No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Actes du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

La combinaison de cycles de l'Enquête sur la santé dans les collectivités canadiennes

Steven Thomas[1]

L'Enquête sur la santé dans les collectivités canadiennes comporte deux enquêtes transversales menées en alternance sur un cycle annuel. Les deux enquêtes recueillent des renseignements généraux sur la santé; la deuxième, de moindre envergure, recueille des renseignements supplémentaires sur des aspects précis de la santé. Même si la taille des échantillons est importante, les utilisateurs sont intéressés à combiner les cycles de l'ESCC afin d'améliorer la qualité des estimations et de produire des estimations pour les petits domaines géographiques ou pour des caractéristiques ou des populations rares. Le présent document porte sur certains enjeux liés à la combinaison des cycles de l'ESCC, dont les interprétations possibles du résultat combiné, ainsi que sur certaines méthodes de combinaison des cycles.

MOTS-CLÉS : Enquête transversale; combinaison de cycles; populations rares.

1 Introduction

1 Aperçu

L'***Enquête sur la santé dans les collectivités canadiennes*** (ESCC) comporte deux enquêtes transversales par sondage, menées en alternance sur une base annuelle. Il s'agit d'une enquête au contenu général, menée auprès d'un vaste échantillon, et d'une enquête portant en outre sur un sujet particulier, menée auprès d'un échantillon restreint. La première enquête (tenue en 2001, en 2003, en 2005…), ou le cycle .1, vise à recueillir des renseignements généraux sur la santé auprès d'un échantillon assez important pour produire des renseignements sur plus de cent régions sociosanitaires canadiennes. À cette fin, il faut recueillir des données auprès de plus de 130 000 répondants. La deuxième enquête (tenue en 2002, en 2004…), ou le cycle .2, porte sur un aspect précis de la santé et recueille des données auprès d'un échantillon restreint de 30 000 répondants pour produire des renseignements à l'échelle provinciale.

Malgré la taille importante des échantillons, il existe des cas où un cycle unique de l'enquête ne répond pas aux besoins des utilisateurs. La taille de l'échantillon peut être suffisante pour permettre de publier des estimations sur les populations ciblées, mais pas assez importante pour offrir la capacité statistique de déceler des écarts significatifs. Dans d'autres cas, les chercheurs sont intéressés à repousser les limites d'un cycle unique de l'enquête pour étudier des sous-populations fondées sur des caractéristiques sociodémographiques ou géographiques, ou encore sur des caractéristiques rares, pour lesquelles l'échantillon est tout simplement insuffisant.

Pour accroître les utilisations des données, les utilisateurs sont intéressés par la possibilité de combiner les différents cycles de l'ESCC afin d'estimer leurs paramètres d'intérêt. Le plus souvent, tous les cycles .1 recueillent des données sur les mêmes caractéristiques et les cycles .2 recueillent une partie des mêmes renseignements. Il est donc possible de combiner les données. Toutefois, il faut tenir compte de plusieurs enjeux et, dans certains cas, la combinaison des données est contre-indiquée. Dans d'autres cas, où la combinaison est jugée indiquée, il existe différentes méthodes de combinaison et le choix de la méthode dépend des détails de l'analyse.

[1]Steven Thomas, Statistique Canada, 16e étage, Immeuble R.H. Coats, pré Tunney, Ottawa (Ontario), Canada K1A 0T6.

Dans le présent document, nous résumons les différents enjeux à prendre en compte avant de tenter de combiner des enquêtes, en nous inspirant surtout de l'ESCC, dont nous présentons un aperçu dans la section 2. Dans la section 3, nous étudions les interprétations possibles d'une analyse combinée, puis, dans la section 4, les méthodes de combinaison. Enfin, dans la section 5, nous étudions à l'aide d'exemples les enjeux liés à la combinaison de cycles de l'ESCC et nous présentons un aperçu général des principaux points à prendre en compte à cet égard.

Il convient de préciser que le présent document n'inclut pas toutes les analyses possibles portant sur un ensemble de données combinées. Les méthodes présentées ci-dessous se veulent simplement des lignes directrices pour la plupart des types généraux d'analyse et il peut exister des exceptions aux principes énoncés ci-dessous.

2 L'Enquête sur la santé dans les collectivités canadiennes

L'Enquête sur la santé dans les collectivités canadiennes a été créée dans le cadre du Carnet de route de l'information sur la santé (ICIS, 1999). L'objectif principal de cette enquête consiste à mieux évaluer l'état du système de soins de santé et la santé des Canadiens. À cette fin, deux enquêtes transversales sont menées en alternance sur un cycle annuel, et la taille et l'objectif de ces enquêtes varient d'une année à l'autre.

2.1 Les cycles .1

La première enquête, ou le cycle .1, a pour objectif principal de recueillir des renseignements généraux sur la santé à l'échelle des régions sociosanitaires, niveau géographique infraprovincial. La population cible de cette enquête comprend toutes les personnes de 12 ans et plus qui habitent des logements privés dans les dix provinces et les trois territoires pendant la période de collecte. Sont exclus du champ de l'enquête les habitants des réserves indiennes ou des terres de la Couronne, les résidents des établissements, le personnel à temps plein des Forces canadiennes et les résidents de certaines régions éloignées. Lors de chaque cycle, on a besoin de plus de 130 000 répondants pour estimer adéquatement les caractéristiques de la santé des groupes d'intérêt selon l'âge et le sexe. À ce jour, on a recueilli et publié les données de trois cycles de l'enquête .1. On a recueilli ces données aux périodes suivantes : de septembre 2000 à octobre 2001 (cycle 1.1), de janvier à décembre 2003 (cycle 2.1) et de janvier à décembre 2005 (cycle 3.1).

Pour tous les cycles .1, on a utilisé trois bases de sondage : une base aréolaire de l'Enquête sur la population active (EPA), une base fondée sur une liste téléphonique créée en combinant des annuaires téléphoniques à l'échelle du pays et une base fondée sur la composition aléatoire (CA), utilisée dans les régions éloignées. On recueille donc les données au moyen d'un mélange d'interviews téléphoniques et d'interviews sur place. Dans les cycles 2.1 et 3.1 de l'enquête, on a contrôlé l'échantillonnage pour qu'environ 50 % de l'échantillon proviennent de la base aréolaire et 50 % des bases fondées sur la liste téléphonique et sur la CA. Toutefois, ce n'était pas le cas du cycle 1.1, où 83 % de l'échantillon provenaient de la base aréolaire et 17 % des deux autres bases.

Lors de chaque cycle .1, on a apporté de légères modifications au plan d'échantillonnage. Premièrement, les parties intéressées ont toujours eu la possibilité d'acheter des unités d'échantillonnage supplémentaires afin de mieux cibler leurs besoins particuliers. Certaines régions sociosanitaires l'ont fait afin d'obtenir des estimations concernant des sous-régions sociosanitaires. Pour éviter des frais exceptionnels, on a souvent choisi l'échantillon supplémentaire dans la base téléphonique, où il était moins coûteux de mener des interviews. Deuxièmement, lors de chaque cycle .1, il était possible de choisir un contenu facultatif à ajouter au questionnaire. Les régions sociosanitaires qui ont choisi cette option ont donc pu obtenir plus de renseignements sur des enjeux propres à leur territoire. Enfin, il y avait un contenu concernant des sous-échantillons où, pour une partie de l'échantillon principal, on a posé des questions supplémentaires pour obtenir des estimations à un niveau géographique moins détaillé. Ces sous-échantillons différaient d'un cycle à l'autre. Pour plus de renseignements sur les sous-échantillons, on peut consulter le guide de l'utilisateur qui accompagne les données de l'ESCC.

Le plus souvent, le questionnaire est demeuré inchangé entre les cycles .1. Cette mesure visait à maintenir la cohérence et la comparabilité entre les cycles. Toutefois, il n'en a pas toujours été ainsi. Avec le temps, on a modifié certaines questions pour mieux recueillir l'information désirée. On peut modifier le questionnaire lorsqu'on se rend compte que la question est mal interprétée ou lorsque des concepts sont modifiés. L'utilisateur de données doit prendre note que cette modification n'entraîne pas toujours un changement du nom de la variable de l'ESCC.

2.2 Les cycles .2

Les cycles .2 de l'ESCC ont varié considérablement par rapport aux cycles .1, où la cohérence était importante. Les cycles .2 ont été conçus pour enquêter sur un aspect donné de la santé dans une population cible donnée. Leur plan d'enquête n'assure aucunement la comparabilité avec d'autres cycles. Les bases utilisées, les populations ciblées, les questionnaires administrés et la stratégie générale d'échantillonnage sont propres à chaque enquête. À ce jour, on a mené deux cycles, pour lesquels on avait généralement besoin d'estimations à l'échelle provinciale et il fallait plus de 30 000 répondants pour estimer des proportions au niveau de précision souhaité à ce niveau de détail. Les estimations n'étaient pas exigées pour les trois territoires.

Les données du cycle 1.2 ont été recueillies de mai à novembre 2002 et portaient sur la santé mentale. La population cible comprenait toutes les personnes de 15 ans et plus des dix provinces, avec les mêmes exclusions que dans le cas des cycles .1. L'échantillonnage visait à établir des estimations provinciales, sauf en Nouvelle-Écosse et en Ontario, où l'on avait besoin d'estimations infraprovinciales. Pour ce cycle, on a utilisé uniquement la base aréolaire de l'EPA pour cibler cette population et l'on a recueilli l'information au moyen d'interviews sur place.

Les données du cycle 2.2 ont été recueillies de janvier à décembre 2004 et portaient sur la nutrition. La population cible de cette enquête comprenait des personnes de tous les âges. Dans chaque province, en raison de la difficulté de cibler les enfants à partir d'une seule base de sondage, on a utilisé deux bases différentes pour cibler convenablement cette population. Le plus souvent, on a combiné la base aréolaire de l'EPA avec les répondants du cycle 2.1 qui avaient des enfants dans le ménage. Dans le cas du cycle 2.2, on a également administré un questionnaire de suivi pour recueillir un deuxième ensemble de données sur l'alimentation auprès d'environ 10 000 répondants. On a recueilli les données au moyen d'interviews sur place pour la première enquête et d'interviews téléphoniques pour l'enquête de suivi.

3 Analyse combinée

Plusieurs auteurs, dont Kish (1999), ont traité en détail de la combinaison d'enquêtes. Dans le contexte général, il s'agit de combiner des enquêtes qui recueillent les mêmes renseignements et qui, ensemble ou indépendamment, représentent la population d'intérêt. On produit ainsi des estimations plus précises concernant les caractéristiques communes. Cette situation survient très souvent, car des enquêtes différentes recueillent les mêmes renseignements de base. Dans le cas d'enquêtes à passages répétés, on pose souvent aux répondants les mêmes questions lors de chaque cycle. Pour ces enquêtes semblables, il existe différentes méthodes de combinaison des renseignements, dont la meilleure dépend des détails de l'analyse. Ces détails sont très importants et l'on ne saurait en faire abstraction. Le chercheur doit avoir une idée précise de la population d'intérêt et des caractéristiques pour lesquelles on a besoin d'estimations.

3.1 Définition de l'analyse

Lorsqu'un chercheur entreprend son analyse, la première étape consiste à définir l'analyse en fonction de la caractéristique d'intérêt, qui est habituellement celle d'une population donnée. Les deux types généraux d'analyse sont la description et l'inférence. Le chercheur tente de décrire les caractéristiques d'une population finie ou d'établir des inférences au sujet d'un modèle ou d'une superpopulation. Une population finie est un groupe fini de personnes à un moment fixe. Le chercheur qui étudie cette population s'intéresse surtout à des statistiques descriptives comme des moyennes, des proportions et des totaux. Par exemple, il peut s'intéresser à l'indice de masse corporelle moyen de la population. Dans le cas de l'inférence, le chercheur tente de modéliser la relation entre les caractéristiques. Il s'intéresse aux paramètres du modèle au lieu de tenter de décrire une population. On peut citer comme exemple le lien entre le tabagisme et le cancer : ici, le chercheur ne s'intéresse pas à confiner sa recherche à une population donnée à un moment donné. Pour plus de renseignements sur les notions de population finie et de population infinie, on peut consulter Binder et Roberts (2006).

Au moment de combiner des renseignements provenant d'enquêtes différentes, le chercheur doit définir soigneusement l'analyse à effectuer. En général, la méthode de combinaison suppose que les statistiques combinées estiment la même chose. Pour des populations finies, il existe souvent de légers écarts entre les populations ainsi que des écarts entre les caractéristiques de ces populations qui peuvent rendre cette hypothèse difficile à confirmer. Dans

le cas des enquêtes à passages répétés, les populations et leurs caractéristiques varient en raison d'une évolution temporelle. Si les cycles de l'enquête mesuraient exactement la même chose, leur répétition serait superflue. Ces variations peuvent être minimes si les cycles sont relativement rapprochés, mais peuvent être non ignorables si le temps écoulé entre les enquêtes est plus long. Compte tenu des écarts intrinsèques dans la réalité mesurée par une enquête à passages répétés, il n'y a sans doute pas lieu de mener une analyse descriptive autre qu'une analyse chronologique de moyennes mobiles, mais si l'analyse est décrite avec soin, il peut être justifié de mener des analyses successives. En général, le problème ne se pose pas lorsqu'on établit des inférences, car on s'intéresse au modèle plutôt qu'à une population donnée. On suppose que ce modèle s'appliquera aux deux enquêtes et qu'on peut expliquer les écarts grâce au modèle.

4 Méthodes de combinaison

Pour combiner des renseignements provenant de différentes sources de données, il existe plusieurs méthodes, dont chacune présente des avantages et des inconvénients. On peut répartir ces méthodes en deux catégories générales définies par Binder et Roberts (2006). La méthode individuelle, ou d'estimation composite, consiste à calculer des estimations distinctes pour chaque enquête, puis à les combiner, alors que la méthode de mise en commun consiste à combiner les données de l'échantillon à l'échelle des microdonnées et à traiter l'ensemble de données ainsi obtenu comme s'il s'agissait d'un échantillon provenant d'une seule population.

4.1 La méthode individuelle

Selon la méthode individuelle, on estime les caractéristiques d'intérêt à partir de chaque source de données et l'on calcule une moyenne pondérée des estimations. Cette méthode d'estimation composite est attrayante en théorie, mais sa mise en œuvre peut être longue si l'on a besoin de nombreuses estimations (Chu, Brick et Kalton, 1999). Selon ce scénario, supposons que deux enquêtes recueillent indépendamment des renseignements sur un paramètre de population θ. On peut calculer des estimations de θ à partir de chaque source de données pour obtenir deux estimations distinctes $\hat{\theta}_1$ et $\hat{\theta}_2$. On peut alors calculer la moyenne pondérée de ces deux estimations comme suit :

$$\hat{\theta}_C = \alpha \hat{\theta}_1 + (1-\alpha) \hat{\theta}_2 .$$

Dans ce modèle, si $\hat{\theta}_1$ et $\hat{\theta}_2$ sont des estimations sans biais de θ, alors $\hat{\theta}_C$ est également sans biais pour n'importe quelle valeur choisie de α.

Si les estimations ne mesurent pas tout à fait la même chose, il peut s'avérer difficile d'interpréter une estimation combinée. Tel est le problème inhérent à la combinaison d'enquêtes qui couvrent des moments différents puisque, souvent, ces enquêtes ne mesurent pas la même chose. Ce n'est qu'après avoir effectué une analyse attentive pour vérifier que les mesures sont les mêmes que l'on peut combiner les résultats et les interpréter selon cette méthode.

Cela ne signifie pas qu'on ne peut pas combiner les estimations provenant de différentes enquêtes ou de différents cycles de la même enquête. Toutefois, on ne peut formuler l'interprétation décrite plus haut puisqu'on ne peut formuler l'hypothèse selon laquelle chaque enquête estime le même paramètre de population θ. Dans ce cas, on peut interpréter l'estimation combinée selon la méthode individuelle seulement comme l'estimation d'une moyenne pondérée de deux valeurs différentes. Un autre problème peut survenir lorsqu'il s'agit de combiner des enquêtes dépendantes, car il faut tenir compte de la dépendance au moment d'estimer la variance d'une analyse combinée.

On peut choisir de nombreuses valeurs pour α. On pourrait choisir 0,5, ce qui donnerait une moyenne simple, mais ce choix serait inefficace. Il serait préférable de choisir une valeur de α qui réduit au minimum la variance de $\hat{\theta}_C$. Ici, en supposant que les deux enquêtes sont indépendantes, $\alpha=V_2/(V_1+V_2)$, où V_i est la variance de l'échantillon de $\hat{\theta}_i$. Comme V_i est inconnu, il convient mieux d'appliquer une fonction de la taille effective des échantillons n_i* où

$$\alpha=n_1^*/(n_1^*+n_2^*)$$

et

$$n_i^* = n_i / D_i$$

où n_i est la taille de l'échantillon et D_i est l'effet de plan de $\hat{\theta}_i$.

Les méthodes plus efficaces de calcul de α ont pour inconvénient qu'il faudrait calculer α séparément pour chaque variable comprise dans l'analyse. Pour certaines analyses, où la comparabilité des valeurs calculées est importante, on pourrait utiliser la même valeur de α pour toutes les caractéristiques, mais elle serait moins efficace qu'une valeur spécifique pour l'analyse. Dans ce cas, une valeur de α fondée sur une fonction de la taille initiale des échantillons serait sans doute plus appropriée.

4.2 La méthode de mise en commun

La méthode de combinaison fondée sur la mise en commun consiste à combiner différentes enquêtes au niveau des microdonnées. Ici, on intègre les différents ensembles de données afin d'obtenir un seul ensemble qu'on peut alors analyser comme un seul échantillon. Cette option est attrayante en raison de la taille accrue de l'échantillon dont on dispose pour effectuer des analyses. Afin d'obtenir des statistiques comme des paramètres de régression, il peut être préférable d'utiliser une méthode de mise en commun pour calculer les paramètres, plutôt que de prendre une moyenne des paramètres calculée à partir des différentes enquêtes.

Lorsqu'on utilise la méthode de mise en commun, il convient de rééchelonner les poids. En supposant que les deux enquêtes estiment la même population, l'analyse des totaux a pour effet de surestimer le total proportionnellement à un facteur du nombre d'enquêtes à combiner (Korn et Graubard, 1998). En divisant les poids d'échantillonnage initiaux par ce facteur pour obtenir de nouveaux poids w_i^*, on obtiendrait un estimateur sans biais du chiffre de population, mais il serait inefficace. Il vaudrait mieux rééchelonner les poids en fonction des variances, comme nous l'avons décrit en 4.1. En appliquant simplement et séparément les mêmes valeurs calculées de α aux poids d'échantillonnage initiaux pour chaque enquête, puis en combinant les ensembles de données initiaux avec ces poids, le chercheur obtiendrait un ensemble de données qui se prêterait à des analyses.

Dans le cas de statistiques linéaires où l'on peut exprimer $\hat{\theta}_i$ sous la forme $\hat{\theta}_i = \sum_{i \in S} w_i y_i$, les estimations seraient les mêmes selon les deux méthodes. Dans le cas d'un total selon la méthode de mise en commun, $\hat{Y}_p = \sum_{i \in S} w_i^* y_i$ et, selon la méthode individuelle :

$$\begin{aligned}
\hat{\theta}_C &= \alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2 \\
&= \alpha \sum_{i \in S1} w_i y_i + (1-\alpha) \sum_{i \in S2} w_i y_i \\
&= \sum_{i \in S1} \alpha w_i y_i + \sum_{i \in S2} (1\text{-}\alpha) w_i y_i \\
&= \sum_{i \in S1} w_i^* y_i + \sum_{i \in S2} w_i^* y_i \\
&= \sum_{i \in S} w_i^* y_i
\end{aligned}$$

Toutefois, il en va autrement des statistiques non linéaires, auquel cas les résultats de la méthode individuelle et ceux de la méthode de mise en commun sont très différents à moins que certains critères ne soient respectés. Dans le cas d'une moyenne, il faudrait que $\sum_{S1} w_i = \sum_{S2} w_i$.

La méthode de mise en commun offre l'avantage qu'après avoir calculé les poids, on peut utiliser l'ensemble de données combinées pour effectuer des analyses multiples. Par contre, il serait malheureusement difficile d'obtenir un ensemble de poids qui serait efficace pour toutes les variables d'intérêt dans toutes les analyses. En effet, une valeur choisie de α qui est très efficace pour une variable peut avoir l'effet opposé pour d'autres variables d'intérêt. Un choix de α fondé sur les effets moyens du plan d'échantillonnage offre l'avantage de prendre en compte plus d'une variable dans la création des poids et serait donc presque optimal pour chaque variable. Toutefois, s'il existe de nombreuses variables d'intérêt dans un vaste ensemble de données, il peut s'avérer difficile de trouver une

solution optimale. Dans ce cas, il peut être plus approprié, et tout aussi efficient, de calculer des poids rajustés en fonction de la taille des échantillons ou de calculer la moyenne simple des poids.

5 Combinaison des cycles de l'ESCC

Comme nous l'avons mentionné plus haut, il existe de nombreux aspects à prendre en compte dans la combinaison d'enquêtes et de nombreuses possibilités de combinaison. Dans la présente section, nous tentons d'éclaircir ces aspects à l'aide de quelques exemples empruntés à l'ESCC. Le premier aspect à souligner est qu'il ne convient pas vraiment de combiner le contenu général des cycles des enquêtes .2 à ceux de l'enquête .1 ou à d'autres cycles .2. Ainsi que nous l'avons montré dans la section 2.2, l'enquête .2 porte sur un aspect précis de la santé et les résultats ne sont pas nécessairement comparables à ceux d'autres cycles. Les problèmes de comparabilité qui peuvent survenir sont attribuables aux effets du mode de collecte, du questionnaire, du temps et des modifications apportées au plan d'échantillonnage. L'effet du mode de collecte intervient puisque les cycles .2 sont habituellement menés au moyen d'interviews sur place et les cycles .1, au moyen d'un mélange d'interviews sur place et d'interviews téléphoniques. Les effets de questionnaire existent puisqu'on pose souvent des questions de manière différente ou dans un ordre différent. Les modifications apportées au plan d'échantillonnage peuvent intervenir puisque des domaines non compris par la base aréolaire, mais compris par les bases téléphoniques, ne sont pas inclus dans la population échantillonnée. Enfin, l'effet de temps peut causer des écarts temporels dans la population et dans les caractéristiques de cette population. Pour la plupart des analyses, il est probable que ni les caractéristiques mesurées ni les populations observées ne sont les mêmes. Par conséquent, bon nombre d'hypothèses nécessaires à la combinaison risquent de ne pas se vérifier.

La combinaison des cycles .2 présente un autre problème lorsqu'on tente de combiner les données du cycle 2.2 à celles du cycle 2.1. Ces cycles ne sont pas indépendants, puisque l'échantillon du cycle 2.1 a servi de base de sondage pour le cycle 2.2. La méthodologie actuelle ne permet pas d'estimer la variance en raison des hypothèses de l'indépendance des cycles. Il convient de souligner que ce problème ne devrait pas se présenter avec les cycles .1. Il existe une faible possibilité qu'on ait visité les mêmes grappes dans plus d'un cycle mais, le plus souvent, on peut considérer les cycles .1 comme indépendants. En raison des réserves énoncées ci-dessus, il est fortement recommandé de ne pas utiliser les cycles .2 dans la combinaison des données de l'ESCC. Quant aux autres cycles et au contenu de certains sous-échantillons, ils peuvent encore présenter certains problèmes mais, grâce à une réflexion et à des vérifications attentives, il est possible d'effectuer une analyse combinée.

5.1 Exemple d'une population finie

Le premier exemple consiste à combiner les cycles .1 de l'ESCC afin d'obtenir de meilleures statistiques descriptives pour le Nunavut. Il pourrait s'agir de statistiques comme les taux de tabagisme ou de diabète. Avec cet exemple, nous allons étudier la proportion de répondants jugés en mauvaise santé selon l'indice de la description de la santé. Lorsqu'un cycle est pris isolément, il arrive parfois que le nombre d'observations soit insuffisant pour produire des estimations de bonne qualité. Dans le cas de cet indice, pour le cycle 3.1 de l'enquête, 23 répondants ont été jugés en mauvaise santé, ce qui représente une estimation de 3,1 % de la population. Cette estimation était de piètre qualité avec un coefficient de variation de 39,26 %. Afin d'améliorer la qualité des estimations, on peut envisager de calculer une estimation combinée, puisqu'on peut considérer les cycles comme indépendants. Toutefois, comme nous l'avons souligné plus haut, les premières étapes d'une analyse consistent à déterminer la population d'intérêt ainsi que les caractéristiques à étudier. Ces étapes permettent de choisir la meilleure méthode de combinaison des données; faute d'en tenir compte, on risque de mal interpréter les résultats.

Dans ce scénario, il faut se demander comment composer avec le fait que le chercheur combine trois aperçus différents d'une population en évolution sur une période de cinq ans. En effet, chaque échantillon représente une population finie différente pour un moment différent. Il ne semble pas approprié de supposer que chaque échantillon est choisi à partir de la même population finie ni, par conséquent, d'utiliser une méthode de mise en commun. Il peut être avantageux de considérer la population finie de chaque cycle comme une strate tirée d'une population plus vaste, mais ce concept est assez compliqué. Dans ce scénario, le concept le plus simple est celui d'une moyenne mobile des estimations obtenues à partir des trois cycles. D'après les cycles antérieurs, on a estimé que 2,2 % et 1,6 % de la population étaient en mauvaise santé, ce qui donne une moyenne mobile de 2,3 %. On accepte des écarts

entre les valeurs puisqu'il n'y a pas d'hypothèse d'égalité entre les cycles. Quant à la variance, étant donné que les échantillons sont indépendants, on peut l'estimer comme suit :

$$\hat{V}\left[(\hat{p}_1 + \hat{p}_2 + \hat{p}_3)/3\right] = \left[\hat{V}(\hat{p}_1) + \hat{V}(\hat{p}_2) + \hat{V}(\hat{p}_3)\right]/9$$

Dans ce cas, le coefficient de variation combiné est de 19,6 %, ce qui correspond à une estimation publiable selon les normes de l'ESCC. Il serait peut-être plus efficicace de calculer une moyenne pondérée. On obtiendrait ainsi une combinaison linéaire d'estimations qui serait difficile à interpréter, sauf si l'on peut supposer que l'estimation de chaque échantillon est une estimation sans biais du même paramètre de population. Dans ce cas, on obtiendrait une estimation sans biais du même paramètre. Il faudrait recourir à des tests statistiques ou à des connaissances spécialisées pour vérifier cette hypothèse.

5.2 Exemple d'une population infinie

Le deuxième exemple est la combinaison de renseignements sur le temps d'attente pour des soins de santé, qu'on a recueillis à chaque cycle .1 auprès d'un sous-échantillon de près de 35 000 répondants à l'ESCC. Ce module de l'ESCC recueille suffisamment de renseignements pour permettre d'effectuer des analyses des soins de santé génériques à l'échelle du Canada. Toutefois, lorsqu'on veut effectuer des analyses de soins de santé plus détaillés, on ne dispose pas d'un nombre suffisant d'observations. Par exemple, on a repéré seulement dix personnes ayant connu des temps d'attente inacceptables pour une chirurgie cardiaque dans le cycle 3.1, et huit personnes dans le cycle 2.1. Chaque échantillon pris isolément ne renferme peut-être pas suffisamment de renseignements à analyser mais, étant donné que les échantillons sont indépendants, il est possible de combiner les cycles afin de disposer d'un plus vaste échantillon en vue d'une analyse.

Le principal problème lié à la combinaison des données réside dans la conception du questionnaire de chaque cycle de l'enquête. Les modifications les plus importantes ont été apportées entre les cycles 1.1 et 2.1 et nous estimons que la conception des questionnaires est si différente que les cycles ne mesurent pas nécessairement les mêmes rapports. Il est donc déconseillé de combiner le cycle 1.1 avec les autres cycles. Pour plus de renseignements sur les modifications, le chercheur peut consulter le guide d'utilisation de l'ESCC qui accompagne chaque cycle de l'enquête.

Avant de se demander comment combiner les données, le chercheur doit préciser le résultat escompté. Si l'on combine des ensembles de données, c'est habituellement pour être en mesure de mieux estimer les paramètres d'un modèle de régression. Dans ce cas, on peut considérer chaque échantillon comme généré indépendamment à partir du même modèle. La méthode de combinaison dite de mise en commun peut s'avérer préférable. En combinant les ensembles au niveau des microdonnées, on disposera d'un échantillon plus vaste pour estimer les paramètres de régression. Dans cet exemple, on arriverait presque à doubler la taille de l'échantillon disponible en combinant les deux cycles. On pourrait aussi utiliser la méthode individuelle. Toutefois, on obtiendrait alors des estimations différentes de celles de la méthode de mise en commun et il pourrait être difficile d'interpréter une combinaison linéaire de coefficients de régression.

Lorsqu'on estime un modèle par la méthode de mise en commun, on doit tenir compte de la présence d'échantillons provenant de différents cycles en incluant dans le modèle un facteur qui indique de quel cycle proviennent les données. On doit également tenir compte des interactions entre cette variable et les variables clés de l'analyse. En incluant ce terme dans le modèle avec les interactions, on tient compte dans le modèle d'aspects inconnus du plan, notamment de modifications apportées au mode de collecte.

Dans ce genre de situation, on se demande souvent s'il y a lieu d'utiliser des poids déterminés par le plan d'échantillonnage lorsqu'on estime les paramètres de régression du modèle. Il est recommandé d'utiliser les poids déterminés par le plan d'échantillonnage pour tenir compte du fait que les données sont générées par un plan d'enquête pouvant avoir un effet sur ce qui est mesuré. Dans ce cas, on pourrait comparer les estimations pondérées et non pondérées pour voir si le plan d'enquête a bel et bien un effet. Il est également suggéré d'utiliser les poids bootstrap, qui sont fournis, pour estimer les variances.

5.3 Facteurs à prendre en compte au moment de combiner des cycles de l'ESCC

Les exemples présentés plus haut témoignent de la possibilité de combiner des cycles de l'ESCC. Le chercheur doit avoir une idée précise de ce qu'il désire lorsqu'il combine les données de différents cycles et se rappeler que chaque échantillon est tiré d'une population observée différente. On peut surmonter cette contrainte si l'on peut considérer la population cible comme infinie mais, dans le cas d'une population finie, l'analyse doit tenir compte du fait que chaque population est différente. Un résultat combiné ne représente clairement aucune des populations finies observées dans un cycle donné. Si tel était le résultat désiré, on pourrait envisager les techniques d'estimation pour petits domaines ou des corrections de séries chronologiques. Le reste de la présente section porte en détail sur certains facteurs à prendre en compte avant de combiner des cycles de données de l'ESCC.

La première chose à prendre en compte est le questionnaire. On apporte souvent des modifications au questionnaire dans l'espoir d'améliorer la qualité des réponses aux questions. En général, les concepts restent inchangés malgré les modifications, mais le fait qu'une variable porte le même nom ne signifie pas nécessairement qu'on a posé la même question. Par exemple, dans le module de l'ESCC portant sur l'usage du tabac, la question du cycle 1.1 de l'enquête était « Avez-vous utilisé un timbre à la nicotine pour cesser de fumer? » alors que dans le cycle 2.1, on a modifié la question comme suit : « Au cours des 12 derniers mois, avez-vous utilisé un timbre à la nicotine pour cesser de fumer? » et le nom de la variable est resté le même.

Il importe également de s'assurer que les populations cibles de l'enquête sont les mêmes. Le module du cycle 1.1 de l'ESCC portant sur le comportement sexuel visait les personnes de 15 à 59 ans. Dans le cycle 2.1, on a ciblé plutôt les personnes de 15 à 49 ans pour réduire le fardeau de réponse des personnes âgées de 50 à 59 ans. De même, la plupart des renseignements sur les maladies transmises sexuellement ont été recueillis auprès des personnes appartenant au groupe d'âge de 15 à 49 ans.

Dans certaines analyses, il importe de s'assurer que les paramètres d'intérêt estimés d'après les échantillons combinés sont les mêmes. On peut effectuer des tests statistiques pour s'assurer qu'il n'y a pas d'écarts statistiques entre les éléments combinés provenant de chaque enquête. Toutefois, il est peu probable qu'on trouve des écarts statistiquement significatifs puisque les échantillons étaient de si petite taille qu'il a fallu les combiner. Si tel est le cas, il peut être préférable d'avoir recours à des connaissances spécialisées.

À chaque cycle .1 de l'ESCC, on a apporté des modifications aux limites géographiques des régions sociosanitaires. Ces modifications ont été tantôt légères, une partie d'une région sociosanitaire devenant une partie d'une autre région, tantôt plus importantes, les régions sociosanitaires définies par l'administration provinciale étant complètement transformées en fonction du mandat de cette administration en matière de santé. Si l'on a besoin d'un ensemble de données combinées pour établir des estimations infraprovinciales, il importe de prendre en compte toutes les modifications géographiques. Comme il est probable qu'on s'intéresse aux régions sociosanitaires définies le plus récemment, il serait souhaitable de recoder les limites géographiques des cycles précédents en fonction des nouvelles. La tâche n'est pas simple et Statistique Canada entend recoder prochainement les fichiers de données antérieures selon les nouvelles limites géographiques. Ces ensembles de données permettraient également d'établir des comparaisons entre les cycles où les limites des régions sociosanitaires ont été modifiées.

Un problème peut se poser pour certaines variables : le mode de collecte a un effet important sur les réponses obtenues. Dans le cas de l'ESCC, les modes de collecte possibles sont l'interview téléphonique et l'interview sur place et, en matière de santé, les gens répondent différemment selon le mode de collecte. Selon une étude menée en 2004 par St-Pierre et Béland, les caractéristiques les plus problématiques sont les données autodéclarées sur la taille et le poids, l'indice d'activité physique, la consultation de médecins et les besoins non satisfaits en matière de santé. Pour résoudre ce problème, on s'assure, dans toute la mesure du possible, que le mélange d'interviews téléphoniques et d'interviews sur place reste constant. Toutefois, vu l'importance de l'achat d'échantillons supplémentaires, habituellement recueillis par téléphone, la proportion d'interviews téléphoniques et d'interviews sur place peut différer d'un cycle à l'autre. Nous avons souligné dans la section 2.1 que la proportion d'interviews téléphoniques était très faible pour le cycle 1.1, ce qui crée un problème lorsqu'on veut combiner ce cycle avec les autres cycles. En raison de ces écarts, il peut être difficile d'interpréter le résultat combiné.

6 Conclusion

L'objet du présent document est de montrer qu'en effectuant une analyse attentive et détaillée, il est possible de combiner les cycles de l'ESCC, grâce à l'abondance de données semblables qui ont été recueillies. Toutefois, l'auteur tient à signaler au chercheur qu'il ne s'agit pas d'une tâche simple à entreprendre aveuglément. Afin de pouvoir combiner les cycles de l'ESCC, le chercheur doit avoir bien défini le résultat escompté et analysé avec soin les données provenant de l'ESCC pour s'assurer que ce résultat est réalisable. Il faudra perfectionner l'ESCC pour s'assurer que les données se prêtent à la combinaison. Cet aspect deviendra impératif lorsqu'on entreprendra la collecte continue avec le cycle 4.1 de l'enquête en partant du principe qu'on peut combiner des données recueillies sur n'importe quelle période pour créer des estimations de la population.

7 Remerciements

L'auteur tient à remercier tous les collègues qui l'ont aidé dans la rédaction du présent document, et particulièrement Georgia Roberts, Karla Fox et Michelle Simard pour leurs précieuses observations.

8 Références

Binder, D., et Roberts, G. (2006), "Issues Relating to Methods for Analysis of Survey Data", 2006 *Proceedings of the session on Complex Data Structures, Joint Statistical Meetings*, manuscrit soumis pour publication.

Canadian Institute for Health Information (1999), "Health Information Roadmap: Beginning the Journey" (1-895581-32-X).

Chu, A., Brick, J.M., et Kalton, G.(1999). "Weights for Combining Surveys across Time or Space". *Proceedings from the 1999 International Statistical Institute*, p. 103-104.

Kish, L. (1999). "Le cumul ou la combinaison d'enquêtes démographiques". *Techniques d'enquête, vol. 25,* p. 147 – 158.

Korn, E. L. et Graubard, B. I. (1998). *Analysis of Health Surveys.* Wiley.

St-Pierre, M. et Béland, Y. (2004). "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, Canada: American Statistical Association.