

LA CONVERSION DE DONNÉES HISTORIQUES SELON UN NOUVEAU SYSTÈME DE CLASSIFICATION POUR L'ENQUÊTE MENSUELLE SUR LE COMMERCE DE GROS ET DE DÉTAIL

Susie Fortier¹

RÉSUMÉ

Les systèmes de classification industrielle fournissent un cadre commun de concepts permettant de décrire l'activité économique. Au cours des années, Statistique Canada a utilisé différentes versions de la Classification type des industries (CTI) et, plus récemment, le Système de classification des industries de l'Amérique du Nord (SCIAN). Les changements de système de classification créent des brisures dans les séries chronologiques. Afin de préserver la continuité de ces séries, il devient nécessaire de convertir les estimations historiques obtenues sous un ancien système en estimations historiques selon la classification plus récente. Une enquête majeure de Statistique Canada, l'Enquête mensuelle du commerce de gros et de détail (EMCGD), subit actuellement un remaniement complet qui inclut l'utilisation de la nouvelle classification et doit relever le défi d'une telle conversion. La version actuelle de l'EMCGD a été développée à la fin des années 1980 afin de produire principalement des estimations sur les ventes et les stocks pour des secteurs industriels définis par la CTI de 1980. L'enquête remaniée produira dorénavant des estimations selon le SCIAN à partir de 2004. Diverses options ont été considérées afin de convertir les données des secteurs de la CTI en estimations pour les secteurs du SCIAN. Les avantages et inconvénients de ces options seront discutés. L'option choisie, étant donné les contraintes reliées à l'EMCGD, est présentée.

MOTS CLÉS : Rétropolation ; système de classification ; coefficient de conversion.

ABSTRACT

Industrial classification systems provide a conceptual framework that allows the description of economic activities. Throughout the years, Statistics Canada has used different versions of the Standard Industrial Classification (SIC) system and the North American Industry Classification System (NAICS) for industrial classification. Changes in industrial classification systems create disruptions in associated time series of population estimates. To maintain continuity, existing series under the previous classification must be converted according to the new classification. The Monthly Wholesale and Retail Trade Survey (MWRTS), a major survey conducted by Statistics Canada, is in the midst of a redesign and faces the challenges of classification conversion. The current MWRTS was developed in the late 1980's to produce sales and inventories estimates for SIC-based industrial sectors. The redesigned survey will produce NAICS-based estimates starting in 2004. This paper summarizes the options considered for backcasting the SIC-based data into NAICS-based estimates. The pros and cons of each method are discussed and the final strategy is presented within the specific context of the MWRTS.

KEY WORDS: Backcasting; Classification system; Concordance coefficient

1. INTRODUCTION

L'Accord de libre-échange, signé en janvier 1994, a créé le besoin d'une classification des activités économiques commune au Canada, au Mexique et aux États-Unis. Le Système de classification des industries de l'Amérique du Nord (SCIAN) de 1997 a été conçu à cette fin par les organismes statistiques des trois pays (Statistique Canada 2002). Le SCIAN s'appuie sur un cadre conceptuel fondé sur la production ou l'offre ; les établissements y sont groupés par classe en fonction de la similitude des procédés qu'ils appliquent à la production de biens et de services. L'utilisation du nouveau système permet une meilleure comparaison des statistiques industrielles entre les trois pays. Toutefois, ceci interrompt la continuité historique des séries nationales basées sur d'autres systèmes de classification. Depuis 1948, Statistique Canada utilisait principalement le système de la Classification type des industries (CTI) et ses différentes révisions (en 1960, 1970 et 1980). La conversion au SCIAN touche plusieurs enquêtes majeures de Statistique Canada, dont l'Enquête mensuelle sur le commerce de gros et de détail (EMCGD).

¹ Susie Fortier (susie.fortier@statcan.ca), Immeuble R.-H.Coats, 11^e étage, Ottawa (Ontario), Canada, K1A 06T.

Depuis 1988, les données de l'EMCGD sont produites en fonction des définitions de la CTI version 1980. L'enquête a récemment été remaniée pour produire des estimations de qualité selon le SCIAN à coûts réduits. Ce remaniement permet également de réduire le fardeau de réponse, d'implanter des systèmes informatiques plus performants et de favoriser l'utilisation de données administratives nouvellement disponibles, telles que les ventes selon la Taxe sur les produits et services (TPS) perçue (Bérard 2001). Les premiers résultats de l'enquête remaniée seront publiés en 2004. Ils seront accompagnés d'estimations historiques mensuelles selon le SCIAN obtenues grâce à la conversion des données historiques recueillies selon la CTI.

Cet article présente les défis associés à la production de données historiques selon un nouveau système de classification. La section 2 définit le commerce de gros et le commerce de détail selon chacun des deux systèmes de classification. Les différentes méthodes considérées pour la rétopolation sont brièvement discutées dans la section 3. La section 4 identifie la méthode choisie et traite de son implantation.

2. DÉFINITIONS DU COMMERCE DE GROS ET DU COMMERCE DE DÉTAIL

Sous la CTI et le SCIAN, les secteurs du commerce de gros et du commerce de détail ont la même fonction principale ; soit l'achat de marchandises aux fins de revente. La distinction entre les deux secteurs dépend par contre du système de classification utilisé (Meyer 2001). Selon la CTI, la distinction est fondée sur la catégorie de clients. Les marchandises des détaillants sont destinées au public, pour usage personnel ou ménager. Les grossistes revendent des marchandises à des détaillants, des industriels, des commerçants, des établissements publics, des agriculteurs, des professionnels ou à d'autres grossistes. La différence entre le commerce de gros et le commerce de détail, selon le SCIAN, réside plutôt sur le processus de production utilisé, à savoir le fait que les ventes sont effectuées dans un magasin ou non. Ce changement de concept entraîne un changement de secteurs pour certains magasins. Par exemple, les magasins d'ordinateurs, les marchands de matériaux de construction, y compris les centres de rénovation, ainsi que les magasins de fournitures de bureau et de papeterie, qui étaient tous des grossistes selon la CTI, deviennent des détaillants sous le SCIAN. Les établissements dont l'activité principale est l'installation et la réparation, activité appartenant au secteur du commerce de détail selon la CTI, sont maintenant classés dans le secteur des services selon le SCIAN. Le tableau 1 présente un aperçu des mouvements entre les secteurs de la CTI et du SCIAN. On y présente la distribution des ventes des secteurs du commerce de gros et du commerce de détail de la CTI selon la nouvelle classification du SCIAN.

Tableau 1 – Pourcentage moyen de ventes selon le secteur d'industrie (1998-2001)

Secteur selon la classification	SCIAN : commerce de détail	SCIAN : commerce de gros	SCIAN : autres secteurs	total
CTI : commerce de détail	96,3%	-	3,7%	100%
CTI : commerce de gros	4,2%	94,3%	1,5%	100%

L'activité reste fortement concentrée dans le même secteur industriel. Par contre, il faut souligner que l'EMCGD publie ses données à un niveau plus raffiné que le secteur, soit les groupes de commerce. Alors que l'enquête courante sous la CTI a 11 groupes de commerce d'intérêt pour le commerce de gros et 18 pour le commerce de détail, l'enquête remaniée sous le SCIAN en aura respectivement 15 et 19. Les changements entre les anciens et les nouveaux groupes de commerce sont nombreux et variés. Afin de permettre l'analyse des 34 nouvelles séries selon le SCIAN, il est nécessaire d'estimer leurs valeurs historiques à l'aide de la rétopolation des données obtenues sous la CTI.

3. MÉTHODES CONSIDÉRÉES

3.1 Mise en contexte

Le Registre des entreprises (RE) de Statistique Canada est une base de données sur la population des entreprises canadiennes. Les établissements de ces entreprises sont répertoriés sur le RE et classés entre autres selon leur type d'activités industrielles et leur classification géographique. L'EMCGD recourt à un plan d'échantillonnage stratifié aléatoire simple sans remise qui est en place depuis 1988 et qui utilise le RE comme base de sondage. La stratification industrielle est en fonction de la CTI. Depuis 1998, les établissements sur le RE sont classés selon la CTI et le SCIAN. Cette double classification permet donc d'obtenir facilement des estimations selon le SCIAN même si le plan de sondage

est basé sur la CTI. Il suffit d'assigner la classification SCIAN aux unités échantillonnées selon l'information disponible sur le RE et d'effectuer une estimation par domaine. La précision des estimations est toutefois difficile à contrôler.

Puisque toutes les unités de la population sont classées selon les deux systèmes, il aurait également été possible de procéder par poststratification afin d'améliorer la précision des estimations selon le SCIAN. Cette option a été rejetée car la classification SCIAN des unités hors échantillon est de moins bonne qualité que celle des unités échantillonnées ; et ce particulièrement en 1998. Les estimations selon le SCIAN seront plutôt produites à l'aide de l'estimation par domaine pour la période allant de 1998 jusqu'à la mise en œuvre de l'enquête remaniée. Certains utilisateurs, comme le système de comptabilité nationale, ont besoin de séries historiques antérieures à janvier 1998. La rétopolation doit commencer en janvier 1991 pour le commerce de détail, alors que pour le commerce de gros, elle peut commencer en janvier 1993. Différentes options pour la conversion de données de la CTI au SCIAN ont été étudiées pour plusieurs enquêtes de Statistique Canada. (Hidiroglou, Quenneville et Huot 2001). Pour l'EMCDG, deux approches ont été considérées.

3.2 Approche « micro » : ajustements des microdonnées

Dans l'approche « micro », l'objectif est d'assigner à chaque unité échantillonnée selon la CTI un groupe de commerce selon le SCIAN puis de totaliser les estimations. Le défi réside dans la reclassification pour la période antérieure à 1998. Certains codes de la CTI ont un lien simple (un à un) avec un code du SCIAN. Une reclassification automatisée peut alors être effectuée. Pour les codes avec liens multiples, une reclassification manuelle par des spécialistes en la matière aurait été idéale mais impraticable, étant donné les contraintes de temps et de budget. Diverses méthodes d'imputation utilisées à des fins de reclassification ont été étudiées. Par exemple, pour les unités qui sont encore présentes sur la base de sondage en 1998, nous avons considéré la possibilité d'imputer les mêmes codes du SCIAN pour les années antérieures. Cette option est efficace si aucun changement d'activité ne s'est produit depuis 1991. Toutefois, seulement 55 % des unités visées au début de la période de rétopolation (soit 1991 ou 1993) sont présentes en janvier 1998. De plus, les unités toujours vivantes ont bien souvent subi des changements de structure importants mettant en doute l'hypothèse de stabilité de l'activité principale. Pour les unités qui ne sont plus sur la base de sondage en 1998, un code du SCIAN peut être attribué de manière probabiliste. Pour chaque code de la CTI, les probabilités d'assignation à un code du SCIAN sont déterminées de manière empirique, c'est-à-dire par la fréquence de chacun des liens CTI-SCIAN présents en 1998 et les années suivantes. Enfin, l'assignation d'un code du SCIAN pour les unités classées sous un code de la CTI avec un lien « un à plusieurs » peut se faire à l'aide d'une méthode de partage. Un certain pourcentage de la variable d'intérêt est alors recodé dans chacun des codes du SCIAN éligibles. Les facteurs de partage sont également dérivés des données pour lesquelles la classification est connue sous les deux systèmes.

L'avantage principal de l'approche « micro » est au niveau de la précision. Les méthodes considérées sont cependant assez complexes. De plus, on suppose que les microdonnées sous la CTI concordent parfaitement aux données agrégées. Ce n'est pas le cas pour l'EMCGD. Entre autres, une restratification importante a eu lieu en décembre 1997 et a causé un saut dans les séries historiques. Ce saut a été lissé graduellement dans les mois antérieurs à la restratification à l'aide d'ajustements « macro ». L'utilisation de l'approche « micro » nécessiterait donc un traitement supplémentaire pour les ajustements « macro » liés à la restratification et pour d'autres ajustements « macro » minimes mais non négligeables. Nous avons rejeté cette approche.

3.3 Approche « macro » : coefficients de conversion

Dans l'approche « macro », les unités n'ont pas besoin d'être reclassifiées individuellement. Pour chaque groupe d'intérêt selon le SCIAN, on utilise plutôt une combinaison linéaire pondérée du total de chacun des groupes d'intérêt selon la CTI. Le total $X_j(a, m)$ du groupe de commerce j selon le SCIAN pour l'année a et le mois m est donné par

$$X_j(a, m) = \sum_i \alpha_{ij}(a, m) X_i(a, m)$$

où $X_i(a, m)$ est la somme du groupe de commerce i selon la CTI. Les poids de la combinaison linéaire, soit les coefficients de conversion $\alpha_{ij}(a, m)$, représentent le pourcentage du total du groupe CTI i qui est attribué au groupe SCIAN j . Ils sont estimés à l'aide des données pour lesquelles la double classification est connue. Pour l'EMCGD, ces données sont celles de janvier 1998 et des mois suivants.

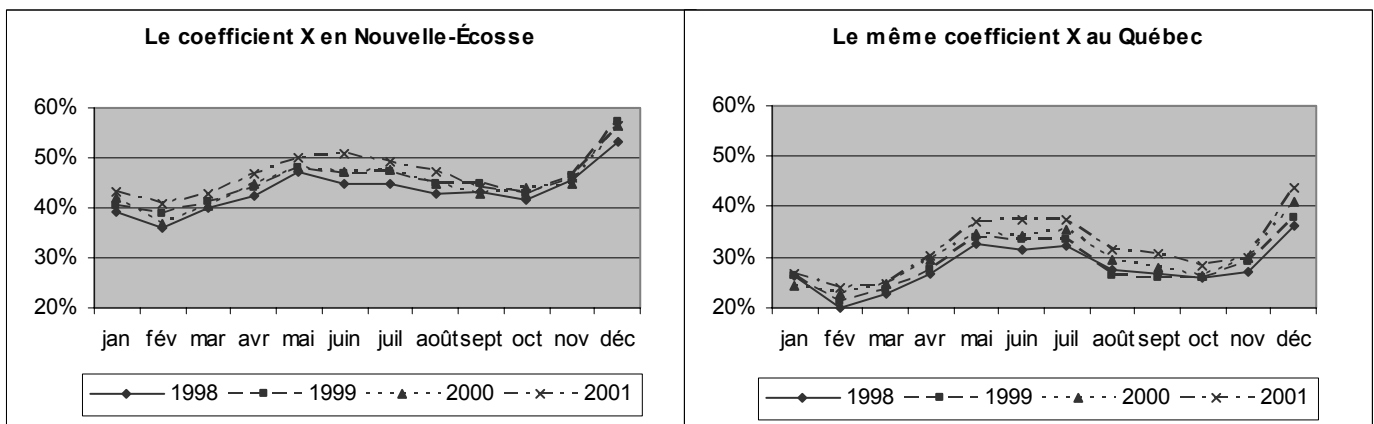
L'avantage majeur de cette approche est qu'aucun traitement supplémentaire n'est nécessaire pour les ajustements à des niveaux agrégés (lissage ou autres) effectués aux séries sous la CTI. Son désavantage principal découle de l'introduction d'erreurs lors de l'estimation des coefficients de conversion. Les sources d'erreurs sont détaillées dans la section 4.2. L'approche « macro » a été retenue pour l'EMCGD.

4. IMPLANTATION

4.1 Estimation des coefficients de conversion

Les valeurs des coefficients de conversion peuvent être dérivées des données échantillonnées pour janvier 1998 et les mois suivants. Les coefficients antérieurs doivent être estimés à l'aide des coefficients connus. Environ 1000 séries de coefficients de conversion, obtenues par le croisement des groupes de la CTI et des groupes du SCIAN, ont été étudiées à l'aide de 48 observations (janvier 1998 à décembre 2001). Les coefficients jugés invalides par des spécialistes en la matière ou en valeur absolue inférieurs à 0,3 % ont été éliminés et réalloués. Les 230 séries de coefficients non nuls restantes ont été analysées graphiquement pour détecter la présence de différences régionales, de saisonnalité ou de valeurs aberrantes. Puisque les différences régionales et la saisonnalité étaient facilement décelables (graphique 2), il a été décidé d'estimer les coefficients mensuels pour les années 1991 à 1997 à l'aide de la moyenne des coefficients calculés sur les mois homologues des années 1998 à 2001 et ce, pour chaque région.

Graphique 2 - Coefficients de conversion dans le temps



Les coefficients de conversion peuvent donc s'écrire sous la forme

$$\hat{\alpha}_{ij}^r(1991, m) = \dots = \hat{\alpha}_{ij}^r(1997, m) = \frac{1}{k} \sum_{a=1998}^{2001} \delta_{ij}^r(a, m) \alpha_{ij}^r(a, m) \quad \text{pour } m = 1, \dots, 12$$

où la variable k est la somme sur les 4 années des indicatrices $\delta_{ij}^r(a, m)$ définies par

$$\delta_{ij}^r(a, m) = \begin{cases} 0 & \text{si } \alpha_{ij}^r(a, m) \text{ est jugé aberrant;} \\ 1 & \text{sinon.} \end{cases}$$

Les valeurs aberrantes sont ainsi retirées du calcul de la moyenne. Les coefficients obtenus sont réajustés pour sommer à 100% pour chaque combinaison d'année a , de mois m , de région r et de groupe de commerce i selon la CTI.

4.2 Sources d'erreurs

Différentes sources d'erreurs affectent l'efficacité de l'approche « macro » (Hidioglou et al. 2001). Une première source d'erreur potentielle est la base de sondage elle-même. Une erreur de classification selon la CTI ou le SCIAN pour un mois donné entre 1998 et 2001 affecte évidemment le mois en question ainsi que tous les mois homologues des années 1991 à 1997. Afin de réduire l'impact des unités mal classées, les gros contributeurs ont été vérifiés manuellement et recodés au besoin. Les corrections apportées à la base de sondage depuis 1998 ont été répertoriées dans le but d'évaluer leur impact ; lorsque jugés nécessaires, des ajustements aux séries estimées sous le SCIAN ont été apportés.

Le second type d'erreur provient de l'utilisation des coefficients de conversion calculés sur des années récentes (1998-2001) pour estimer les coefficients de conversion des années antérieures. Cette méthode est efficace si la répartition selon le SCIAN est stable d'une année à l'autre ; si elle ne l'est pas, nous considérons quand même que le risque d'erreur est plus faible en 1997 qu'en 1991. L'hypothèse de stabilité a été acceptée dans la majorité des cas. Il faut cependant proscrire l'utilisation des coefficients de conversion estimés uniquement par la moyenne lorsqu'une industrie a subi un changement important tels que les centres de rénovation à grande surface. Puisque que le secteur de la rénovation a connu une croissance marquée au cours des dernières années et que les centres à grandes surfaces sont récents dans l'économie canadienne, leur contribution au groupe de commerce de la CTI en 1998 représente mal la contribution qu'ils avaient au début des années 1990. On doit donc ajuster les coefficients de conversion qui leur sont associés à la baisse pour les années 1991 à 1997. Les centres de rénovation représentent une partie d'un groupe de commerce de la CTI. Lorsque leur contribution est révisée à la baisse, la contribution des unités résiduelles doit être révisée à la hausse. La valeur de l'ajustement lui-même est basée sur l'analyse des experts et les résultats d'une classification partielle sous le SCIAN au niveau micro. Ce type d'ajustement permet de modéliser les variations dans le temps des coefficients. D'autres éléments liés à la saisonnalité des coefficients tels les jours ouvrables et l'effet de Pâques n'ont pas été considérés étant donné le peu d'observations disponibles au moment de l'analyse.

Une source d'erreur additionnelle découle de l'utilisation de coefficients de conversion basés sur une variable et utilisés sur une autre. L'EMCGD a deux variables d'intérêt ; les ventes et les stocks – notons que les stocks ne sont présentement publiés que pour les grossistes selon la CTI. Toute l'analyse des coefficients de conversion a été effectuée au niveau des ventes. Au lieu de convertir selon le SCIAN les séries des stocks selon la CTI, nous avons converti les séries rétrolées des ventes en stocks à l'aide de ratios. Des ratios stocks/ventes sont calculés pour chaque mois depuis 1993 au niveau des groupes de commerce CTI et appliqués aux groupes de commerce SCIAN correspondants. Les correspondances sont établies en fonction de la règle suivante : un groupe de commerce selon le SCIAN correspond à un groupe de commerce selon la CTI si ce dernier contribue à plus de 99 % du total du groupe selon le SCIAN. Sept groupes du commerce de gros selon le SCIAN ont un groupe correspondant selon la CTI. Les groupes du SCIAN sans groupe correspondant selon la CTI sont traités à l'aide d'un ratio stocks/ventes global.

4.3 Continuité des séries sous le SCIAN

Les séries sous la classification du SCIAN se divisent en trois parties. Une première de janvier 1991 à décembre 1997 où les estimations sont obtenues à l'aide de coefficients de conversion estimés. La seconde partie commence en janvier 1998 et se terminera lors de l'arrêt de l'enquête actuelle. Dans cette partie, les séries sous le SCIAN sont obtenues par estimation par domaine, autrement dit à l'aide de coefficients de conversion *observés*. La troisième partie débute en même temps que l'enquête remaniée. La deuxième et la troisième parties se chevauchent de quelques mois lorsque l'ancienne et la nouvelle enquête seront toutes les deux en production (test en parallèle). On prévoit la présence d'un bris dans les séries lors du passage à l'enquête remaniée. Ce saut s'explique par le changement de classification mais aussi par d'autres changements méthodologiques. Les résultats du test en parallèle permettront d'ajuster le niveau des séries rétrolées. On prévoit un ajustement multiplicatif constant dans le temps pour ajuster les séries historiques aux niveaux publiés dans la nouvelle enquête.

En tenant compte de la saisonnalité, des bris dans les séries sous le SCIAN étaient également observables en janvier 1998. On passe alors de coefficients estimés en décembre 1997 à des coefficients observés. Pour amoindrir l'effet, toutes les données rétrolées de 1998 ont été recalculées en utilisant les coefficients estimés, c'est-à-dire en calculant la moyenne des quatre années, incluant 1998. Notons que les coefficients de 1998 diffèrent plus souvent de la moyenne que les trois autres années et que les coefficients très aberrants étaient enlevés du calcul de la moyenne. En allongeant la première partie de la série jusqu'en décembre 1998, on annule la présence de bris entre les deux premières parties.

5. ACTIVITÉS À VENIR

Les premières estimations sous le SCIAN de l'enquête remaniée sont prévues en 2004. Les données historiques antérieures seront ajustées pour le niveau à l'aide des résultats du test en parallèle. Grâce aux données réropolées, les nouvelles estimations pourront être désaisonnalisées et analysées.

Le SCIAN a connu une première mise à jour en 2002 qui ne touchait pas les secteurs du commerce de gros et du commerce de détail. Une seconde révision, attendue en 2007, affectera le commerce de gros. Une rétopolation des données selon la nouvelle version du SCIAN est prévue.

REMERCIEMENTS

L'auteure désire remercier François Lavoie, Benoît Quenneville, Julie Trépanier et particulièrement Hélène Bérard pour leurs commentaires et leurs suggestions fort utiles.

RÉFÉRENCES

- Bérard, H.(2001), The Redesign of the Monthly Wholesale and Retail Trade Survey of Statistics Canada, *Recueil 2001 de la section des méthodes d'enquêtes*, Société statistique du Canada, pp 81-86.
- Hidioglou, M., Quenneville, B. et Huot G. (2001), *Methodological Problem and Options for SIC-NAICS Conversion*, Document de travail de Statistique Canada, Division des méthodes d'enquêtes auprès des entreprises, octobre 2001.
- Meyer, B.(2001), *Conversion de la CTI de 1980 au SCIAN; Commerce de gros et de détail; aperçu dans la perspective d'un secteur d'enquête*, Document de travail de Statistique Canada, Division de la statistique du commerce, septembre 2001.
- Statistique Canada (2002), *Système de classification des industries de l'Amérique du Nord : SCIAN, Canada 2002*, produit n° 12-501-XPF au catalogue de Statistique Canada, Ottawa, Ministre de l'industrie, avril 2002.