous pouvons le constater pour le facteur g de l'estimateur par le quotient (qui n'est pas l'estimateur que nous utilisons, mais qui est plus facile à observer), nous avons pour le groupe modèle p,

$$g_{kp} = \frac{X_p}{\sum_{k \in s_p} a_k x_k} = \frac{X_p}{\hat{X}_p}, \qquad (5.4)$$

où a_k est le poids du plan pour l'unité k et x_k est la variable auxiliaire pour l'unité k. Ce facteur g sera toujours supérieur à 1 car il y aura plus de $x_k=0$ dans l'échantillon qu'à l'extérieur de l'échantillon dans chaque groupe modèle p. L'objectif est d'avoir des facteurs g qui sont proches de 1 en moyenne. Deux options sont offertes pour le traitement des cessations identifiées.

Option 1 Exclusion des cessations identifiées du calage. On impose simplement 1 à leur g_k. On applique ensuite le calage aux unités « jugées actives » seulement. Toutefois, si nous excluons les cessations identifiées du calage, nous devons modifier les totaux de contrôle connus puisqu'ils englobent la contribution des unités inactives et non identifiées qui sont largement présentes dans le volet hors échantillon de la base de sondage, mais non dans le volet échantillonné. Autrement, nous allons surestimer les ventes mensuelles parce que le total de contrôle de la variable auxiliaire sera trop grand comparativement au total estimé.

Afin de surmonter cette difficulté, nous corrigeons (déflation) le vecteur total de contrôle de la variable auxiliaire de chaque groupe modèle à l'aide d'un facteur qui représente le rapport entre le nombre estimatif d'unités actives dans chaque groupe modèle et le nombre d'unités actives dans la base de sondage (y compris les cessations non identifiées), de façon à obtenir le total de contrôle corrigé pour chaque groupe modèle p :

$$\mathbf{X}^*_p = \mathbf{X}_p \times \frac{\hat{N}_{p(actives)}}{N_{p(actives)}} \tag{5.5}$$

où $\hat{N}_{p(actives)} = \sum_{s_p} a_k I_k$ (I_k=1 si k est considéré comme une unité active; I_k =0 autrement) est le nombre estimatif d'unités actives dans le groupe modèle p, et $N_{p(actives)} = \sum_{U_p} I_k$ (I_k=1 si k est une unité active; I_k =0 autrement) est le nombre d'unités actives dans la base de sondage pour le groupe modèle p. Le recours à ce facteur de déflation est un prolongement du processus proposé par Hidiroglou et coll. (1995). Le total de contrôle corrigé $\mathbf{X}^*_p$ servira alors de total de contrôle dans l'estimateur de régression.

Le facteur proposé fait diminuer le total de contrôle de 20 % environ dans chaque groupe modèle du secteur du commerce de détail, et de 30 % environ dans chaque groupe modèle du secteur du commerce de gros. Cela nous permet d'avoir de meilleurs totaux de contrôle. Un inconvénient de cette façon de procéder est que les nouveaux totaux de contrôle sont maintenant des valeurs estimatives comportant une variabilité inconnue, et que nous ne tenons pas compte de cette variabilité dans

l'estimation de la variance. En réalité, l'hypothèse qui sous-tend notre estimateur de la variance est que les totaux de contrôle sont constants.

Option 2 Inclusion des cessations identifiées dans le calage mais moyennant un traitement spécial. La variable auxiliaire (total des ventes annuelles de biens assujettis à la TPS) est imputée à l'aide d'une valeur positive pour toutes les cessations identifiées à moins qu'une valeur soit accessible (comme c'est possible pour les unités hors du champ d'observation). Le processus d'imputation se déroule à l'aide d'un modèle de régression fondé sur une valeur historique du revenu figurant dans des versions antérieures du RE. Tous les totaux de contrôle sont alors recalculés à l'aide de ces valeurs imputées pour des cessations identifiées et à l'aide des valeurs x originales des autres unités.

Un avantage de cette façon de procéder est qu'il n'est pas nécessaire de rajuster les totaux de contrôle et que le facteur g a de nouveau une valeur moyenne de 1. Un inconvénient est que le réglage du modèle qui sous-tend l'estimateur de régression est moins efficace puisque nous avons des unités inactives (y_k=0 et x_k>0) dans le modèle. Il devrait en résulter des estimations plus élevées de la variance. De plus, il est à noter que nous ne tenons pas compte de la variabilité des x_k imputés dans l'estimation de la variance. Pour le moment, les deux options restent possibles pour notre étude.

5.5 Résultats

Nous avons retenu deux stratégies pour l'estimation. Un estimateur de régression pour lequel les groupes modèles se fondent sur la stratification initiale est utilisé pour les deux stratégies, la différence entre les deux relevant du traitement des cessations , suivant l'option 1 et l'option 2 décrites à la section 5.4.2. Nous effectuons toujours une estimation par domaine. Pour les deux stratégies, nous avons calculé des estimations et des estimations de la variance. Le tableau 5 indique les intervalles de confiance pour un niveau de confiance de 95 % et les coefficients de variation pour certains domaines du secteur du commerce de détail. Le tableau 6 présente les résultats pour certains domaines du secteur du commerce de gros. Ces intervalles de confiance s'appliquent à des estimations mensuelles des ventes pour le mois de référence avril 1999. Toutes les valeurs sont exprimées en milliards. Les intervalles de confiance obtenus de l'estimateur par dilatation simple sont inclus à titre de comparaison. Comme il a été expliqué à la section 2, l'estimateur par dilatation simple est non biaisé, tandis que l'estimateur de régression est approximativement non biaisé. On peut constater que les intervalles de confiance des deux estimateurs

de régression donnent lieu à des estimations qui ne sont pas appréciablement différentes de celles de l'estimateur par dilatation simple.

Comme on pouvait s'y attendre, l'estimateur de régression, pour les deux options de traitement des cessations, produit dans presque chaque cas des estimations comportant des coefficients de variation inférieurs à ceux de l'estimateur par dilatation simple. Il en va de même pour presque chaque groupe de commerce et chaque région géographique des secteurs du commerce et de détail. De plus, lorsque l'on exclut les cessations du calage (option 1), le modèle de régression qui sous-tend l'estimateur de régression est meilleur, et les estimations correspondantes de la variance comportent donc une variance moins élevée que celle de l'estimateur de régression lorsque les cessations sont incluses dans le calage (option 2).

Tableau 5
Intervalles de confiance avec α=0.05 (IC) et coefficients de variation (CV) des estimations des ventes mensuelles de détail d'avril 1999

Commerce de détail, Secteurs	Estimateur de régression Option 1		Estimateur de régression Option 2		Estimateur par dilatation simple	
	IC (en milliards)	CV (%)	IC (en milliards)	CV (%)	IC (en milliards)	CV (%)
Supermarché et épicerie	[4,26, 4,39]	0,74	[4,30, 4,45]	0,90	[4,25 , 4,47]	1,24
Pharmacies	[1,04, 1,11]	1,67	[1,08, 1,16]	1,88	[1,05 , 1,14]	2,20
Véhicules récréatifs et automobile	[5,91, 6,47]	2,34	[6,22, 6,94]	2,80	[6,08 , 6,96]	3,47

Tableau 6
Intervalles de confiance avec α=0.05 (IC) et coefficients de variation (CV) des estimations des ventes mensuelles de gros d'avril 1999

Commerce de gros, Secteurs	Estimateur de régression Option 1		Estimateur de régression Option 2		Estimateur par dilatation simple	
	IC (en milliards)	*CV* (%)	IC (en milliards)	CV (%)	IC (en milliards)	CV (%)
Pièces et accessoires de véhicules automobiles	[6,15 , 6,48]	1,34	[6,12 , 6,57]	1,82	[5,89 , 6,22]	1,40
Alimentation	[4,36 , 4,50]	0,80	[4,43 , 4,63]	1,12	[4,39 , 4,66]	1,52
Électronique, ordinateurs et machines connexes	[2,78 , 2,96]	1,58	[2,69 , 2,96]	2,48	[2,61 , 2,93]	2,99

5.6 Estimation par calage sur plusieurs mois

L'étude sur un mois de données donnaient des résultats très intéressants. Nous avons donc décidé de poursuivre l'étude en utilisant plusieurs mois de données. La période de janvier 1999 à juillet 1999 a été considérée.

5.6.1 Méthodes

Pour poursuivre nos tests et donc soumettre les deux options retenues sur plusieurs mois de données, deux objectifs, en quelque sorte opposés, sont visés :

1) Stabilité dans le temps, c'est-à-dire éviter que le facteur de calage varie grandement d'un mois à l'autre. Pour ce faire, il faut que l'estimateur retenu et ses options soient le moins affectés par le passage d'un mois à l'autre. Par exemple, si on change le fichier de ventes de TPS d'un mois à l'autre, il est évident que le facteur de calage changera conséquemment d'un mois à l'autre. Nous ne voulons pas introduire des sauts dans la série qui seraient simplement

dus à des changements au facteur de calage. En résumé, cet objectif se préoccupe d'abord et avant tout de l'aspect longitudinal.

2) Meilleur calage possible, c'est-à-dire utiliser de façon optimale l'information auxiliaire disponible pour effectuer la meilleure estimation par calage. Cet objectif se préoccupe d'abord et avant tout de l'aspect transversal.

Pour limiter notre charge de travail, nous avons décidé, dans un premier temps, de viser le meilleur calage possible. Ainsi, pour chaque mois d'estimation, la source la plus à jour disponible est utilisée. Ceci signifie que la valeur de la variable auxiliaire associée à chaque unité échantillonnée peut varier d'un mois à l'autre, de même que les totaux de contrôle au niveau des groupes modèles. Pour les estimations des mois de janvier 1999 à juillet 1999, nous avons utilisé la TPS du mois précédent ou sinon celle la plus récente. Les estimations ne couvrent uniquement que la partie commerce de détail.

Pour le traitement des cessations, nous avons retenu les deux options des tests sur le mois d'avril c'est-à-dire, les inclure ou les exclure. Le traitement des naissances vient ajouter une nouvelle dimension à l'estimation par calage lorsque répétée sur plusieurs mois. Nous n'avions pas eu à tenir compte de cet aspect dans les tests sur le mois d'avril 1999. En ayant choisi de réaliser le meilleur calage possible, nous rencontrons immédiatement notre objectif d'inclure les naissances puisque toutes les étapes sont refaites pour chaque mois indépendamment des autres mois.

5.6.2 Analyse des résultats

Les résultats sont présentés dans les graphiques 1 et 2. Les variables REG1 et REG2 correspondent respectivement à l'estimateur par régression des options 1 et 2 pour le traitement des cessations. Rappelons que l'option 1 consiste à exclure les cessations du calage alors que ces dernières sont incluses dans l'option 2. Les deux estimations sont comparées à l'estimation publiée et l'on retrouve les limites supérieures et inférieures de l'intervalle de confiance (95%) de cette dernière.

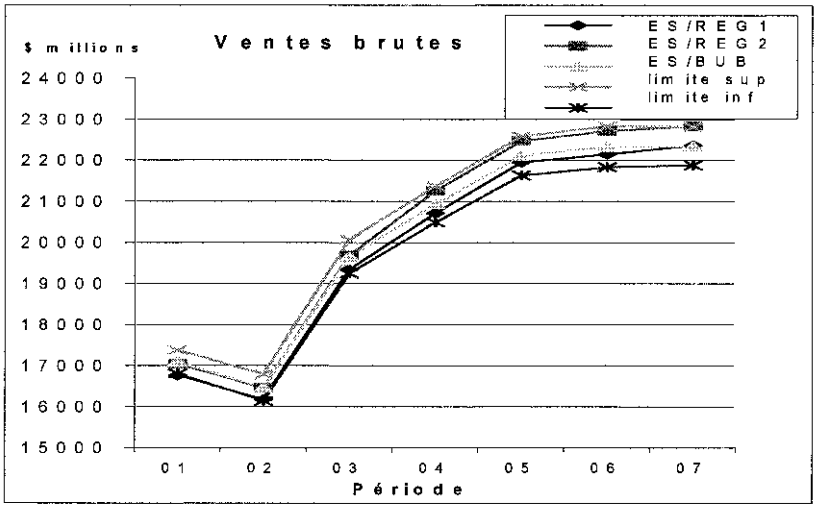
Le graphique 1 présente les résultats pour tous les groupes de commerce à l'échelle nationale. Nous constatons que les deux estimateurs se retrouvent dans les limites de l'intervalle de confiance de l'estimation publiée. Cependant, il est important d'étudier les résultats au niveau plus détaillé. Le graphique 2 présente les résultats pour les détaillants du groupe « autres magasins de vêtement ». La corrélation entre la variable TPS et les ventes de l'EMCGD est de 91%. Il s'agit du groupe de commerce qui possède la plus forte corrélation au niveau des petites compagnies à tirage partiel. Nous constatons que les résultats provenant des deux estimateurs par régression sont en dehors de l'intervalle de confiance et que le niveau des ventes estimées est supérieur à la valeur publiée. Cependant, pour plusieurs autres groupes

de commerce, les estimations par calage sont toutes à l'intérieur de l'intervalle de confiance.

L'EMCGD publie toujours au niveau des groupes de commerce. Il est donc important d'avoir un estimateur très stable à ce niveau. Ainsi, il n'est pas clair si le problème est causé par l'estimation par calage ou par l'estimateur par dilatation. Nos analyses nous permettent de croire que le modèle de calage est adéquat mais que la variable auxiliaire n'est pas très stable et varie d'un groupe de commerce à un autre. Il est donc important de poursuivre les analyses pour mieux comprendre les enjeux que représentent l'utilisation de la variable TPS. L'une des avenues que nous explorons est de faire une expérience contrôlée et de fixer des paramètres. Il faudrait faire une simulation et fixer les ventes et calculer l'estimateur par calage dans le temps. Ceci permettrait de mesurer la robustesse de notre modèle.

Graphique 1

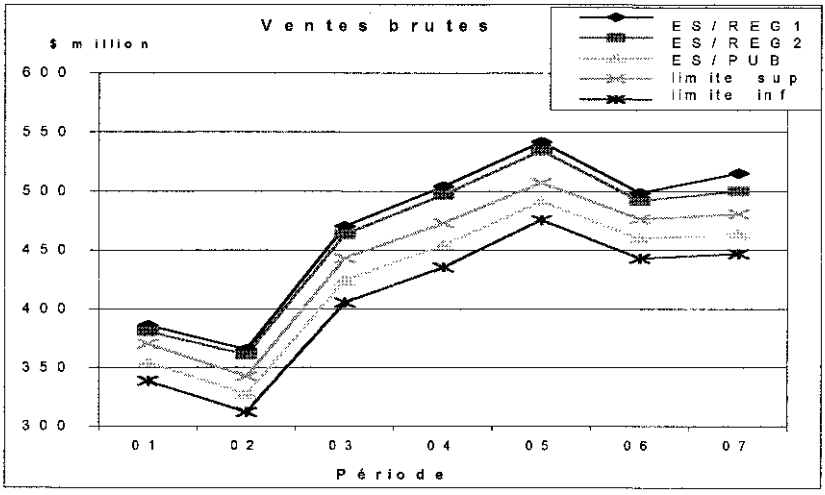
Comparaison des estimations calibrées avec les estimations publiées pour le commerce détail
Tous les groupes de commerce



Graphique 2

Comparaison des estimations calibrées avec les estimations publiées pour le commerce détail

Groupe 070 – Autres magasins de vêtements



6. Conclusion

L'utilisation des données de la TPS dans le cadre du remaniement de l'EMCGD semble être très prometteuse. Au niveau de la stratification, les gains sont importants puisque les problèmes d'unités mal classées représentent souvent une lourde charge de travail d'un mois à l'autre. Cependant, l'analyse des estimations provenant de plusieurs mois de données ne nous permet pas de comprendre les mécanismes qui régissent les deux types d'estimateurs, par dilatation et par calage. Il est nécessaire de poursuivre les analyses afin de mieux comprendre les faiblesses des données de la TPS. L'estimateur par calage réduit le coefficient de variation mais, la qualité varie beaucoup d'un groupe de commerce à l'autre à cause de différentes périodes de remise, du type de compagnies et, pour les compagnies complexes, de la réconciliation avec le RE. Il faudra éventuellement choisir l'une des deux options pour les cessations.

Le projet d'expérience contrôlée permettra de mieux évaluer les estimateurs. Parallèlement à ce projet, il est primordial d'améliorer la qualité des données de la TPS avant de se commettre à utiliser un estimateur par calage. D'ailleurs, Statistique Canada a entrepris un vaste programme pour améliorer la qualité. Ce nouveau programme vise à développer une base de données longitudinales, un système sophistiqué de vérification et d'imputation des données aberrantes et un processus pour simplifier la réconciliation avec le RE. Puisque le remaniement va être complété en 2003, il est fort possible que l'estimateur par calage soit reconsidéré lorsque tous les aspects de la TPS seront mieux maîtrisés.

7. Remerciements

Les auteurs tiennent à remercier tous ceux qui ont contribué à la réalisation de la partie méthodologique de cette enquête dont, Julie Girard, Martin St-Pierrre, Michel Ferland et Naïma Gouzi. Les auteurs tiennent également à remercier Michel Hidiroglou et Jocelyn Tourigny pour leurs précieux commentaires.

8. Bibliographie

Bissonnette, J., I. Marchand, M. St-Pierre, and J. Trépanier (1998), "Amélioration de la série d'estimations mensuelles des ventes au détail au moyen d'un estimateur par calage", *1998 Proceedings of the Survey Methods Section,* Statistical Society of Canada, 151-157.

Deville, J.C., and C.-E. Särndal (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, 376-382.

Hidiroglou, M. A. (1986). "The Construction of a Self-Representing Stratum of Large Units in Survey Design", *The American Statistician*, 40, 27-31.

Hidiroglou, M. A., Choudhry, G. H., et Lavallée, P. (1991). Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises. *Techniques d'enquête, 17, 211-227.*

Hidiroglou, M. A., C.-E. Särndal, and D. A. Binder (1995), "Weighting and Estimation in Business Surveys", *Business Survey Methods*, New York : Wiley, pp. 477-502.

Lavallée, P. (1996), "Frame Update Problems with Panel Surveys", *Collection of Papers, Statistical Day '96 Economy - Slovenia - European Union*, Statistikni Urad Republike Slovenije, Radenci, November 1996, pp. 252-261.

Särndal, C.-E., B. Swensson, and J. Wretman (1992), *Model Assisted Survey Sampling,* New York : Springler-Verlag.

Trépanier, J., C. Babyak, I. Marchand, J. Bissonnette, and M. St-Pierre (1998), "Enhancements to the Canadian Monthly Wholesale and Retail Trade Survey", *1998 Proceedings of the Section on Survey Research Methods*, American Statistical Association, 487-492.