

PONDÉRATION DU CYCLE 2.1 DE L'ENQUÊTE SUR LA SANTÉ DANS LES COLLECTIVITÉS CANADIENNES

Marco Grenier¹ and François Brisebois

RÉSUMÉ

La composante régionale de l'Enquête sur la santé dans les collectivités canadiennes est menée par Statistique Canada auprès d'environ 130 000 répondants et vise à produire des données à l'échelle de 133 régions sociosanitaires. Pour la pondération du cycle 2.1, en plus d'avoir à composer avec les mêmes complexités qu'au cycle 1.1 telles que l'utilisation de deux bases de sondage, de nouveaux défis se sont présentés. Mentionnons, entre autres, l'intégration à l'enquête d'une étude visant à déterminer l'effet du mode de collecte (personne vs téléphone) sur les estimations. La stratégie développée pour la pondération est présentée en mettant l'accent sur les défis rencontrés.

MOTS CLÉS : Bases duales; pondération; sous-échantillon.

ABSTRACT

The regional component of the Canadian Community Health Survey is conducted by Statistics Canada with a total sample of 130,000 respondents in order to provide region-level data for 133 health regions. For the weighting of cycle 2.1, in addition to dealing with the same complexities as in cycle 1.1, such as the use of two survey frames, new challenges arose. One example is the integration of a study measuring the collection mode effect (in person vs. telephone) on estimates. The strategy developed to produce the survey weights is presented with emphasis placed on the new challenges.

KEY WORDS: Dual frames, Sub-sample, Weighting.

1. INTRODUCTION

L'Enquête sur la santé dans les collectivités canadiennes (ESCC) est une enquête menée par Statistique Canada dont l'objectif est de procurer des estimations sur les déterminants de la santé, sur l'état de la santé et sur l'utilisation du régime de santé. L'ESCC comprend deux enquêtes effectuées au cours d'un cycle répétitif de deux ans : une enquête régionale la première année (.1) auprès d'un échantillon de plus de 130 000 répondants et une enquête provinciale la deuxième année (.2) auprès d'un échantillon de 30 000 répondants. En 2003, l'ESCC en était à la première année de son 2^e cycle (2.1). Pour ce cycle, certains ajouts se sont greffés aux paramètres habituels de l'ESCC. Effectivement, en plus de l'échantillon régulier de plus de 130 000 répondants, trois sous-échantillons ont été sélectionnés dans le but de produire des estimations nationales et provinciales sur des sujets choisis. Notons également l'intégration à l'enquête d'une étude visant à déterminer l'effet du mode de collecte utilisé lors de l'interview sur les réponses fournies par les répondants.

Le présent article décrit la pondération du cycle 2.1 de l'ESCC et donne un aperçu des défis rencontrés. La section 2 donne des détails sur le plan de sondage de même que sur les deux bases utilisées par l'ESCC. La section 3 décrit la stratégie de pondération en passant à travers toutes les étapes. Finalement, la section 4 traite des défis supplémentaires associés au cycle 2.1 de l'ESCC. Notons que la stratégie de pondération qui est présentée est celle développée pour les 10 provinces du Canada et que la stratégie de pondération pour les trois territoires diffère légèrement étant donné certaines différences dans le plan de sondage. Pour plus de détails sur la pondération de l'ESCC (cycle 2.1), consulter le guide du fichier de microdonnées à grande diffusion (FMGD) du cycle 2.1 de l'ESCC (Statistique Canada, 2005).

2. PLAN ET BASES DE SONNAGE

¹ Marco Grenier, Statistique Canada, Immeuble R.H. Coats, Ottawa (ON), Canada, K1A 0T6, gremarc@statcan.ca

Afin d'obtenir un échantillon suffisant dans chacune des 133 régions sociosanitaires (RSS), l'ESCC a eu recours à deux bases de sondage. Une base aréolaire utilisée à titre de base principale et une base téléphonique comme base complémentaire. Les deux bases sont utilisées simultanément pour 127 des 133 RSS alors qu'une seule base est utilisée pour les six autres RSS. La base aréolaire est celle conçue pour l'Enquête sur la Population Active (EPA) et a fourni environ 60 000 répondants pour le cycle 2.1 de l'ESCC. Les interviews pour les répondants de la base aréolaire sont des interviews personnelles assistées par ordinateur (IPAO). De son côté, la base téléphonique est constituée de deux composantes mutuellement exclusives : une base liste de numéros de téléphone qui est utilisée dans 125 des 133 RSS et une base de composition aléatoire (CA) de numéros de téléphone qui est utilisée pour seulement sept RSS. La base téléphonique a fourni environ 70 000 répondants pour le cycle 2.1 de l'ESCC et les interviews ont été réalisées en utilisant la méthode d'interviews téléphoniques assistées par ordinateur (ITAO). Notons que la proportion de répondants provenant de la base aréolaire a considérablement diminué pour le cycle 2.1 par rapport au cycle 1.1 justifiant ainsi la réalisation d'une étude sur l'effet du mode de collecte. La section 4.2 discute plus en détails des particularités de cette étude et de son impact sur la pondération.

2.1 Base aréolaire

La base aréolaire de l'EPA utilise un plan d'échantillonnage stratifié à plusieurs degrés. Notons que, étant donné la complexité du plan, le présent article ne fera qu'un survol nécessaire à la compréhension de la base utilisée pour l'ESCC. Plus de détails sur le plan d'enquête sont donnés dans le document « Méthodologie de l'Enquête sur la population active du Canada » (Statistique Canada, 1998).

Tout d'abord, le pays est divisé en strates selon des critères géographiques, économiques et démographiques. Chaque strate est subdivisée en grappes habituellement définies par une fraction ou un ensemble de secteurs de dénombrement. Ces grappes représentent les unités primaires d'échantillonnage. Le premier degré d'échantillonnage consiste à sélectionner des grappes proportionnellement à leur taille. Ensuite, tous les logements faisant partie des grappes sélectionnées sont listés et un échantillon systématique de logements est sélectionné. Cette étape représente le deuxième degré d'échantillonnage. Finalement, au troisième degré, une personne est sélectionnée avec probabilité inégale à l'intérieur de chaque ménage qui avait été choisi au second degré.

2.2 Base téléphonique

Contrairement au cycle 1.1, la base liste de numéros de téléphone représente la base téléphonique principale. En fait, la base de CA est utilisée seulement dans les RSS pour lesquelles la base aréolaire n'est pas disponible pour compléter la base liste. L'utilisation de la base liste plutôt que de la base de CA est justifiée par les faibles taux de succès des intervieweurs à rejoindre un ménage à partir de la base de CA obtenus au cours du cycle 1.1 de l'ESCC. Il est vrai que la couverture de la base liste est plus faible que celle de la base de CA, mais cette sous-couverture est prise en compte dans la pondération (voir section 3). De plus, l'impact de cette sous-couverture est amoindri par la présence de la base aréolaire partout où la base liste est utilisée.

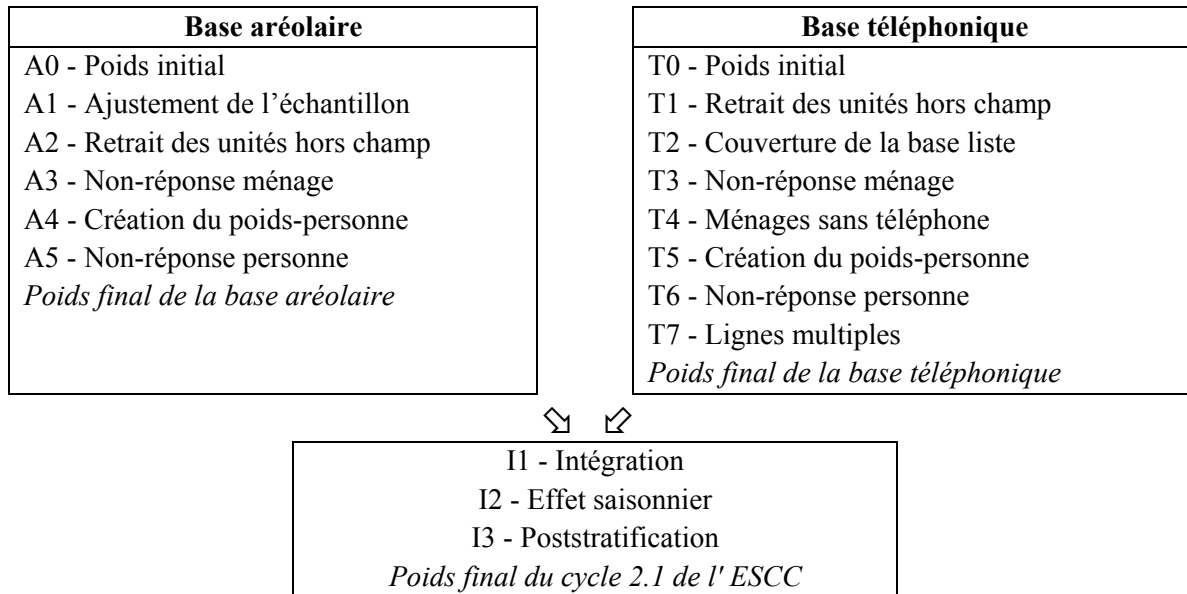
La base liste consiste en une liste de numéros de téléphone obtenue à partir du CD-ROM « Infobase Telephone Directories, Canadian edition ». Une RSS est assignée à chaque numéro de téléphone en se basant sur le code postal fourni sur le CD-ROM. La base liste utilise un plan stratifié pour lequel les RSS représentent les strates. Dans chaque strate, on sélectionne le nombre requis de numéros de téléphone selon un plan d'échantillonnage aléatoire simple. Pour ce qui est de la base de CA, l'échantillonnage a été réalisé selon la méthode d'élimination des banques non valides adoptée par l'Enquête sociale générale, décrite dans le document de Norris et Paton (1991). Une banque de cent numéros (les 8 premiers chiffres d'un numéro à 10 chiffres) est considérée non valide si elle ne contient aucun numéro résidentiel. Toutes les banques valides sont ensuite regroupées ensemble pour former des strates de CA pour former le mieux possible les RSS impliquées. On choisit au hasard une banque et on génère un numéro entre 00 et 99 pour compléter le numéro à 10 chiffres. Ce procédé est répété jusqu'à ce que la taille visée soit atteinte.

3. PONDÉRATION

Étant donné les différences entre les deux types de bases de sondage, les répondants de la base aréolaire sont traités séparément des répondants de la base téléphonique. Les répondants provenant de ces deux bases sont ensuite combinés à

l'étape d'intégration. Plusieurs ajustements sont appliqués aux poids initiaux de l'ESCC (cycle 2.1) afin de tenir compte de toutes les particularités relatives aux bases de sondage, à la méthode d'échantillonnage et au déroulement des activités de collecte. La présente section décrit les ajustements utilisés dans la pondération dans l'ordre auquel ils sont appliqués. La description de ces ajustements est très brève mais plus de détails sont disponibles en consultant le guide du FMGD (Statistique Canada, 2005). Le diagramme A présente la liste des ajustements pour chaque base. Les lettres A et T sont utilisées comme préfixe pour faire référence aux bases aréolaire et téléphonique respectivement. De son côté, le préfixe « I » est utilisé pour les ajustements appliqués aux unités intégrées.

Diagramme A – Sommaire de la stratégie de pondération de l'échantillon complet



3.1 Pondération de l'échantillon provenant de la base aréolaire

A0 – Poids initial

Puisque le mécanisme utilisé pour sélectionner l'échantillon de la base aréolaire a été celui établi pour l'EPA, le poids initial a dû être calculé selon les particularités de cette enquête. Le produit des probabilités de sélection des deux premiers degrés d'échantillonnage représente la probabilité de sélection du logement et son inverse représente le poids initial du logement. Pour plus de détails sur le mécanisme de sélection, de même qu'une définition plus complète des strates et des grappes, se référer à Statistique Canada (1998).

A1 – Ajustement de l'échantillon

Certaines modifications ont dû être faites au mécanisme standard de l'EPA lors de la sélection de l'échantillon pour le cycle 2.1 de l'ESCC. Étant donné que les besoins en échantillon pour l'ESCC sont légèrement différents de ceux de l'EPA, l'échantillon provenant de la base aréolaire de l'EPA a dû être ajusté pour en tenir compte. Un ajustement aux poids a été appliqué pour refléter cet ajustement.

A2 – Retrait des unités hors champ

Parmi tous les logements échantillonnés, une certaine proportion de ceux-ci est, lors de la collecte, identifiée comme étant hors du champ de l'enquête. Des logements détruits ou en construction, des logements vacants, saisonniers ou secondaires, de même que des établissements, sont tous des exemples de cas hors champ pour l'ESCC. Ces logements sont tout simplement retirés de l'échantillon, ne laissant plus que les logements faisant partie du champ de l'enquête.

A3 – Non-réponse ménage

Lors de la collecte, une certaine proportion des ménages interviewés a inévitablement résulté en non-réponse. Ceci survient habituellement lorsque le ménage refuse de participer à l'enquête, fournit des données inutilisables, ou encore, ne peut être rejoint pour réaliser l'interview. Le poids des ménages non-répondants est redistribué aux répondants à l'aide de classes de réponse formées de façon indépendante à l'intérieur de chaque RSS. L'algorithme CHAID (Chi-Square Automatic Interaction Detector), disponible dans Knowledge Seeker (Angoss Software, 1995), a permis d'identifier les

caractéristiques qui divisent le mieux l'échantillon en groupes selon leurs propensions à répondre. Puisque l'information disponible auprès des non-répondants est très limitée, seules quelques caractéristiques disponibles sur les bases de sondage comme le mois de collecte et l'indicateur rural/urbain ont pu être utilisées pour la création des classes de réponse.

A4 – Création du poids-personne

Puisque l'unité d'échantillonnage finale pour l'ESCC est la personne, le poids-ménage calculé jusqu'ici doit être converti en un poids-personne. Celui-ci est obtenu en multipliant le poids A3 par l'inverse de la probabilité de sélection de la personne choisie dans le ménage. Cette probabilité de sélection dépend de la composition du ménage et du groupe d'âge auquel chaque membre du ménage appartient. En effet, afin d'atteindre les tailles d'échantillon visées pour les moins de 20 ans, il a fallu augmenter la probabilité de sélection pour ces personnes. Pour les ménages sans personne de moins de 20 ans, la probabilité de sélection était la même pour tous les membres du ménage. Noter que seulement une personne a été sélectionnée par ménage contrairement au cycle 1.1 pour lequel deux personnes étaient sélectionnées dans certains ménages.

A5 – Non-réponse personne

Même si le ménage a accepté de participer à l'enquête, il est possible que la personne sélectionnée refuse de répondre aux questions ou que les intervieweurs ne réussissent pas à entrer en contact avec la personne sélectionnée. De tels cas sont définis comme étant des non-réponses à l'échelle de la personne. Tout comme pour la non-réponse à l'échelle du ménage, un ajustement est appliqué aux poids des répondants à l'intérieur de classes définies à partir des caractéristiques disponibles pour les répondants et non-répondants. L'algorithme CHAID a encore une fois été utilisé pour obtenir la définition des classes. Après avoir appliqué successivement tous les ajustements, le poids A5 représente le poids final pour la base aréolaire.

3.2 Pondération de l'échantillon provenant de la base téléphonique

T0 – Poids initial

Tel que mentionné précédemment, la base téléphonique est en fait composée de deux bases : la base de composition aléatoire, puis la base liste de numéros de téléphone. Les unités provenant de ces deux bases sont traitées ensemble et sont donc toutes soumises aux mêmes ajustements à quelques exceptions près. Le poids initial est calculé différemment selon que l'échantillon provienne de la base de CA ou de la base liste. Dans les deux cas, le poids initial est défini comme étant l'inverse de la probabilité de sélection, mais puisque les méthodes de sélection diffèrent, les probabilités diffèrent aussi. Pour la base de CA, la probabilité de sélection est le ratio entre le nombre d'unités échantillonnées et cent fois le nombre de banques présentes dans la strate de CA. Pour ce qui est de la base liste, la probabilité de sélection correspond au ratio entre le nombre d'unités échantillonnées et le nombre de numéros de téléphone dans la liste pour la RSS.

T1 - Retrait des unités hors champ

Les numéros de téléphone associés à des logements hors du champ de l'enquête (ex. : entreprises), de même que les numéros hors service sont tous des exemples de cas hors champ pour la base téléphonique. Comme pour la base aréolaire, ces cas sont retirés de l'échantillon, ne laissant ainsi que les logements dans le champ de l'enquête.

T2 – Couverture de la base liste

Puisque la base liste ne couvre pas certains numéros de téléphone qui sont toutefois couverts par la base de CA, un ajustement doit être apporté au poids initial des unités de la base liste de façon à ce que les deux bases soient comparables en terme de couverture. Cet ajustement consiste à gonfler le poids des unités de la base liste proportionnellement au taux de couverture dans chaque RSS. L'estimation de ce taux de couverture a pu être faite à l'aide des données recueillies auprès de l'échantillon de la base aréolaire. Pour dériver le taux de couverture désiré, on a calculé le pourcentage des numéros de téléphone recueillis auprès des répondants de la base aréolaire qui étaient présents sur la base liste.

T3 - Non-réponse ménage

L'ajustement fait ici pour compenser l'effet de la non-réponse ménage est identique à celui appliqué pour la base aréolaire (ajustement A3).

T4 - Ménages sans téléphone

Les ménages canadiens n'ayant pas accès à une ligne téléphonique résidentielle privée ne sont pas représentés par la base téléphonique alors qu'ils le sont sur la base aréolaire. Afin d'obtenir une couverture comparable pour les deux bases, un facteur d'ajustement est calculé pour gonfler les poids des répondants de la base téléphonique. Ce facteur d'ajustement est dérivé à l'échelle de la RSS en utilisant l'information recueillie auprès des répondants de la base aréolaire.

T5 - Création du poids-personne

Tout comme l'ajustement A5, cet ajustement permet de convertir ce qui était jusqu'à cette étape-ci un poids-ménage en un poids-personne. L'algorithme de sélection de la personne à l'intérieur du ménage étant le même que pour la base aréolaire, le calcul du facteur d'ajustement est effectué de la même façon.

T6 - Non-réponse personne

L'ajustement fait ici pour compenser l'effet de la non-réponse ménage est identique à celui appliqué pour la base aréolaire (ajustement A5).

T7 - Lignes multiples

Le fait que certains ménages aient plus d'une ligne téléphonique résidentielle a un impact sur la pondération; plus le ménage possède de lignes, plus grande est sa probabilité d'être sélectionné. Conséquemment, les poids doivent être ajustés pour tenir compte du nombre de lignes résidentielles que le ménage possède. Le facteur d'ajustement représente l'inverse du nombre de lignes téléphoniques résidentielles. Après avoir appliqué successivement tous les ajustements, on obtient un poids final pour la base téléphonique. Ce poids sera plus tard intégré au poids final de la base aréolaire pour créer le poids final du cycle 2.1 de l'ESCC.

3.3 Intégration et ajustements finaux

I1 - Intégration des bases aréolaire et téléphonique

Cette étape consiste à intégrer les poids finaux des échantillons aréolaire et téléphonique créés jusqu'à maintenant en un seul poids en appliquant une méthode d'intégration. La méthode choisie est la même que celle utilisée pour le cycle 1.1. Un facteur d'ajustement, compris entre 0 et 1, est déterminé de façon à représenter l'importance relative de chaque échantillon dans l'échantillon total. Cette importance relative est mesurée en termes de taille d'échantillon et d'effet de plan. Plus la proportion d'échantillon qu'une base représente dans l'échantillon total est grande, plus grande sera son importance relative dans l'échantillon total. Pour les détails théoriques sur l'intégration, consulter Skinner et Rao (1996).

I2 - Effet saisonnier

La collecte des données devait initialement être répartie également sur les douze mois de l'enquête. Certains événements survenus pendant la collecte, combinés aux ajustements appliqués aux poids, ont eu un impact sur la distribution saisonnière de l'échantillon. Un ajustement additionnel a dû être ajouté pour assurer qu'il n'y ait pas d'effet saisonnier dans les estimations produites. Cet ajustement a été fait de façon à ce que la somme des poids des unités interviewées lors d'une des quatre saisons représente exactement 25 % de la somme des poids de l'échantillon total.

I3 - Poststratification

La dernière étape nécessaire afin d'obtenir le poids final du cycle 2.1 de l'ESCC est la poststratification. La poststratification est appliquée de sorte que la somme des poids finaux corresponde aux estimations de population définies à l'échelle des RSS, pour chacun des 10 groupes d'âge-sexe d'intérêt, c'est-à-dire les cinq groupes d'âge 12-19, 20-29, 30-44, 45-64, 65+, pour chacun des deux sexes. Les estimations de population utilisées sont basées sur les comptes du Recensement de 1996, de même que sur les comptes de naissance, décès, immigration et émigration.

4. DÉFIS SUPPLÉMENTAIRES

4.1 Sous-échantillons

Afin de pouvoir estimer des caractéristiques basées sur les sous-échantillons, un poids supplémentaire a dû être créé pour chacun de ces trois sous-échantillons. En résumé, la stratégie de pondération utilisée est la même que pour l'échantillon total, à quelques différences près. Le premier ajout consiste à redistribuer le poids des ménages non sélectionnés pour le sous-échantillon aux ménages sélectionnés et ce à l'intérieur de chaque base puisque les sous-échantillons proviennent des deux bases. Cette redistribution des poids est réalisée à partir de l'échantillon total juste après les étapes A2 et T1 et se fait par RSS puisque la sélection des sous-échantillons s'est faite par RSS. Les ajustements qui suivent sont les mêmes que pour la pondération de l'échantillon total mais appliqués à l'échelle de la province plutôt que de la RSS étant donné le caractère provincial des sous-échantillons. Notons également que des particularités propres à chacun des sous-échantillons font en sorte que la stratégie de pondération varie quelque peu d'un sous-échantillon à l'autre.

4.2 Étude sur l'effet du mode de collecte

Cette étude, qui avait pour but de déterminer si le mode de collecte pouvait avoir une influence sur certaines caractéristiques mesurées par l'ESCC, était complètement intégrée à l'enquête. Cela signifie que les cas faisant partie de l'étude faisaient également partie de l'échantillon régulier. Cette étude a eu lieu auprès de 11 RSS sélectionnées de façon à représenter toutes les régions du pays (Est, Québec, Ontario et Ouest). Pour ces 11 RSS, l'étude a eu comme conséquence d'ajouter une troisième base de sondage. En fait, l'échantillon de l'étude a été sélectionné à partir de la base liste de numéros de téléphones mais les différences dans le plan d'échantillonnage imposent un traitement différent aux répondants sélectionnés pour l'étude. Plus de détails sur l'étude sont disponibles dans Béland et St-Pierre (2004).

5. CONCLUSION

Le cycle 2.1 de l'ESCC est un exemple qui démontre bien que chaque décision prise au moment de la conception du plan d'enquête ou pendant les activités de collecte peut avoir un impact sur les méthodes d'estimation. Tel que présenté dans cet article, l'utilisation de bases duales, la sélection de sous-échantillons et l'intégration d'une étude sur le mode de collecte sont tous des facteurs ayant contribué à complexifier la pondération du cycle 2.1. C'est cependant le prix à payer pour atteindre les objectifs et répondre aux attentes des utilisateurs tout en respectant les contraintes opérationnelles.

RÉFÉRENCES

- ANGOSS Software (1995). *Knowledge Seeker IV for Windows - User's Guide*. ANGOSS Software International Limited.
- Béland, Y. et St-Pierre M. (2004). « *Mode effects in the Canadian Community Health Survey: a comparison of CAPI and CATI* ». 2004 Proceedings of the Survey Research Methods Section, American Statistical Association. À paraître.
- Norris, D.A. et Paton, D.G. (1991). « *Canada's General Social Survey: Five Years of Experience* ». *Survey Methodology*, 17, 227-240.
- Skinner, C.J. et Rao, J.N.K. (1996). « *Estimation in Dual Frame Surveys With Complex Designs* », *Journal of the American Statistical Association*, 91, pp. 349-356.
- Statistique Canada (2005). « *Guide du fichier de microdonnées à grande diffusion de l'ESCC (cycle 2.1)* ». Statistique Canada. À paraître.
- Statistique Canada (1998). « *Méthodologie de l'Enquête sur la population active du Canada* ». Statistique Canada. Cat. No. 71-526-XPB.