ENQUETE MENSUELLE SUR LE COMMERCE DE GROS (EMCG) – ÉNONCÉ DE LA QUALITÉ DES DONNÉES

1. OBJECTIFS, UTILISATIONS ET UTILISATEURS

1.1. Objectifs

L'Enquête mensuelle sur le commerce de gros (EMCG) fournit des renseignements sur performance du secteur du commerce de gros et constitue un important indicateur de la santé de l'économie canadienne. Le milieu des affaires utilise aussi les données pour analyser le comportement du marché.

1.2 Utilisation

Les estimations fournissent une mesure de la santé et de la performance du secteur du commerce de gros. L'information recueillie est utilisée pour estimer le niveau et la tendance mensuelle des ventes et des stocks des marchands en gros. À la fin de chaque année, les estimations donnent un premier aperçu de la valeur annuelle des ventes de gros et de la performance du secteur.

1.3 Utilisateurs

Divers organismes, associations sectorielles et gouvernements utilisent l'information.
Les grossistes utilisent les résultats de l'enquête pour comparer leurs résultats à ceux d'entreprises similaires, ainsi qu'à des fins de marketing. Les associations de grossistes peuvent surveiller la performance de leur industrie et promouvoir les industries du commerce de gros. Les investisseurs peuvent surveiller la croissance de l'industrie, ce qui peut donner aux grossistes un meilleur accès au capital d'investissement. Les données de l'enquête aident les administrations à comprendre le rôle des grossistes dans l'économie, ce qui facilite l'élaboration des politiques et des encouragements fiscaux. Le commerce de gros étant une industrie importante dans l'économie canadienne (de 5 % à 6 % du produit intérieur brut, selon l'année), les données permettent aux administrations de déterminer plus exactement la santé globale de l'économie grâce à l'utilisation des estimations dans le calcul du produit intérieur brut (PIB) national.

2. CONCEPTS, VARIABLES ET CLASSIFICATIONS

2.1. Concepts

Le commerce de gros est généralement l'étape intermédiaire dans la distribution des marchandises. Le secteur comprend les établissements dont l'activité principale consiste à acheter et à vendre des marchandises et à fournir des services connexes de logistique, de marketing et de soutien.

Les grossistes sont organisés pour vendre des marchandises en grande quantité à des détaillants, à des entreprises et à une clientèle institutionnelle. Cependant, certains grossistes, notamment ceux qui fournissent des biens d'équipement qui ne sont pas de grande consommation, vendent des marchandises à la pièce aux utilisateurs finals. Le secteur comprend deux grands types de grossistes, soient, d'une part, les marchands en gros et, d'autre part, les agents et les courtiers en gros. Les marchands en gros achètent et vendent des marchandises pour leur propre compte; autrement dit, ils s'approprient les marchandises qu'ils vendent. Ils travaillent habituellement à partir d'entrepôts ou de bureaux et ils peuvent expédier les marchandises qu'ils ont en stock, ou les faire expédier directement du fournisseur au client. En plus de vendre des marchandises, ils peuvent assurer ou faire le nécessaire pour que soient assurés des services de logistiques, de marketing et de soutien, tels que l'emballage et l'étiquetage, la gestion des stocks, l'expédition, le traitement des réclamations au titre de la garantie, la promotion interne ou la promotion coop et la formation requise par le produit. Entrent aussi dans cette catégorie les négociants en machines et en matériel, comme les négociants en machines agricoles et en poids lourds. Les établissements de ce secteur sont

connus sous diverses appellations selon les liens qu'ils entretiennent avec les fournisseurs ou les clients, ou selon la méthode de distribution qu'ils emploient. Ils peuvent se faire appeler, par exemple, grossistes, distributeurs en gros, intermédiaires en gros, concessionnaires de rayon, marchands d'import-export, groupes d'acheteurs, coopératives de marchands et grossistes d'une bannière particulière. Aux fins de la classification des branches d'activité, les marchands en gros sont classés d'après les principales gammes de marchandises vendues. La description de chaque groupe de commerce inclus dans les données statistiques d'accompagnement figure à l'annexe IV. Comme la plupart des entreprises vendent plusieurs types de marchandises, la classification attribuée reflète généralement la marchandise unique ou le groupe de archandises qui est la source principale des recettes de l'établissement, ou un ensemble de marchandises qui caractérise l'activité de l'établissement. Les agents et courtiers en gros achètent et vendent des marchandises pour le compte de tiers moyennant le versement d'honoraire ou d'une commission. Ils ne deviennent pas propriétaires de ces marchandises, et ils travaillent habituellement à partir d'un bureau. Les agents et les courtiers en gros sont connus sous diverses appellations, dont agents d'import-export, agents en gros à la commission, courtiers en gros et agents commerciaux.

2.2. Variables

Les ventes sont définies comme étant les ventes de toutes les marchandises achetées pour la revente, nettes des rendus et des escomptes. Sont incluses les pièces qui sont à l'origine des revenus d'entretien et de réparation, les revenus de main-d'oeuvre pour l'entretien et la réparation, les ventes de biens fabriqués par le grossiste à titre d'activité secondaire, et les recettes provenant de la location ou de la location à bail de locaux à bureaux, d'autres biens immobiliers, et de biens et d'équipement. Sont également incluses les recettes sous forme de commissions et d'honoraires résultant de l'achat et de la vente de marchandises par des marchands en gros pour le compte de tiers.

Sont exclues les autres recettes d'exploitation, comme les subventions d'exploitation, ainsi que les recettes provenant de l'expédition, de la manutention et du stockage de biens pour des tiers.

Les stocks sont définis comme étant la valeur comptable, c'est-à-dire la valeur inscrite dans les livres comptables, de tous les stocks possédés par un établissement à la fin du mois et destinés à la revente. Sont inclus les stocks détenus dans les points de vente, les entrepôts, en transit ou en consignation par des tiers. Sont également inclus les stocks possédés au Canada et à l'étranger.

Sont exclus les stocks détenus en consignation pour des tiers (non possédés), ainsi que les fournitures de magasin et de bureau et tout autre type de fournitures non destinées à la vente. L'emplacement d'affaires comprend le ou les emplacements physiques où a lieu l'activité commerciale dans chaque province et territoire, et dont les ventes sont créditées ou reconnues dans les états financiers de la compagnie. Pour les grossistes, il s'agit habituellement d'un centre de distribution. Le prix courant s'entend du prix en vigueur durant la période de référence. Le prix constant est la valeur exprimée au prix en vigueur durant une période de référence fixe ou période de base.

2.3. Classifications

L'Enquête mensuelle sur le commerce de gros est fondée sur la définition du commerce de gros adoptée dans le SCIAN (Système de classification des industries de l'Amérique du Nord). Le SCIAN est le cadre commun reconnu pour la production de statistiques comparables par les organismes statistiques du Canada, du Mexique et des États-Unis. L'accord définit les limites de 20 secteurs. Le SCIAN est fondé sur un cadre conceptuel axé sur la production, ou l'offre, en ce sens que les établissements sont regroupés par industrie d'après la similarité des processus utilisés pour produire les biens et les services.

Les estimations sont calculées pour 24 groupes fondés sur le Système de classification des industries de l'Amérique du Nord (SCIAN) de 2007. Les 24 groupes sont en outre agrégés en

sept sous-secteurs qui correspondent exactement aux codes à trois chiffres du SCIAN pour les classes du secteur du commerce de gros, à l'exception des agents et courtiers en gros, et des grossistes-distributeurs de pétrole, et de graines oléagineuses et de céréales. Du point de vue géographique, les estimations des ventes sont produites pour le Canada et pour chaque province et territoire. Les estimations des stocks sont produites uniquement pour le Canada dans son ensemble.

3. COUVERTURE ET BASES DE SONDAGE

La base de sondage de l'Enquête mensuelle sur le commerce de gros (EMCG) est le Registre des entreprises (RE) de Statistique Canada. Ce dernier est une liste structurée d'entreprises productrices de biens et de services au Canada. Cette base de données tenue à jour centralement contient des renseignements détaillés sur la plupart des entités commerciales exploitées au Canada. Le RE couvre toutes les entreprises constituées en société, avec ou sans employés. Pour les entreprises non constituées en société, le RE comprend toutes les entreprises ayant des employés, ainsi que les entreprises sans employés ayant des ventes annualisées provenant d'un compte de la taxe sur les produits et services (TPS) ou un revenu annuel provenant de la déclaration d'impôt individuelle.

Dans le RE, les entreprises sont représentées selon une structure hiérarchique à quatre niveaux ayant pour sommet l'entreprise statistique suivie, par ordre décroissant, par la compagnie statistique, l'établissement statistique et l'emplacement statistique. Une entreprise peut être reliée à une ou à plusieurs compagnies statistiques, une compagnie statistique à un ou à plusieurs établissements statistiques et un établissement statistique à un ou à plusieurs emplacements statistiques.

La population cible de l'EMCG comprend tous les établissements statistiques figurant dans le RE, excluant les entreprises non constituées en société n'ayant pas d'employés dont les ventes annuelles sont inférieures à 30 000 $, qui sont classés dans le secteur du commerce de gros d'après le Système de classification des industries de l'Amérique du Nord (SCIAN) (environ 90 000 établissements). La fourchette de codes du SCIAN pour le secteur du commerce de gros varie de 410000 à 419999. Un établissement statistique est l'entité de production ou le plus petit groupe d'entités de production qui produit un ensemble de biens ou de services homogènes, dont les activités ne débordent pas les frontières provinciales/territoriales, et qui est en mesure de fournir des données sur la valeur de la production, ainsi que sur le coût des matières utilisées et le coût et l'importance de la main-d'oeuvre affectée à la production. L'entité de production est l'unité physique où se déroulent les activités de l'entreprise. Elle doit avoir une adresse de voirie et une main-d'oeuvre directement affectée au processus de production.

Sont exclus de la population cible les établissements auxiliaires (producteurs de services de soutien de l'activité de production de biens et services destinés au marché de plus d'un établissement au sein de l'entreprise, et qui sont considérés comme un centre de coûts ou un centre de dépenses discrétionnaires pour lequel les données sur tous les coûts, y compris la main-d'oeuvre et l'amortissement, peuvent être déclarées par l'entreprise), les futurs établissements, les établissements pour lesquels les signaux économiques indiquent un revenu manquant ou nul, et les établissements appartenant aux catégories du SCIAN non couvertes qui suivent :

- 41112 (graines oléagineuses et céréales)
- 412 (produits pétroliers)
- 419 (agents et courtiers du commerce de gros)

4. ÉCHANTILLONNAGE

L'échantillon de l'EMCG est formé de 7 500 groupes d'établissements (grappes) classés dans le secteur du commerce de gros et sélectionnés à partir du Registre des entreprises de Statistique Canada. Par définition, une grappe d'établissements comprend tous les établissements appartenant à une entreprise statistique qui font partie d'un même groupe industriel et d'une

même région géographique. L'EMCG est fondée sur un plan d'échantillonnage stratifié avec sélection d'un échantillon aléatoire simple dans chaque strate. La stratification est faite selon des groupes industriels (majoritairement mais non exclusivement des SCIAN à quatre chiffres) et selon la région géographique, c'est-à-dire selon la province ou le territoire. Ensuite, la population est stratifiée selon la taille de l'établissement. La mesure de taille est créée en combinant des données provenant d'enquêtes indépendantes et trois variables administratives, à savoir le revenu annuel profilé, les ventes assujetties à la TPS exprimées sur une base annuelle et le revenu de la déclaration d'impôt (T1 ou T2).

Les strates de taille comptent une strate à tirage complet (recensement), au moins deux strates à tirage partiel (échantillonnées partiellement) et une strate à tirage nul (non échantillonnée). La strate à tirage nul est destinée à réduire le fardeau de réponse en excluant les entreprises les plus petites de la population observée. Ces entreprises représentent, en principe, au plus 10 % du total des ventes. Au lieu d'envoyer un questionnaire à ces entreprises, on produit les estimations d'après des données administratives.

L'échantillon est réparti de façon optimale afin d'atteindre les coefficients de variation cibles au niveau du Canada dans son ensemble, de la province ou du territoire, de l'industrie et des groupes industriels selon la province ou le territoire. On procède aussi à un suréchantillonnage pour tenir compte des unités disparues, non répondantes ou classées incorrectement.

L'EMCG est une enquête répétée avec maximisation du chevauchement des échantillons mensuels. On retient l'échantillon d'un mois à l'autre et, chaque mois, on y ajoute de nouvelles unités (naissances). Pour découvrir les nouvelles unités visées par l'EMCG, c'est-à-dire les nouvelles grappes d'établissement(s), on examine chaque mois l'univers le plus récent du RE. On stratifie ces nouvelles unités conformément aux mêmes critères que ceux appliqués à la population initiale, puis on les échantillonne conformément à la fraction d'échantillonnage de la strate à laquelle elles appartiennent et on les ajoute à l'échantillon mensuel. Des disparitions d'entité surviennent également chaque mois. Une entité disparue peut être une grappe d'établissements qui ont cessé leurs activités (fermeture) ou dont les activités principales ne se rattachent plus au commerce de gros (hors du champ). La situation de ces entreprises est mise à jour dans le RE d'après des renseignements de source administrative et les commentaires reçus lors des enquêtes, y compris ceux des entreprises prenant part à l'EMCG. Les méthodes suivies pour traiter les unités disparues et les unités classées incorrectement font partie des procédures d'échantillonnage et de mise à jour de la population.

5. CONCEPTION DU QUESTIONNAIRE

Le questionnaire est conçu pour recueillir mensuellement auprès d'un échantillon de grossistes des données sur les ventes en gros, sur le nombre d'emplacements d'affaires par province ou territoire et sur les stocks de biens possédés et destinés à la revente. Lors du remaniement de 2004, à part l'inclusion du Nunavut, la plupart des questionnaires n'ont subit que des changements de présentation. Les modifications ont été discutées avec les intervenants et les répondants ont eu l'occasion de faire des commentaires avant que le nouveau questionnaire ne soit finalisé. Si d'autres modifications devaient être apportées à l'un des questionnaires, les changements proposés seraient soumis à un comité d'examen et ferait l'objet d'un essai sur le terrain auprès de répondants et d'utilisateurs des données pour s'assurer de leur pertinence.

6. RÉPONSE ET NON RÉPONSE

6.1. Réponse et non-réponse

Bien que les gestionnaires d'enquête et les employés des opérations fassent tout leur possible pour maximiser la réponse à l'EMCG, un certain degré de non-réponse a lieu. Pour qu'un établissement statistique soit considéré comme répondant, il faut que le degré de réponse partielle (situation où une réponse exacte n'est obtenue que pour certaines questions posées au répondant) atteigne un seuil minimal au-dessous duquel la déclaration fournie par l'établissement

serait rejetée et l'établissement, considéré comme une unité non répondante. Le cas échéant, on considère que l'entreprise n'a pas répondu du tout.
La non-réponse a deux effets sur les données : premièrement, elle introduit un biais dans les estimations si les non-répondants diffèrent des répondants en ce qui concerne les caractéristiques mesurées et, deuxièmement, elle fait augmenter la variance d'échantillonnage des estimations, parce que la taille effective de l'échantillon est réduite comparativement à celle considérée au départ.

L'ampleur des efforts déployés pour obtenir une réponse auprès d'un non-répondant dépend des contraintes budgétaires et de temps, de l'effet de la non-réponse sur la qualité globale et du risque de biais dû à la non-réponse.

La méthode principalement utilisée pour réduire l'effet de la non-réponse à l'étape de l'échantillonnage consiste à augmenter la taille de l'échantillon en appliquant un taux de suréchantillonnage déterminé d'après les résultats d'enquêtes similaires.

Les cas de non-réponse qui surviennent malgré les méthodes appliquées aux étapes de l'échantillonnage et de la collecte pour réduire l'effet de la non-réponse sont traités par imputation.

Afin de déterminer l'importance de la non-réponse qui a lieu chaque mois, on calcule divers taux de réponse. Pour un mois de référence donné, on produit les estimations au moins deux fois (estimations provisoires et estimations révisées). Entre les deux exécutions, certaines données fournies par les répondants peuvent être jugées inutilisables et des valeurs imputées peuvent être corrigées au moyen de données fournies par les répondants. Par conséquent, les taux de réponse sont calculés après chaque exécution du processus d'estimation.

Pour l'EMCG, deux types de taux sont calculés (non pondérés et pondérés). Afin d'évaluer l'efficacité du processus de collecte, on calcule les taux de réponse non pondérés. Les taux pondérés, fondés sur le poids d'estimation et la valeur de la variable d'intérêt, évaluent la qualité de l'estimation. À l'intérieur de chacun de ces types de taux, il existe des taux distincts pour les unités faisant partie de l'échantillon et pour les unités qui sont uniquement modélisées à partir de données administratives qui ont été extraites des fichiers de TPS.

Afin d'obtenir une meilleure idée du succès du processus de collecte de données, on calcule deux taux non pondérés appelés « taux de résultat de la collecte » et « taux de résultat de l'extraction ». On calcule ces taux en divisant le nombre de répondants par le nombre d'unités avec lesquelles on a essayé de prendre contact ou pour lesquelles on a essayé de recevoir des données extraites. Les déclarants non mensuels (répondants bénéficiant de modalités de déclaration spéciales leur permettant de ne pas produire de déclaration chaque mois, mais pour lesquels des données réelles sont disponibles lors des révisions subséquentes) sont exclus du numérateur ainsi que du dénominateur pour les mois où aucun contact n'est pris avec eux.

Brièvement, les divers taux de réponse se calculent comme suit :

Taux pondérés :

Taux de réponse des unités faisant partie de l'échantillon (estimation) =
Somme des ventes pondérées des unités avec situation de réponse *i*
Somme des ventes pondérées des unités faisant partie de l'échantillon
où *i* = unités pour lesquelles il existe des données déclarées qui seront utilisées dans l'estimation ou qui sont des refus convertis, ou pour lesquelles il existe des données déclarées qui n'ont pas encore été évaluées pour l'estimation.
Taux de réponse des unités modélisées à partir de données administratives (estimation) =
Somme des ventes pondérées des unités avec situation de réponse *ii*
Somme des ventes pondérées des unités modélisées à partir de données administratives
où *ii* = unités pour lesquelles il existe des données extraites des fichiers administratifs et qui sont utilisables pour l'estimation.

Taux de réponse total (estimation) =
Somme des ventes pondérées des unités avec situation de réponse *i* ou situation de réponse *ii*
Somme de toutes les ventes pondérées

Taux non pondérés :

Taux de réponse des unités faisant partie de l'échantillon (collecte) =
Nombre de questionnaires avec situation de réponse *iii*
Nombre de questionnaires avec situation de réponse *iv*
où *iii* = unités pour lesquelles il existe des données déclarées (dont le cas n'est pas résolu, utilisées ou non utilisées pour l'estimation) ou qui sont des refus convertis;
où *iv* = toutes les unités susmentionnées, ainsi que les unités qui ont refusé de répondre, les unités avec lesquelles on n'a pas pris contact et d'autres types d'unités non répondantes.
Taux de réponse des unités modélisées à partir de données administratives (extraction) =
Nombre de questionnaires avec situation de réponse *vi*
Nombre de questionnaires avec situation de réponse *vii*
où *vi* = unités dans le champ d'observation pour lesquelles il existe des données (utilisables ou non utilisables) extraites des fichiers administratifs;
où *vii* = toutes les unités susmentionnées, ainsi que les unités qui ont refusé de déclarer la source de données administratives, les unités avec lesquelles on n'a pas pris contact et d'autres types d'unités non répondantes.
(% *de questionnaires recueillis par rapport à l'ensemble des questionnaires dans le champ d'observation*)
Taux de résultat de la collecte =
Nombre de questionnaires avec situation de réponse *iii*
Nombre de questionnaires avec situation de réponse *viii*
où *iii* = même que *iii* défini plus haut;
où *viii* = même que *iv*, à part l'exclusion des unités avec lesquelles on a pris contact, parce que leur réponse n'est pas disponible pour un mois particulier, puisqu'il s'agit de déclarants non mensuels.
Taux de résultat de l'extraction =
Nombre de questionnaires avec situation de réponse *ix*
Nombre de questionnaires avec situation de réponse *vii*
où *ix* = même que *vi*, avec l'ajout des unités extraites qui ont été imputées ou qui étaient hors du champ de l'enquête;
où v*ii* = même que v*ii* défini plus haut.
(% de questionnaires recueillis par rapport à l'ensemble des questionnaires dans le champ d'observation que nous avons tenté de recueillir)
Tous les taux pondérés et non pondérés susmentionnés sont calculés au niveau du groupe industriel, de la région et du groupe de taille, ainsi que pour toute combinaison de ces niveaux.

Utilisation des données administratives

Réduire le fardeau de réponse est un défi à long terme pour Statistique Canada. Afin d'alléger le fardeau de réponse et de réduire les coûts reliés à l'enquête, notamment en ce qui a trait aux petites entreprises, l'EMCG a réduit le nombre d'établissements simples de l'échantillon qui sont enquêtés directement et dérive plutôt les chiffres de vente pour ces établissements à partir des fichiers de la TPS en utilisant un modèle statistique. Le modèle explique les différences entre les ventes et les recettes déclarées aux fins de la TPS, ainsi que le décalage entre la période de référence de l'enquête et celle de la TPS.

Les stocks des entreprises dont les ventes sont tirées du fichier de la TPS sont imputés par le système d'imputation de l'EMCG. L'imputation se rapporte aux chiffres du mois précédent, aux variations de mois en mois et d'année en année, pour des entreprises enquêtées de même taille.

Pour en savoir plus sur la méthode utilisée lors de la modélisation des ventes tirées de fichiers administratifs, veuillez consulter le document intitulé Enquête mensuelle sur le commerce de gros : Utilisation de données administratives sous la rubrique 'Documentation' du BMDI.

6.2. Méthodes utilisées pour réduire la non-réponse durant la collecte

Beaucoup d'efforts sont déployés en vue de réduire au minimum la non-réponse durant la collecte. Les méthodes utilisées incluent des techniques d'interview, comme l'utilisation de questions d'approfondissement et des techniques de persuasion, la replanification répétée des appels téléphoniques pour obtenir l'information et la mise en place de procédures indiquant aux intervieweurs comment s'y prendre avec les répondants qui refusent de participer à l'enquête.
Si les données demandées ne sont pas disponibles au moment de la collecte, la meilleure estimation fournie par le répondant est acceptée et est révisée par la suite, quand les données réelles sont disponibles.

Pour réduire au minimum la non-réponse totale pour toutes les variables, des réponses partielles sont acceptées. En outre, les questionnaires sont personnalisés pour la collecte de certaines variables, comme les stocks, de sorte que la collecte ait lieu durant les mois où les données sont disponibles.

Enfin, pour établir un climat de confiance entre les intervieweurs et les répondants, les cas sont généralement affectés au même intervieweur chaque mois. Ce dernier peut ainsi établir une relation personnelle avec le répondant et renforcer sa confiance.

7. OPÉRATIONS DE COLLECTE ET DE SAISIE DES DONNÉES

La collecte des données est réalisée par les bureaux régionaux de Statistique Canada.
Ces derniers envoient un questionnaire aux répondants ou communiquent avec ceux-ci par téléphone afin d'obtenir les valeurs de leurs ventes et de leurs stocks, et de confirmer l'ouverture ou la fermeture des emplacements d'affaires. Ils effectuent aussi un suivi auprès des non-répondants. La collecte des données débute environ sept jours ouvrables après la fin du mois de référence et se poursuit pendant tout le mois en question.
Les entités qui participent à l'enquête pour la première fois reçoivent une lettre d'introduction en vue d'informer le répondant qu'un représentant de Statistique Canada l'appellera. Cet appel a pour but de présenter l'enquête, de confirmer l'activité de l'entreprise, d'établir et de commencer la collecte des données, et de répondre à toutes questions que le répondant pourrait avoir.

8. VÉRIFICATION

La vérification des données est l'application de contrôles pour déceler les entrées manquantes, invalides ou incohérentes, ou pour repérer les enregistrements de données susceptibles d'être erronés. Durant le processus d'enquête de l'EMCG, les données sont vérifiées à deux moments distincts.

Premièrement, une vérification est faite durant la collecte des données. Après leur collecte par téléphone ou au moyen du questionnaire à renvoyer par la poste, les données sont saisies à l'aide d'applications informatiques personnalisées. Toutes sont soumises à une vérification. Les contrôles réalisés durant la collecte des données, appelés contrôles sur le terrain, comprennent généralement des contrôles de validité et certains contrôles de cohérence simples. Ils servent aussi à déceler les erreurs commises durant l'interview par le répondant ou par l'intervieweur et de repérer l'information manquante à l'étape de la collecte en vue de réduire le besoin d'un suivi ultérieur. Les contrôles sur le terrain ont également pour but d'épurer les réponses. Dans le cas de l'EMCG, les réponses du mois courant sont comparées aux réponses fournies par le répondant le mois précédent et (ou) l'année précédente pour le mois courant. Les contrôles sur le terrain permettent de repérer les problèmes que posent les procédures de collecte des données et la conception des questionnaires, et de déterminer s'il faut offrir une formation supplémentaire aux intervieweurs.

Tout enregistrement de données rejeté lors des contrôles préliminaires fait l'objet d'un suivi auprès du répondant afin de valider les données soupçonnées d'être incorrectes. Une fois validé, les données recueillies sont transmises de façon régulière au Bureau central à Ottawa.

Deuxièmement, après la collecte, les données sont soumises à une vérification statistique dont la nature est plus empirique. On exécute la vérification statistique avant l'imputation, afin de repérer les données qui serviront de base pour l'imputation de valeurs pour les non-répondants. Les valeurs très extrêmes risquant de perturber une tendance mensuelle sont exclues des calculs de tendance lors de la vérification statistique. Il convient de souligner qu'aucun ajustement n'est fait à cette étape pour corriger les valeurs extrêmes déclarées.

La première étape de la vérification statistique consiste à repérer les réponses qui seront soumises aux règles de vérification statistique. Les données déclarées pour le mois de référence courant sont soumises à divers contrôles.

Le premier ensemble de contrôles est fondé sur la méthode d'Hidiroglou-Berthelot qui consiste à examiner le rapport des données du mois courant fournies par un répondant à des données historiques (c.-à-d. dernier mois ou même mois l'année précédente) ou administratives. Si le rapport calculé pour le répondant diffère significativement de ceux obtenus pour des répondants dont les caractéristiques sont comparables en ce qui concerne le groupe industriel et/ou la région géographique, la réponse est considérée comme une valeur extrême.

Le deuxième ensemble de contrôles est basé sur la vérification de la part de marché.

Cette méthode, qui s'appuie sur les données du mois courant uniquement, permet de vérifier les données fournies par tous les répondants, mêmes ceux pour lesquels on ne dispose pas de données historiques ou de données auxiliaires. Par conséquent, parmi un groupe de répondants présentant des caractéristiques similaires en ce qui concerne le groupe industriel et (ou) la région géographique, toute valeur dont la contribution pondérée au total du groupe est trop importante sera considérée comme une valeur extrême.

Pour les contrôles fondés sur la méthode d'Hidiroglou-Berthelot, les données jugées extrêmes ne sont pas incluses dans les modèles d'imputation (ceux fondés sur les ratios). En outre, les données considérées comme des valeurs extrêmes lors de la vérification de la part de marché ne sont pas incluses dans les modèles d'imputation où les moyennes et les médianes sont calculées pour imputer des valeurs pour les réponses pour lesquelles il n'existe pas de données historiques.

Conjointement avec les vérifications statistiques effectuées après la collecte de données, on procède à la détection d'erreurs des données extraites des fichiers administratifs. Les données modélisées de la TPS sont également assujetties à une phase de vérification approfondie. Chaque fichier sur lequel les données modélisées sont fondées est vérifié de même que les valeurs modélisées. Les vérifications sont effectuées au niveau agrégé (industrie, géographie) afin de détecter les fichiers qui dévient de la norme (soit en exhibant des différences d'un mois à l'autre trop importantes ou qui diffèrent considérablement des autres unités. Toutes les données qui faillissent ces étapes de contrôle sont sujettes à une vérification manuelle, et si nécessaire, à une action corrective.

9. IMPUTATION

Le processus d'imputation de l'EMCG a pour but de remplacer les données manquantes par des valeurs imputées. Des valeurs sont attribuées aux enregistrements pour lesquels la vérification a révélé des valeurs manquantes afin de s'assurer que les estimations soient de haute qualité et d'établir une cohérence interne plausible. Pour des raisons de fardeau de réponse, de coût et d'actualité des données, il est généralement impossible de réaliser auprès des répondants tous les suivis nécessaires pour résoudre les problèmes de réponses manquantes. Puisqu'il est souhaitable de produire un fichier de microdonnées complet et cohérent, on recourt à l'imputation pour traiter les cas persistants de données manquantes.

Dans le cas de l'EMCG, on peut fonder l'imputation des valeurs manquantes sur des données historiques ou sur des données administratives. Le choix de la méthode appropriée est fondé sur

une stratégie qui dépend de l'existence de données historiques ou de données administratives et (ou) du mois de référence en question.

Il existe trois types de méthode d'imputation d'après des données historiques. Le premier est l'application d'une tendance générale qui s'appuie sur une source unique de données historiques (mois précédent, données recueillies pour le mois suivant ou données recueillies pour le même mois l'année précédente). Le deuxième est un modèle de régression dans lequel sont utilisées simultanément les données provenant du mois précédent et celles provenant du même mois l'année précédente. La troisième méthode consiste à remplacer directement les valeurs manquantes par des données historiques.

Selon le mois de référence, il existe, pour le choix de la méthode, un ordre de préférence en vue d'assurer une imputation de haute qualité. Le troisième type de méthode d'imputation historique est toujours la dernière option considérée pour chaque mois de référence.

Les méthodes d'imputation fondées sur des données administratives sont sélectionnées automatiquement lorsqu'on ne dispose pas de données historiques pour un non-répondant. La source de données administratives (ventes annuelles assujetties à la TPS) est le fondement de ces méthodes. Les ventes annuelles assujetties à la TPS sont utilisées pour deux types de méthode. L'une est une tendance générale que l'on utilise pour les structures simples, comme les entreprises ne comptant qu'un seul établissement et l'autre, appelée méthode de la médiane-moyenne, est utilisée pour les unités dont la structure est plus complexe.

En dernier lieu, on doit noter que les stocks des entreprises dont les ventes sont tirées des fichiers administratifs, sont également imputés par le système d'imputation de l'EMCG. Les valeurs imputées sont calculées par le même système d'imputation qui existe pour remplacer les données manquantes dues à la non-réponse.

10. ESTIMATION

L'estimation est un processus qui consiste à calculer une valeur approximative des paramètres de population inconnus en utilisant uniquement la partie de la population qui est incluse dans un échantillon. Des inférences sont ensuite faites au sujet des paramètres inconnus en utilisant les données d'échantillon et les renseignements connexes sur le plan de sondage. Cette étape fait usage du Système généralisé d'estimation (SGE) de Statistique Canada.

Pour les ventes des marchands en gros, la population est divisée en une partie observée (strates à tirage complet et à tirage partiel) et une partie non observée (strate à tirage nul). D'après l'échantillon tiré à partir de la partie observée, on calcule une estimation pour la population au moyen d'un estimateur d'Horvitz-Thompson où les réponses concernant les ventes sont pondérées par l'inverse des probabilités d'inclusion des unités échantillonnées. Ces poids (appelés poids d'échantillonnage) peuvent être interprétés comme étant le nombre de fois que chaque unité échantillonnée devrait être répétée pour représenter la population complète. Les valeurs pondérées des ventes ainsi calculées sont totalisées par domaine, pour produire une estimation du total des ventes pour chaque combinaison des groupes industriels/région géographique. Un domaine est défini comme correspondant aux valeurs de classification les plus récentes disponibles dans le RE pour l'unité et la période de référence de l'enquête. Les domaines peuvent différer des strates d'échantillonnage originales, parce que les unités peuvent avoir changé de taille, de groupe de commerce ou d'emplacement. Les changements de classification sont reflétés immédiatement dans les estimations et ne sont pas cumulés au cours du temps. Pour la partie non observée de la population, les ventes sont estimées à l'aide de modèles statistiques exploitant les ventes assujetties à la TPS exprimées sous forme mensuelle.

Pour les stocks des marchands en gros, on se sert de l'échantillon sélectionné pour estimer les ventes pour calculer l'estimation au moyen d'un estimateur d'Horvitz-Thompson pour la partie observée de la population. Puis, on utilise un ratio fondé sur l'échantillon pour produire

l'estimation pour la partie non observée et on obtient l'estimation du total des stocks en additionnant les estimations obtenues pour les parties observée et non observée.

Pour en savoir plus sur la méthode utilisée lors de la modélisation des ventes tirées de fichiers administratifs, veuillez consulter le document intitulé Enquête mensuelle sur le commerce de gros : Utilisation de données administratives sous la rubrique ‘Documentation’ du BMDI.

La variance est la mesure de précision utilisée dans le cas de l'EMCG pour évaluer la qualité de l'estimation des paramètres de population et pour obtenir des inférences valides. Pour la partie observée de la population, la variance est calculée directement à partir d'un échantillon aléatoire simple stratifié sans remise.

Les estimations d'échantillon peuvent différer de la valeur prévue des estimations.

Cependant, puisque l'estimation est fondée sur un échantillon probabiliste, il est possible d'évaluer la variabilité de l'estimation d'échantillon par rapport à la valeur prévue. La variance d'une estimation est une mesure de la précision de l'estimation d'échantillon qui est définie comme étant la moyenne, sur tous les échantillons possibles, de l'écart quadratique de l'estimation par rapport à sa valeur prévue.

11. RÉVISIONS ET DÉSAISONNALISATION

Des révisions des données brutes doivent être effectuées pour corriger les erreurs non dues à l'échantillonnage qui sont décelées. Ceci comporte généralement le remplacement de données imputées par des données déclarées, la correction de données déclarées précédemment, et de procéder à des estimations pour les nouvelles entreprises créées dont on ne connaissait pas l'existence au moment des estimations originales.

Les données brutes sont révisées, sur une base mensuelle, pour le mois précédant immédiatement le mois de référence en cours qui fait l'objet de la publication. C'est donc dire que lorsque les données pour décembre sont publiées pour la première fois, on procédera aussi à des révisions, au besoin, à l'égard des données brutes pour novembre. En outre, des révisions sont aussi effectuées une fois par année, au moment de la première publication des données de février, pour tous les mois de l'année précédente. On vise ainsi à corriger tout problème important que l'on ait décelé et qui s'applique pour une période prolongée. La période de révision proprement dite dépend de la nature du problème décelé, mais elle ne dépasse rarement trois ans.

Les séries temporelles ou chronologiques comportent les éléments essentiels à la description, l'explication et la prévision du comportement d'un phénomène économique. « Ce sont des dossiers statistiques de l'évolution des processus économiques dans le temps1 ». Les séries temporelles socio-économiques comme celles de l’Enquête mensuelle sur le commerce de gros peuvent habituellement être décomposées en cinq composantes principales : la tendance-cycle, la saisonnalité, l’effet des jours ouvrables, l’effet de la fête de Pâques et la composante irrégulière.

La tendance représente l’évolution à long terme de la série, tandis que le cycle représente un mouvement lisse, quasi périodique, autour de la tendance qui met en évidence une succession de phases de croissance et de décroissance (ex. le cycle des affaires). Les deux composantes tendance et cycle sont estimées ensemble et la tendance-cycle reflète l'évolution fondamentale de la série. Les autres composantes traduisent des mouvements passagers à court terme.
La composante saisonnière représente des fluctuations infra-annuelles, mensuelles ou trimestrielles, qui se répètent plus ou moins régulièrement d'une année à l'autre. Les variations saisonnières sont le produit des effets directs et indirects des saisons climatiques et d’éléments de type institutionnel (attribuable aux conventions sociales ou aux règles administratives, Noël par exemple).

L’effet des jours ouvrables provient du fait que l'importance relative des jours varie systématiquement à l'intérieur de la semaine et que le nombre de chacun des jours dans un mois donné varie d'une année à l'autre. Cet effet est présent lorsque l’activité change en fonction du jour de la semaine. Par exemple, dimanche connaît typiquement moins d'activité que les autres jours, et le nombre de dimanches, lundis, etc., dans un mois donné change d'année en année.[1]

[1] La désaisonnalisation des séries temporelles économiques : quelques remarques; tiré de la Revue statistique du Canada , août 1974

[2] Pour plus de renseignements, voir X-12-ARIMA Reference Manual Version 0.3 (2007), U.S. Census Bureau.

[3] Ladiray, D. and Quenneville, B. (2001). Seasonal Adjustment with the X-11 Method. New York: Springer-Verlag, Lecture Notes in Statistics #158.

L'effet de la fête de Pâques est la variation due au déplacement d'une partie de l'activité d'avril vers mars quand Pâques tombe en mars plutôt qu'en avril.

Enfin, la composante irrégulière regroupe toutes les autres fluctuations plus ou moins erratiques non prises en compte dans les composantes précédentes. Elle représente un résidu qui incorpore, entre autres, les erreurs de mesure sur la variable elle-même ainsi que des événements inhabituels (ex. grèves, sécheresse, inondations, panne d'électricité majeure ou d'autres variations inattendues dans les activités des répondants).

Ainsi, les composantes saisonnière et irrégulière, l'effet des jours ouvrables et l'effet de la fête de Pâques masquent la composante fondamentale de la série, qui est la tendance-cycle. La désaisonnalisation (correction des variations saisonnières) consiste à retirer de la série la composante saisonnière, l'effet des jours ouvrables et l'effet de la fête de Pâques. Elle contribue donc à révéler la tendance-cycle. Bien que la désaisonnalisation permette de mieux comprendre la tendance-cycle fondamentale d'une série, la série désaisonnalisée n'en contient pas moins une composante irrégulière. De légères variations d'un mois à l'autre dans la série désaisonnalisée peuvent n'être que de simples mouvements irréguliers. Pour avoir une meilleure idée de la tendance fondamentale, les utilisateurs doivent donc examiner les séries désaisonnalisées sur un certain nombre de mois.

Depuis avril 2008, l'Enquête mensuelle sur le commerce de gros utilise le logiciel X-12-ARIMA2 pour la désaisonnalisation. La technique utilisée consiste essentiellement, dans un premier temps, à corriger la série initiale de toute sorte d'effets indésirables, tels l'effet des jours ouvrables et l'effet de Pâques, par un module appelé regARIMA. L'estimation de ces effets se fait grâce à l'utilisation de modèles de régression à erreurs ARIMA (modèles autorégressifs à moyennes mobiles intégrées). On peut également extrapoler la série d'au moins une année à l'aide du modèle. Dans un deuxième temps, la série brute, pré-ajustée et extrapolée s'il y a lieu, est désaisonnalisée par la méthode X-11.

La méthode X-11, qui permet d'analyser des séries mensuelles et trimestrielles, repose sur un principe itératif d'estimation des différentes composantes, cette estimation étant faite à chaque étape grâce à des moyennes mobiles adéquates3. Les moyennes mobiles utilisées pour estimer les principales composantes, la tendance et la saisonnalité, sont avant tout des outils de lissage conçus pour éliminer une composante indésirable de la série. Puisque les moyennes mobiles réagissent mal à la présence de valeurs atypiques, la méthode X-11 incorpore un outil de détection et de correction des points atypiques utilisé pour nettoyer la série au cours de la désaisonnalisation. Les valeurs atypiques peuvent également être détectées et corrigées d'avance, à l'aide du module regARIMA.

Finalement, les données désaisonnalisées sont ajustées aux totaux annuels des données brutes. Malheureusement, la désaisonnalisation supprime l'additivité infra-annuelle d'un système de séries; de légères différences peuvent alors être observées entre la somme de séries désaisonnalisées et la désaisonnalisation *directe* de leur total. Afin d'assurer ou de rétablir l'additivité d'un système de séries, un processus de réconciliation est appliqué ou une désaisonnalisation *indirecte* est employée, c.-à-d. la désaisonnalisation d'un total est obtenu en faisant la somme des séries désaisonnalisées individuellement.

12. ÉVALUATION DE LA QUALITÉ DES DONNÉES

La méthodologie de l'enquête a pour objectif de contrôler les erreurs et de réduire leurs effets éventuels sur les estimations. Les résultats de l'enquête peuvent néanmoins contenir des erreurs dont l'erreur d'échantillonnage n'est que l'une des composantes.

L'erreur d'échantillonnage survient lorsque les observations sont faites uniquement sur un échantillon et non sur l'ensemble de la population. Toutes les autres erreurs commises aux diverses phases de l'enquête sont appelées erreurs non dues à l'échantillonnage. Des erreurs de ce type peuvent survenir, par exemple, quand un répondant fournit des renseignements erronés ou qu'il ne répond pas à certaines questions; quand une unité du champ de l'enquête y est incluse erronément ou que des erreurs sont commises lors du traitement des données, comme des erreurs de codage ou de saisie.

Avant la publication, on analyse les résultats combinés de l'enquête afin d'en évaluer la comparabilité; il s'agit généralement d'un examen détaillé des réponses individuelles (particulièrement celles des grandes entreprises), de la conjoncture économique générale et des tendances historiques.

Une mesure habituelle de la qualité des données des enquêtes est le coefficient de variation (CV). Le coefficient de variation, défini comme étant l'erreur-type divisée par l'estimation d'échantillon, est une mesure de la précision relative. Puisque le coefficient de variation est calculé d'après les réponses des unités individuelles, il mesure aussi certaines erreurs non dues à l'échantillonnage.

La formule utilisée pour calculer le coefficient de variation (CV) en pourcentage est :

$$CV(X) = \frac{S(X)}{X} * 100\%$$

où X représente l'estimation et S(X) représente l'erreur-type de X.

On peut construire les intervalles de confiance autour des estimations en utilisant l'estimation et le CV. Donc, pour notre échantillon, il est possible de déclarer avec un niveau donné de confiance que la valeur prévue sera comprise dans l'intervalle de confiance construit autour de l'estimation. Par exemple, si une estimation de 12 millions de dollars à un CV de 2 %, l'erreur-type sera de 240 000 $ (l'estimation multipliée par le CV). On peut déclarer avec 68 % de confiance que les valeurs prévues seront comprises dans l'intervalle dont la longueur est égale à un écart-type de part et d'autre de l'estimation, c'est-à-dire entre 11 760 000 $ et 12 240 000 $. Ou bien, nous pouvons déclarer avec 95 % de confiance que la valeur prévue sera comprise dans l'intervalle dont la longueur est égale à deux écart-type de part et d'autre de l'estimation, c'est-à-dire entre 11 520 000 $ et 12 480 000 $.

Enfin, étant donné la faible contribution de la partie non observée de la population aux estimations totales, le biais dans la partie non observée a un effet négligeable sur les CV. Par conséquent, on utilise le CV provenant de la partie observée pour l'estimation totale qui est égale à la somme des estimations pour les parties observée et non observée de la population.

13. CONTRÔLE DE LA DIVULGATION

La loi interdit à Statistique Canada de rendre publique toute donnée susceptible de révéler l'information recueillie en vertu de la Loi sur la statistique et se rapportant à toute personne, entreprise ou organisation reconnaissable, sans que cette personne, entreprise ou organisation le sache ou y consente par écrit. Diverses règles de confidentialité s'appliquent à toutes les données diffusées ou publiées afin d'empêcher la publication ou la divulgation de toute information jugée confidentielle. Au besoin, des données sont supprimées pour empêcher la divulgation directe ou par recoupement de données reconnaissables.

L'analyse de la confidentialité des données inclut la détection de la « divulgation directe » éventuelle, qui survient lorsque la valeur figurant dans une cellule d'un tableau ne correspond qu'à quelques répondants ou que la cellule est dominée par un petit nombre d'entreprises.