

ÉTUDES DE COUVERTURE DU RECENSEMENT CANADIEN DE 2006 : NOUVEAUTÉS

Christian Thibault, Josée Morel¹

RÉSUMÉ

Les erreurs de couverture constituent l'un des plus importants types d'erreur lors d'un recensement, étant donné qu'elles touchent non seulement la précision des chiffres des divers univers du recensement mais aussi la précision de toutes les données du recensement portant sur les caractéristiques de ces univers. Des changements majeurs seront apportés aux études de couverture du Recensement canadien de 2006. Ces changements sont rendus possibles grâce à la présence, pour la première fois, des noms et adresses sur la base de réponse du recensement pendant la période de traitement.

L'Étude par appariement automatisé de 2006 visera à détecter tout le surdénombrement dû à des personnes inscrites plus d'une fois sur la base de données du recensement alors qu'elle n'en mesurait que 50% en 2001. De plus l'Étude de la contre-vérification des dossiers reviendra à ses objectifs initiaux, c'est-à-dire l'estimation du sous-dénombrement, la composante du sur-dénombrement étant abandonnée au profit de l'Étude par appariement automatisé. Cette réorientation permettra des modifications aux procédures de cette étude et une diminution des ressources requises. La présente communication expose les changements d'orientation prévus aux études de couverture du Recensement canadien de 2006 et les améliorations qui en découleront.

1. INTRODUCTION

1.1 Description

Depuis 1961, Statistique Canada publie des estimations de l'erreur de couverture pour les recensements de la population tenus à tous les cinq ans. Les erreurs de couverture lors du recensement sont des erreurs qui ont une incidence sur l'exactitude des chiffres des divers univers du recensement. Il existe deux genres d'erreurs de couverture : les erreurs de sous-dénombrement et les erreurs de surdénombrement. Il y a sous-dénombrement lorsqu'on omet complètement de dénombrier une unité faisant partie d'un univers visé par le recensement. Par ailleurs le surdénombrement peut survenir de deux façons. La première, beaucoup plus fréquente et la plus facile à détecter, est lorsqu'une unité faisant partie d'un univers du recensement est dénombrée plus d'une fois. La seconde est lorsqu'une unité ne faisant pas partie de l'univers du recensement (par exemple un résident étranger, une personne fictive ou un logement marginal inoccupé) est dénombrée par erreur.

Depuis 1991, des estimations du surdénombrement et du sous-dénombrement net sont également produites. Ces estimations sont un intrant dans la détermination des estimations officielles de la population qui ont un impact direct sur le partage des transferts fiscaux du gouvernement fédéral aux provinces et territoires. Le taux de sous-dénombrement net au recensement de 2001 a été estimé à 2,99% au niveau canadien.

En 2001, trois études de couverture ont été menées. L'Étude par appariement automatisé visait à identifier les ménages dénombrés plus d'une fois sur la base de données du recensement, l'Étude sur les logements collectifs mesurait le surdénombrement entre les logements collectifs et les ménages privés et l'Étude de la contre-vérification des dossiers mesurait tout le sous-dénombrement et la portion du surdénombrement non mesurée par les deux autres études. L'Étude sur les logements collectifs sera abandonnée en 2006.

Cet article présente les changements de méthodologie proposés pour les études de couverture du recensement de 2006 au moment de son écriture. Des modifications pourraient être apportées d'ici leur mise en place. Une description des études de couverture du Recensement de 2001 peut être consultée

¹ Christian Thibault et Josée Morel, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

dans Morel et Thibault (2002) et dans le Rapport technique sur la couverture du Recensement de 2001 (voir Statistique Canada (2004)²).

2. SURDÉNOMBREMENT

2.1 Étude par appariement automatisé de 2001

Le surdénombrement est attribuable principalement aux cas de personnes qui figurent plus d'une fois dans la base de données du recensement. En 2001 le surdénombrement a été estimé par trois études : l'Étude par appariement automatisé (ÉAA) (49% de l'estimation), la Contre-vérification des dossiers (CVD) (50%) et l'Étude sur les logements collectifs (moins de 1%).

L'ÉAA de 2001 visait à repérer et à estimer le surdénombrement dans les logements privés en recherchant dans la base de données du recensement des paires de ménages résidant dans la même région géographique (Atlantique, Québec, Ontario, Ouest et Territoires) qui comprenaient des personnes de même sexe ayant la même date de naissance complète. Cette étude utilisait des programmes visant à repérer les paires de ménages résidant dans la même région et comptant au moins deux appariements exacts de personnes ainsi que les paires de ménages résidant dans la même circonscription électorale fédérale (CÉF) et comportant un seul appariement exact de personnes. Les paires de ménage identifiées constituaient la base de sondage de l'ÉAA.

Ces paires étaient par la suite stratifiées de façon à produire des strates qui étaient à la fois homogènes et de taille raisonnable en regroupant les paires de ménages présentant des probabilités similaires de surdénombrement.

Les questionnaires du recensement pour un échantillon de 17 000 paires de ménages ont été comparés pour déterminer s'il y avait eu surdénombrement. La liste de personnes inscrites sur le questionnaire d'un des ménages était ainsi comparée à la liste des personnes inscrites sur le second questionnaire. Lorsque les mêmes personnes figuraient sur les deux questionnaires on considérait qu'il s'agissait d'un cas de surdénombrement. Dans certains cas, lorsque le niveau de similarité entre les paires était très élevé indiquant ainsi une probabilité de surdénombrement essentiellement de 1, aucune vérification des questionnaires n'a été faite et les cas ont été classés automatiquement comme surdénombrement.

2.2 Étude par appariement automatisé de 2006

Le Recensement de 2006 innovera en procédant, pour la première fois, à la saisie optique des informations des questionnaires du recensement incluant les noms et prénoms et les adresses. Cette information sera conservée sur la base de données réponses (BDR). La BDR est la base de données initiale du recensement qui sera utilisée pendant les opérations de collecte et de traitement. Cette base contiendra l'information saisie par la lecture optique des questionnaires (et par les opérations de réparation de cette information) ainsi que l'information recueillie des personnes qui répondront directement par internet. On y retrouvera également toute l'information nécessaire pour les opérations de vérification des réponses, du contrôle de la couverture et du suivi. Cette base temporaire sera détruite lorsque les opérations du recensement seront complétées. L'information qu'elle contient sera toutefois conservée pour les besoins d'archivage uniquement. D'autre part la base de données du recensement sera construite à partir des données de la BDR mais elle ne contiendra que l'information nécessaire à la production des statistiques du recensement. Les noms, prénoms et adresses ne seront jamais copiés sur la base de données du recensement.

La disponibilité des noms, prénoms et adresses sous format électronique permet d'envisager une expansion majeure de l'ÉAA. La méthodologie proposée pour l'ÉAA de 2006 comprend deux grandes étapes. La première étape vise à appairer la BDR à un ensemble de fichiers administratifs fiables couvrant la plus grande proportion possible de la population cible du recensement. Des appariements exacts seront effectués à cette étape en utilisant les noms et prénoms, la date de naissance et le sexe ainsi que d'autres informations tel que l'état matrimonial et les coordonnées géographiques. Plusieurs options sont envisagées concernant le traitement des noms allant de l'utilisation complète des noms et prénoms à une codification de ceux-ci (par exemple en utilisant les codes NYSIS ou SOUNDEX³)

² Aussi disponible sur le site WEB du recensement de 2001 de Statistique Canada <http://www12.statcan.ca/francais/census01/home/Index.cfm>

³ Pour une description des codes NYSIS ou SOUNDEX voir Armstrong (2000)

incluant des options intermédiaires comme l'utilisation des premières lettres des noms et prénoms uniquement. Le but de cette première étape est d'identifier de façon unique les personnes dénombrées au recensement et de détecter de façon automatique des cas de surdénombrement lorsque deux personnes dénombrées pointent sur un enregistrement administratif unique. L'appariement à une source de données indépendante vise à éviter d'identifier comme surdénombrement des cas de personnes différentes authentiques ayant des noms, prénoms et date de naissance identiques. Ces cas sont évidemment plus probables dans le cas de prénoms et noms fréquents. Les recherches ont démontrées que de tels cas existent.

L'utilisation de méthodes d'appariement exact appliquées de façon stricte permet d'envisager de classer l'ensemble des cas identifiés lors de cette première étape sans vérification subséquente autre que des opérations de vérification de la qualité. Ainsi, il y aura économie de ressources à cette étape car aucune vérification manuelle ne sera requise. Il y aura également une amélioration de la précision car il n'y aura aucun échantillonnage et par conséquent aucune variance associée à l'estimation des cas de surdénombrement de cette première étape.

On pourra donc identifier plusieurs personnes dénombrées plus d'une fois dont les cas de deux questionnaires reçus et saisis pour le même ménage. Il restera par contre d'autres duplications non identifiées incluant les cas associés aux personnes n'apparaissant pas sur les fichiers administratifs. Une deuxième étape est alors prévue et elle consistera en un appariement probabiliste des personnes non appariées lors de la première étape à l'ensemble de la BDR. Ces appariements fourniront une base de sondage de cas potentiels de surdénombrement qui seront stratifiés puis échantillonnés. Certains cas de duplication pourront être classés automatiquement alors que les autres cas sélectionnés devront être confirmés par des vérifications manuelles avant de procéder à l'estimation du nombre de cas de surdénombrement de la population. Les méthodes pour déterminer si les paires identifiées correspondent effectivement à du surdénombrement ne sont pas encore définies.

L'objectif visé pour 2006 est d'utiliser l'ÉAA pour estimer tout le surdénombrement généré par la duplication de personnes sur la base de données du recensement. Il est prévu d'effectuer les appariements au niveau national sans contrainte régionale. Cette approche augmente toutefois la possibilité d'occurrence de personnes ayant les mêmes caractéristiques et elle renforce le besoin d'une source indépendante pour identifier ces cas. De plus l'ÉAA visera toutes les personnes dénombrées au recensement et non seulement les personnes vivant dans des logements privés comme c'était le cas en 2001. En conséquence l'Étude sur les logements collectifs et la composante du surdénombrement de la CVD seront abandonnées.

L'ÉAA de 2006 devrait permettre une amélioration marquée de la précision des estimations du surdénombrement. En 2001, la composante de l'estimation du surdénombrement provenant de la CVD était caractérisée par une variance élevée. L'abandon de cette composante et son remplacement par la nouvelle ÉAA, combiné au développement de l'appariement exact de l'étape 1 de l'ÉAA de 2006 auquel aucune variance n'est associée, permettra de produire des estimations de surdénombrement plus précises.

3. CONTRE-VÉRIFICATION DES DOSSIERS

3.1 Description générale

Après chaque recensement depuis 1966, la Contre-vérification des dossiers (CVD) a servi à évaluer le sous-dénombrement lors du recensement. Depuis 1991, les résultats de la CVD sont combinés à ceux des études de surdénombrement pour calculer le sous-dénombrement net. En 1996 et 2001 la CVD a également été utilisée pour estimer le surdénombrement mais, on l'a vu à la section précédente, cette fonction est appelée à disparaître en 2006.

L'objectif principal de la CVD de 2006 sera de procéder à l'estimation du sous-dénombrement au niveau national, dans les provinces et les territoires canadiens ainsi que pour certains sous-groupes importants de la population.

La population cible de la CVD est identique à celle du recensement. Ainsi, l'échantillon de la CVD se compose de personnes qui auraient dû être dénombrées lors du recensement et il est choisi à partir de sources indépendantes de ce même recensement.

Lors des dernières CVD, peu après la collecte du recensement, un certain nombre d'opérations de dépistage ont été entreprises pour contacter et interviewer les personnes choisies (PC) dans l'échantillon et déterminer l'adresse de leur résidence habituelle le jour du recensement, ainsi que d'autres adresses où elles auraient pu être dénombrées. Par la suite, des recherches étaient effectuées sur les questionnaires et la base de données du recensement pour déterminer si ces personnes avaient été dénombrées. Cette dernière étape est appelée «traitement».

3.2 Changement d'orientation

La disponibilité des noms sur la BDR et le retrait de la composante du surdénombrement de la CVD permettra de réorienter cette étude et de libérer des ressources permettant d'augmenter la taille effective de l'échantillon car il ne sera plus nécessaire de traiter toutes les adresses recueillies pour chaque PC. En effet, contrairement aux CVD de 1996 et de 2001, le traitement des adresses pourra cesser dès que la PC aura été identifiée comme dénombrée sur la base de données du recensement car le statut de la personne (dénombrée ou omise) aura alors été déterminé. Le traitement des adresses supplémentaires de cette personne en vue de mesurer le surdénombrement sera abandonné.

Une nouvelle approche sera adoptée pour le traitement des adresses. Lors des CVD précédentes, l'information disponible des adresses était utilisée dans le but d'identifier un ou des questionnaires pouvant correspondre à l'adresse traitée. Par la suite le ou les questionnaires correspondants étaient consulté(s) pour confirmer ou infirmer le dénombrement de la PC. Avant 2001, le questionnaire papier était consulté alors, qu'en 2001, les opérations ont été automatisées et une image électronique du questionnaire était consultée. En 2006, avec la disponibilité des noms et des informations démographiques et géographiques sous format électronique, l'accès aux questionnaires ne sera plus nécessaire dans la majorité des cas. Le travail de recherche s'effectuera directement à partir des données électroniques et de nombreux outils de recherche utilisant ces informations seront développés.

Le début du traitement sera devancé et commencera même avant la collecte. En effet, les bases de sondage et l'appariement avec des fichiers administratifs permettront d'accumuler de bonnes pistes où les PC auraient pu être dénombrées. Ces informations, combinées avec l'information des noms, dates de naissance, sexes, numéros de téléphone et codes postaux permettront d'identifier directement un bon nombre de PC sur la BDR. Des outils développés lors des CVD précédentes, tel que le méga-appariement, seront utilisés à cet effet. Le méga-appariement utilise les mêmes principes que l'ÉAA de 2001, c'est-à-dire des appariements au niveau des ménages. Dans le cas de la CVD, on apparie l'échantillon de la CVD à la Base de données du recensement. Il est également prévu de développer de nouveaux outils pour effectuer les recherches dans la BDR. Les noms et les autres variables disponibles permettront de trouver une personne dénombrée même s'il y a des erreurs dans l'épellation de son nom sur la base du recensement ou la base de sondage. Ces outils pourront n'utiliser qu'une partie de l'information du nom pour identifier un ou quelques dossiers potentiels qui devront alors être vérifiés manuellement. Il ne sera pas nécessaire d'envoyer à la collecte les PC ainsi trouvées dénombrées. En conséquence, de nombreuses ressources seront épargnées au niveau de la collecte et du traitement ce qui ouvre la porte à une augmentation potentielle de la taille des échantillons.

Les PC qui ne seront pas identifiées comme dénombrées lors de cette première étape seront envoyées à la collecte pour déterminer si elles appartenaient à la population cible du recensement de même que pour recueillir toutes les adresses possibles où elles auraient pu être dénombrées. Une fois la collecte effectuée, les PC seront retournées au traitement. Ces changements de procédures exigeront une synchronisation serrée entre les opérations de la collecte et du traitement.

3.3 Regard vers le futur

Le nouveau paradigme de l'information disponible sur la BDR permet d'envisager un autre changement de méthodologie pour la CVD de 2011. Un appariement exact entre les bases de sondage de la CVD et la base de réponse du recensement pourrait être effectué avant la sélection des échantillons permettant de classer une bonne proportion de la population cible comme dénombrée dès le départ. Les personnes non appariées seraient alors stratifiées et l'échantillon serait sélectionné parmi celles-ci. L'échantillon comporterait une plus grande proportion de personnes omises qui constituent le principal intérêt de l'étude ainsi que de personnes hors-cible (comme les personnes ayant émigrées et les personnes décédées). La précision des estimations de sous-dénombrement serait alors fortement améliorée. Cette approche ne sera pas mise en oeuvre pour la CVD de 2006 (du moins pour

la majorité des échantillons) car elle soulève des problèmes logistiques importants. En effet les activités suivantes devraient être effectuées dans un temps très court : les appariements entre les bases de sondage et la BDR, la stratification, la sélection et la préparation des échantillons. Toutefois les données recueillies par la CVD de 2006 permettront d'évaluer cette approche et de formuler des recommandations pour le Recensement de 2011.

4. Conclusion

La méthodologie des études de couverture du Recensement de 2006 sont présentement en cours de remaniement. La disponibilité des noms sur la BDR permettra une expansion importante de l'Étude par appariement automatisé. Le but recherché est d'augmenter suffisamment la couverture du surdénombrement de l'ÉAA pour permettre l'abandon des autres études mesurant le surdénombrement. Ceci permettra de réorienter la CVD vers ses objectifs originaux qui visaient l'estimation du sous-dénombrement. Il s'en suivra des changements dans les procédures de la CVD qui permettront une économie appréciable de ressources.

RÉFÉRENCES

- Armstrong, M., (2000). Survol des questions touchant l'utilisation d'identificateurs personnels, Statistique Canada, Numéro 85-602- XIF au catalogue, Ministre de l'Industrie, Ottawa.
- Morel, J. et Thibault, C., (2004). « *Étude de couverture pour le Recensement de 2001 de Statistique Canada.* », contribution, Échantillonnage et méthodes d'enquêtes, Dunod, Paris.
- Statistique Canada, (2004). «Couverture, Rapports techniques du recensement de 2001», Série des produits de référence, Numéro 92-370-XIF au catalogue, Ministre de l'Industrie, Ottawa.