

Interprétation des estimations de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) remaniée

Par Steven Thomas, méthodologiste principal, ESCC
et Sylvain Tremblay, analyste principal, ESCC



Résumé

Afin de pouvoir répondre aux besoins des utilisateurs et de mieux utiliser les ressources consacrées aux interviews, la composante régionale de l'Enquête sur la santé dans les collectivités canadiennes, ou le cycle .1 de l'enquête, a été remaniée, en vue d'inclure divers éléments de contenu et de recueillir des données sur une base continue au fil du temps. Ce changement de structure permet la collecte et la diffusion de divers types de données pour diverses périodes, en vue de l'estimation à divers niveaux géographiques et sociodémographiques. Pour l'utilisateur, cela signifie que plusieurs produits différents seront disponibles pour plusieurs périodes différentes. L'interprétation appropriée des résultats revêt plus d'importance que jamais auparavant, l'utilisateur ayant le choix du produit qu'il utilisera pour l'analyse. Le choix du produit sera fondé sur les caractéristiques à étudier et sur le niveau de détail requis dans les estimations. Le présent document vise à préciser les répercussions que le remaniement aura sur les utilisateurs et à contribuer à l'interprétation appropriée des estimations en résultant.

1. Remaniement de l'ESCC

Après la diffusion des données de la composante régionale de 2005 de l'Enquête sur la santé dans les collectivités canadiennes (ESCC cycle 3.1), l'ESCC a été remaniée en vue de donner suite à deux points principaux : mieux répondre aux besoins des utilisateurs et mieux utiliser les ressources affectées à la collecte¹. La mise en œuvre d'une technique de collecte continue a constitué une étape clé de cette démarche. Parallèlement, on a adopté une structure de contenu souple, afin de permettre la collecte de données différentes pour diverses périodes. Ces changements ont des répercussions sur la stratégie de diffusion, du point de vue des types de contenu qui peuvent être diffusés, ainsi que de la fréquence des diffusions. Une fois ces changements en place, il a été décidé qu'il était temps de mettre en œuvre certaines améliorations méthodologiques, y compris un processus plus efficace sur le plan du temps.

¹ Béland Y., Dale V., Dufour J., Hamel M. The Canadian Community Health Survey: Building on the Success from the Past. 2005 Proceedings of the American Statistical Association Meeting, Survey Research Methods. American Statistical Association, 2005.

1.1 Modifications de la collecte

La modification qui a eu les répercussions les plus importantes sur les utilisateurs est la modification de l'approche de collecte des données de l'ESCC. Par le passé, la composante régionale de l'ESCC servait à recueillir des données auprès d'environ 130 000 répondants, sur une période de 12 mois, tous les 2 ans. Depuis janvier 2007, des données sont recueillies sur une base continue auprès d'environ 65 000 répondants, tout au long de l'année, chaque année. Afin d'assurer la continuité de la collecte, des données sont recueillies auprès d'un nouvel échantillon d'environ 11 000 répondants, tous les deux mois, chacun étant représentatif de la région sociosanitaire pour la période donnée. Les échantillons recueillis pour les territoires sont représentatifs de la population après 12 mois.

1.2 Modifications du contenu

Du fait de la nouvelle approche de collecte continue, il est maintenant possible de recueillir divers types de données (ou contenu) pour diverses périodes. La durée de la collecte dépend des caractéristiques voulues et de la taille de l'échantillon requise. Dans le cas des caractéristiques prévalentes et des domaines généraux, les données doivent être recueillies uniquement pendant une courte période pour que l'on dispose d'un nombre suffisant de répondants pour produire des estimations de qualité. Dans le cas des caractéristiques moins prévalentes et des domaines plus détaillés, les données sont recueillies sur une période prolongée, en vue d'obtenir un échantillon approprié de répondants.

Les principales composantes du contenu de l'ESCC se répartissent toujours entre le contenu commun et le contenu optionnel, même si le contenu commun est maintenant fractionné en deux sous-composantes : de base et thématique. Même si les deux sous-composantes sont utilisées pour tous les répondants de l'ESCC, le contenu de base est destiné à demeurer relativement stable au fil du temps, et le contenu thématique fait l'objet d'une collecte pendant 12 à 24 mois et peut être repris dans l'enquête, après 2, 4 ou 6 ans. La composante de contenu optionnel offre aux régions sociosanitaires de choisir du contenu qui répond aux priorités provinciales et régionales en matière de santé publique. Il est recueilli pendant un ou deux ans avant d'être révisé à nouveau.

Une nouvelle composante de réponse rapide est aussi disponible et permet la collecte de données sur les enjeux émergents en matière de santé, auprès d'un petit échantillon de répondants, sur une période de deux mois (environ 11 000 répondants). Cette composante d'une durée maximale de 2 minutes est offerte aux clients des projets à frais recouvrables qui ont un besoin immédiat de données au niveau national.

1.3 Modifications de la diffusion

Les modifications de la collecte et de la structure du contenu de l'ESCC ont des répercussions sur la stratégie de diffusion. Par le passé, les données étaient diffusées tous les deux ans, après la collecte auprès de tous les répondants de l'enquête. Les fichiers de données (principal, de partage, FMGD) sont disponibles pour les années de référence 2000-2001 (cycle 1.1), 2003 (cycle 2.1) et 2005 (cycle 3.1). Un fichier de six mois (permettant le calcul d'estimations pour 65 000 répondants) a été produit à partir des données du cycle 3.1 recueillies de janvier 2005 à juin 2005.

À partir de juin 2008, par suite de la diffusion des données recueillies au cours de la période de collecte de 2007, les fichiers principal et de partage seront diffusés tous les ans. Ces fichiers de données annuels comprendront environ 65 000 répondants, soit la moitié de l'échantillon disponible dans les fichiers de données précédents de l'ESCC. Ils engloberont le contenu de base, le contenu thématique et le contenu optionnel recueillis tout au long de l'année.

4.4 Contenu des fichiers de microdonnées

En juin 2009, deux fichiers principaux seront disponibles : un fichier principal de données fondé sur la période de collecte de 2008, qui s'apparentera au fichier de données de 2007, ainsi qu'un fichier de données fondé sur la période de collecte de 2007-2008. Le fichier principal de 2007-2008 sera de la même taille que les fichiers des cycles précédents (environ 130 000 répondants). Il comprendra le contenu de base, le contenu optionnel et le contenu thématique recueillis au cours de la période de deux ans. Les thèmes touchant une année ne seront pas disponibles dans le fichier de données de deux ans. Par ailleurs, les modules thématiques collectés auprès d'un sous-échantillon de la population continueront d'être diffusés dans des fichiers séparés. Ces derniers incluent le contenu de base et le contenu des modules thématiques collectés auprès d'un sous-échantillon de répondants. Le tableau 4.2 fournit des précisions concernant ce qui sera disponible avec les diffusions de 2007 et 2008.

Outre les fichiers courants, des fichiers de la composante de réponse rapide seront produits pour les clients des projets à frais recouvrables. Ces fichiers seront mis à la disposition d'autres utilisateurs sur demande et comprendront le contenu de la composante de réponse rapide ainsi que le contenu de base pour une période de deux mois.

Les fichiers de microdonnées à grande diffusion (FMGD) seront diffusés tous les deux ans, sur la base de deux années de collecte. Le premier FMGD sera diffusé à l'été 2009, pour la période de collecte de 2007-2008. On ne disposera pas de FMGD pour des années uniques.

Tableau 1. Composantes de contenu comprises dans les fichiers de données de 2007 et 2008

Fichiers		Contenu de base	Contenu thématique de 2007 ¹	Contenu thématique de 2008 ²	Contenu thématique de 2007-2008	Contenu optionnel ³
2007	Principal	Oui	Non	S/O	Oui	Oui
	Sous-échantillon	Oui	Oui	S/O	Non	Non
2008	Principal	Oui	S/O	Oui ⁴	Oui	Oui
	Sous-échantillon	Oui	S/O	Oui ⁵	Non	Non
2007-2008	Principal	Oui	Non	Non	Oui	Oui

¹ Le thème de 2007 était composé de 3 modules (Satisfaction des patients, Accès aux services de santé et Temps d'attentes) qui ont été posés à un sous-échantillon de répondants.

² Le thème de 2008 comprend un groupe de modules reliés à la détection des maladies chroniques ainsi qu'un module sur le poids et la taille mesurés. Ce dernier est demandé à un sous-échantillon de répondants.

³ Selon le principe que le contenu optionnel demeurera le même pendant deux ans. Autrement, il sera inclus uniquement dans le fichier pour l'année pendant laquelle il a été recueilli.

⁴ Détection des maladies chroniques

⁵ Poids et taille mesurées

1.4 Modifications des méthodes d'enquête

Par suite des modifications de la collecte, du contenu et des stratégies de diffusion, certains changements ont été apportés aux méthodes utilisées pour le calcul des poids d'enquête. Le remaniement a signifié que des poids devaient être produits plus fréquemment et qu'une méthodologie conforme à la collecte continue était requise. Cette évolution a aussi été perçue comme une occasion d'apporter certaines améliorations aux ajustements des poids qui sont utilisés dans le processus².

1.4.1 Pondération selon la période

Les poids sont contrôlés, dans la plus large mesure possible, afin que chaque période de collecte soit représentée également et que les répondants visés par la pondération représentent la population moyenne pour la période prolongée de la diffusion particulière. Les estimations représentent la moyenne pour la période.

² Sarafin C., Simard M., Thomas S. (2007). A Review of the Weighting Strategy for the Canadian Community Health Survey. *Recueil du Groupe des méthodes d'enquête, Congrès annuel de 2007 de la Société statistique du Canada.*

1.4.2 Modifications de l'intégration

L'ESCC utilise une méthode à base double, des répondants étant échantillonnés à partir d'une liste téléphonique et d'une base aréolaire. Les poids sont corrigés/intégrés, afin de veiller à ce que la population soit représentée une fois seulement. Par le passé, les poids de la liste téléphonique ont été ajustés pour tenir compte du sous-dénombrement (sans téléphone à fil, numéros de téléphone non publiés, etc.), avant l'intégration dans la base aréolaire, afin que la base aréolaire et la liste téléphonique couvrent la même population. Cela reposait sur l'hypothèse que les personnes non comprises dans la base téléphonique étaient les mêmes que celles qui y figuraient.

Étant donné que l'on sait que les caractéristiques des répondants par téléphone peuvent différer de celles des répondants non compris dans la liste téléphonique, la méthode d'intégration a été mise à jour³. À l'heure actuelle, les répondants de la liste téléphonique sont intégrés uniquement aux unités de la base aréolaire qui figurent aussi dans la liste téléphonique. Les poids des répondants de la base aréolaire qui ne sont pas dans la liste téléphonique ne sont pas ajustés. Cela signifie que pour les variables touchées par le mode de collecte, les estimations produites devraient être plus représentatives de la population réelle.

1.4.3 Modifications de l'étalonnage

La dernière étape de la méthode de pondération vise à faire en sorte que les poids correspondent aux totaux de population connus, grâce à un processus appelé étalonnage. Les totaux connus se situent habituellement au niveau de la région sociosanitaire, selon le groupe d'âge et le sexe. Il est généralement reconnu qu'en étalonnant les poids, les estimations des totaux sont plus précises que celles qui ne sont pas étalonnées. Toutefois, afin de procéder à un ajustement d'étalonnage approprié, il est suggéré de recueillir au moins 20 observations pour le domaine. Cela ne devrait pas poser de problème dans le cas du fichier de deux ans, mais pour le fichier d'un an, il ne sera pas possible d'assurer la stratification a posteriori dans tous les domaines, en raison du nombre réduit de répondants. Les utilisateurs obtiendront une liste des données stratifiées a posteriori comportant moins de 20 observations, et les cellules correspondantes seront supprimées des tableaux produits par Statistique Canada.

2. Répercussions sur les utilisateurs

2.1 Plus de données, plus souvent

À partir des diffusions des données de 2008 et de 2007-2008, en juin 2009, les utilisateurs auront le choix d'utiliser les fichiers d'un an ou de deux ans. Les utilisateurs pourront combiner ces fichiers normalisés pour produire, par exemple, des fichiers de trois ans ou de quatre ans.

³ Skinner, C.J. et Rao, J.N.K. (1996). « Estimation in Dual Frame Surveys with Complex Designs ». *Journal of the American Statistical Association*, 91, 349-356.

2.2 Estimation de la période

Peu importe si un fichier pluriannuel, un fichier de deux ans ou un fichier d'un an est utilisé, on incite les utilisateurs à considérer les données de l'ESCC comme comportant une *estimation de période*, les interviews correspondant à une période étant combinées et un poids d'échantillonnage mis à jour étant calculé. Une estimation annuelle de caractéristiques données rend compte des caractéristiques moyennes de la population moyenne pour la période. Dans le cas du fichier de 2007, les estimations rendent compte de la moyenne de janvier à décembre 2007. Le résultat correspond à une *estimation de période*, qui diffère de l'instantané qui est souvent présenté dans la plupart des enquêtes transversales. Techniquement, cela est vrai uniquement dans le cas du recensement, dont les estimations représentent un point précis dans le temps.

L'idée de l'estimation de période constitue simplement un prolongement des méthodes utilisées pour les cycles précédents de l'ESCC, c'est-à-dire la combinaison d'un ensemble d'interviews menées sur une période de 12 mois. De même, les techniques utilisées pour les ensembles combinés de données uniformisées pour un an ou pour deux ans, en vue de créer des estimations de période personnalisées, seront très similaires à celles utilisées pour la combinaison des cycles 1.1, 2.1 et 3.1 de l'enquête⁴.

Les décisions concernant la période à utiliser pour une analyse donnée devraient être guidées par le niveau de détail et de qualité requis. Dans le cas d'un fichier d'un an, les estimations ne seront pas toujours disponibles en raison de la qualité liée à la taille limitée de l'échantillon. Dans le cadre de l'ESCC, il est recommandé d'utiliser un coefficient de variation de moins de 33 % et de compter au moins dix répondants qui affichent les caractéristiques dans le domaine, avant de publier une estimation. Cela ne sera pas possible dans le cas des caractéristiques rares et des domaines détaillés pour les fichiers d'un an. Les utilisateurs devront plutôt se fier aux fichiers de deux ans ou aux fichiers pluriannuels.

Lorsque l'utilisation d'un fichier d'un an ou de deux ans est possible, l'utilisateur devrait envisager un compromis entre l'exactitude et l'actualité. S'il est important de rendre compte des caractéristiques courantes d'une population le plus précisément possible, le fichier d'un an sera préférable. Dans le cas des fichiers de deux ans, les tendances d'une année à l'autre seront masquées, tout comme les tendances saisonnières sont masquées dans un fichier d'un an. Toutefois, du fait de l'augmentation de la taille de l'échantillon, des estimations et des analyses plus détaillées peuvent être effectuées.

2.3 Répercussions sur la convention de désignation des variables

La convention de désignation des variables a été modifiée légèrement, afin de rendre compte du fait que la même variable est recueillie chaque année. Par le passé, une lettre

⁴ Thomas S. La combinaison de cycles de l'Enquête sur la santé dans les collectivités canadiennes. *Recueil du Symposium de Statistique Canada* (Statistique Canada, n° 11-522-XIF au catalogue), 2006.

désignant les cycles était incluse dans le nom de la variable. Par exemple, le « e » dans « ccece_101 » signifiait qu'il s'agissait des données recueillies dans le cadre du cycle 3.1. À partir de maintenant, la variable sera étiquetée « ccc_101 ». Afin d'aider les utilisateurs qui souhaitent combiner deux fichiers de données ou plus, une nouvelle variable montrant la période de référence « REFPER » a été ajoutée. Cette variable utilise le format AAAAMM-AAAAMM (année et mois du début de la collecte-année et mois de la fin de la collecte).

2.4 Différences dans les estimations par rapport au passé

Les utilisateurs devraient être conscients que les modifications de l'échantillonnage et la production des poids d'échantillonnage adoptés en 2007 expliquent en partie les différences par rapport aux cycles précédents. Du point de vue de l'échantillonnage, l'échantillon est contrôlé, afin de compter à peu près le même nombre de répondants tout au long de l'année et que chaque moitié de l'échantillon provienne de chacune des deux bases de sondage. Il ne s'agit pas d'un changement marqué par rapport aux diffusions précédentes, dans lesquelles l'échantillon était divisé entre les périodes de collecte mensuelles. Du point de vue de la production des poids, les changements apportés au processus d'intégration des listes téléphoniques et des bases aréolaires pourraient avoir pour effet d'influencer les caractéristiques qui comportent une corrélation étroite avec le fait d'avoir un numéro de téléphone publié⁵. D'autres études de cette possibilité sont prévues.

⁵ St-Pierre M, Béland Y. Mode effects in the Canadian Community Health Survey: a comparison of CAPI and CATI. *2004 Proceedings of the American Statistical Association Meeting, Survey Research Methods*. Toronto : American Statistical Association, 2004.

Faits saillants

- À partir du 18 juin 2008, les fichiers de données principaux et de partages seront diffusés chaque année. Ces fichiers annuels comprendront des données sur environ 65 000 répondants, soit la moitié de la taille de l'échantillon des fichiers de données précédents. Des fichiers fondés sur deux années de données continueront d'être produits et seront de la même taille que les fichiers des cycles précédents (~130 000 répondants).
- Dans le cadre du remaniement de l'ESCC, on a adopté un contenu thématique. Les questions s'y rapportant sont posées à tous les répondants de l'ESCC, et les données sont recueillies pour une ou deux années seulement.
- Les fichiers d'échantillons annuels comprendront le contenu de base, le contenu thématique annuel et le thème de deux années, ainsi que le contenu optionnel recueilli cette année-là. Les fichiers de deux ans comprendront le contenu de base, le thème de deux années et tout le contenu optionnel recueilli pour les deux ans.
- À partir de juin 2009, les utilisateurs auront le choix entre les fichiers d'un an ou les fichiers de deux ans.
- Grâce aux estimations pour une année, on peut calculer des tendances d'une année à l'autre. Dans le contexte de la collecte continue, chaque estimation annuelle rend compte des caractéristiques moyennes de la population moyenne pour la période.
- Dans le cas de l'estimation des caractéristiques plus rares dans des domaines plus détaillés, l'utilisation des fichiers de deux ans, ou même des fichiers pluriannuels, sera nécessaire pour assurer la qualité des données (c.v. de 33 % et minimum de dix répondants comportant les caractéristiques).
- La convention de désignation des variables de l'ESCC a été modifiée légèrement, afin de rendre compte du fait que la même variable est recueillie. La lettre désignant le cycle (p. ex., « e » pour le cycle 3.1) a été supprimée du nom de la variable.