N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2004 : Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

2004

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2004 de Statistique Canada
Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

AMÉLIORATION DE LA QUALITÉ DES ESTIMATIONS POUR UNE POPULATION À FAIBLE REVENU: UTILISATION D'UNE BASE DUALE À L'ENQUÊTE SUR LES DÉPENSES DES MÉNAGES

Bruno Lapierre, Christian Nadeau, Johanne Tremblay et José Gaudet[1]

RÉSUMÉ

L'Enquête sur les dépenses des ménages de Statistique Canada vise principalement à fournir des estimations provinciales fiables des dépenses des ménages. L'échantillon stratifié à plusieurs degrés est sélectionné à partir d'une base aréolaire couvrant l'ensemble de la population et les données sont recueillies à l'aide d'une entrevue personnelle. Lors de l'enquête de 2003, un objectif supplémentaire consistait à améliorer la qualité des estimations pour une sous-population de ménages à faible revenu représentant environ 2,5% de l'ensemble des ménages de la province de Québec. Afin de rencontrer cet objectif, un échantillon supplémentaire a été sélectionné à partir d'une liste de logements situés dans un nombre limité d'aires géographiques préalablement identifiées à l'aide de données auxiliaires. Nous présentons ici le plan de sondage à base duale utilisé de même que différents scénarios envisagés et certains résultats ayant mené aux choix effectués.

MOTS CLÉS: Population rare, répartition optimale, stratification.

1. INTRODUCTION

Dans le cadre du développement d'une stratégie de lutte contre la pauvreté, le Ministère des finances du Québec (MFQ) désire refaire une étude réalisée par le Ministère de la Main d'œuvre et de la Sécurité du Revenu qui visait à établir des seuils de revenu minimum pour le Québec (Fugère et Lanctôt, 1985). Le modèle adopté dans cette étude repose en grande partie sur des estimations de moyennes de dépenses pour une sous-population de ménages à faible revenu ne représentant que 2,5% de l'ensemble des ménages du Québec. Afin de refaire cette étude à partir d'estimations de dépenses à jour plus précises que celles disponibles, le MFQ a financé l'ajout d'un échantillon supplémentaire à l'Enquête sur les dépenses des ménages (EDM) de Statistique Canada (SC).

L'échantillon régulier de l'EDM est sélectionné à partir d'une base aréolaire selon un plan stratifié à plusieurs degrés et sa taille est d'environ 3 000 logements au Québec (Arsenault et Tremblay, 2001). Selon ce plan, on estime à plus de 4 000 logements la taille de l'échantillon supplémentaire nécessaire à l'atteinte de l'objectif du MFQ. Un plan de sondage permettant de mieux cibler la population d'intérêt définie par le MFQ a été mis sur pied afin de réduire la taille d'échantillon totale requise.

Ce document présente les principaux éléments du plan de sondage élaboré pour répondre aux besoins du MFQ de même que certains résultats ayant mené aux choix effectués. Il débute par une brève mise en contexte et une description de l'objectif du MFQ. À la section 3, on discute de certaines notions relatives à l'efficacité d'un plan de sondage lorsqu'on désire estimer des moyennes pour une population rare. Divers résultats et analyses ayant mené au choix d'une base de sondage sont présentés à la section 4. À la section 5, on présente et justifie le choix des méthodes de stratification de la base de sondage, de répartition de l'échantillon et de sélection des logements. On conclut à la section 6 avec une description sommaire du plan de sondage adopté.

[1]Bruno Lapierre, Statistique Canada, Édifice principal, pièce 2500, Ottawa, Canada, K1A 0T6, Bruno.Lapierre@Statcan.ca; Christian Nadeau, Statistique Canada, Édifice R.H. Coats, 16ième étage, Ottawa, Canada, K1A 0T6, Christian.Nadeau@Statcan.ca; Johanne Tremblay, Statistique Canada, Édifice R.H. Coats, 16ième étage, Ottawa, Canada, K1A 0T6; José Gaudet, Statistique Canada, Édifice R.H. Coats, 11ième étage, Ottawa, Canada, K1A 0T6.

2. MISE EN CONTEXTE ET OBJECTIF

L'EDM est une enquête annuelle qui vise à recueillir des données détaillées sur les dépenses des ménages pour l'année précédant la période de collecte au moyen d'entrevues personnelles d'une durée moyenne de 110 minutes. Cette enquête impose un lourd fardeau aux répondants et l'utilisation d'interviewers expérimentés contribue à contrôler l'impact des erreurs non dues à l'échantillonnage sur la qualité des données.

L'objectif du MFQ consiste à améliorer la précision des estimations de dix variables de dépenses[2] pour les ménages du premier décile de revenu parmi les ménages d'un seul gagne-pain pour lesquels au moins 50% des revenus proviennent de la rémunération. Ceux-ci ne représentent que 2,5% des ménages du Québec selon le Recensement de 2001. Les coefficients de variation (CV) de ces estimations, obtenus lors de l'EDM 2000 et de l'EDM 2001, sont présentés au tableau 2.1. Selon les critères de qualité établis dans l'entente entre SC et le MFQ, on vise à obtenir des CV inférieurs à 15% pour les estimations d'au moins sept des dix variables. On constate que les CV obtenus à partir de l'échantillon régulier de l'EDM sont clairement supérieurs à 15% pour quatre des dix variables et qu'ils ne sont nettement inférieurs au seuil de 15% que pour deux variables.

Tableau 2.1 : CV des estimations de moyennes pour la population d'intérêt, EDM 2000 et 2001

	EDM 2000	EDM 2001
Alimentation	8,4%	6,6%
Ameublement	24,3%	24,0%
Communications	10,7%	13,9%
Entretien ménager	12,2%	14,5%
Habillement	13,9%	14,8%
Lecture	21,5%	33,4%
Logement	5,4%	4,4%
Loisirs	23,4%	17,9%
Soins personnels	15,0%	14,7%
Transport	31,8%	20,9%

On désire définir un plan de sondage afin de sélectionner un échantillon supplémentaire permettant l'atteinte de l'objectif du MFQ. On juge important de contrôler la taille de l'échantillon supplémentaire afin de limiter l'embauche de nouveaux interviewers et la prolongation de la période de collecte. On souhaite ainsi minimiser les risques d'une réduction de la qualité des données recueillies et de retard par rapport à l'échéancier de diffusion. Dans ce contexte, l'entente entre le MFQ et SC stipule que la taille de l'échantillon supplémentaire doit se situer entre 2 000 et 3 000 logements. Elle mentionne également que les plans de sondage permettant de répondre aux critères de qualité établis à partir d'un échantillon de 2 000 logements doivent être privilégiés. L'option d'augmenter la taille de l'échantillon régulier de l'EDM sans apporter de modifications au plan de sondage a été rejetée dès le départ puisque la taille d'échantillon supplémentaire requise a été évaluée à plus de 4 000 logements.

3. EFFICACITÉ D'UN PLAN DE SONDAGE POUR L'ESTIMATION D'UNE MOYENNE POUR UNE POPULATION RARE

Afin de mettre sur pied un plan de sondage qui permet d'estimer efficacement des moyennes pour une population rare, on désire augmenter la fraction de sondage là où cette population rare est plus concentrée. Kalton (2001) traite du cas de la répartition d'un échantillon entre deux strates lorsqu'on désire estimer une moyenne pour une population rare sous l'hypothèse que la moyenne et la variance de la variable d'intérêt pour la population rare sont les mêmes dans les deux strates. On définit d'abord la prévalence par $P = M/N$, où M et N représentent

[2]Les définitions des variables de dépenses utilisées pour le MFQ sont définies dans Nadeau et coll. (2005) et peuvent différer des définitions en cours à l'EDM.

respectivement la taille de la population rare et la taille de la population. De façon similaire, on définit la prévalence dans la strate h par $P_h = M_h / N_h$ et la couverture de la population rare par la strate h par $A_h = M_h / M$. Lorsque l'échantillonnage aléatoire simple est utilisé dans chacune des strates, la répartition optimale correspond à une répartition de l'échantillon proportionnelle à $N_h \sqrt{P_h}$. Kalton (2001) démontre que, pour que la répartition optimale procure une réduction de la variance appréciable par rapport à une répartition proportionnelle à la taille, la strate présentant la plus forte prévalence (h=1) doit être telle que les valeurs de P_1 / P et A_1 soient relativement élevées. Autrement dit, pour que le gain en efficacité soit appréciable, une telle strate doit présenter une prévalence nettement plus forte que celle observée dans la population tout en couvrant une grande partie de la population rare. Ces principes sont utilisés lors des choix de la base de sondage, des méthodes de stratification et de répartition de l'échantillon présentés dans les sections suivantes.

4. CHOIX DE LA BASE DE SONDAGE

La première étape de l'élaboration du plan de sondage consiste à identifier la base de sondage à partir de laquelle l'échantillon supplémentaire sera sélectionné. Les options de base de sondage considérées afin d'augmenter la taille de l'échantillon de l'EDM au Québec peuvent être regroupées en deux grandes catégories.

Une première catégorie d'options consiste à sélectionner l'échantillon supplémentaire à partir de la base aréolaire qui sert à la sélection de l'échantillon régulier de l'EDM. Parmi cette catégorie, une approche souvent utilisée à SC consiste à sélectionner des ménages ayant déjà participé à d'autres enquêtes dont l'échantillon est choisi de la même base de sondage. Une telle option permet de bénéficier d'information déjà disponible sur les répondants et ainsi de mieux cibler les ménages faisant partie de la population cible. Considérant le fardeau de réponse déjà lourd de l'EDM, cette option a été éliminée dès le départ. Une autre approche envisagée consistait à sélectionner l'échantillon supplémentaire à partir de la base aréolaire en adoptant un plan d'échantillonnage différent de celui en vigueur pour l'échantillon régulier. Cette approche a été rejetée en raison de la difficulté à assigner de l'information auxiliaire à jour et pertinente à des unités géographiques de la base aréolaire suffisamment petites pour cibler efficacement la population visée (Nadeau et coll., 2005).

L'autre catégorie d'options consiste à sélectionner l'échantillon de l'EDM à partir d'une base duale, c'est-à-dire de n'utiliser la base aréolaire que pour la sélection de l'échantillon régulier de l'EDM et d'identifier une autre base ne couvrant qu'une partie de la population du Québec pour la sélection de l'échantillon supplémentaire. Cette deuxième base de sondage doit être conçue de façon à respecter le principe énoncé à la section précédente, c'est-à-dire qu'elle devra présenter à la fois une prévalence et une couverture de la population d'intérêt du MFQ suffisamment grandes pour que le gain en efficacité soit appréciable.

Afin de limiter les coûts et la complexité des opérations et d'assurer une certaine qualité des adresses sur les bases de sondage considérées pour la sélection de l'échantillon supplémentaire, celles-ci correspondent toutes à des listes de logements extraites du Registre des adresses (RA) de Statistique Canada (Swain et coll., 1992). Le RA est tenu à jour par SC dans les aires géographiques à plus forte densité de population. La version du RA utilisée est celle résultant de la mise à jour qui a suivi le Recensement de 2001 et couvre 86% de la population du Québec.

Les principales options considérées sont d'abord présentées en 4.1 et 4.2 suivies d'une évaluation de la précision des estimations pour des échantillons tirés de ces bases de sondage en 4.3.

4.1 Approche post-censitaire

Les enquêtes post-censitaires sont habituellement des enquêtes identifiées et planifiées avant le Recensement. Le questionnaire du Recensement est utilisé comme première étape d'un plan de sondage à deux phases afin de mieux cibler la population d'intérêt. Des ménages ou encore des personnes possédant certaines caractéristiques forment la base de sondage et un échantillon est sélectionné à partir de celle-ci. L'approche étudiée diffère des enquêtes post-censitaires traditionnelles en ce sens que l'échantillon de deuxième phase est composé de logements et non de ménages et ce afin d'éviter les opérations de dépistage nécessaires pour retracer les ménages ou les membres de

ménages ayant déménagé depuis le Recensement de 2001. Pour la première fois lors du Recensement de 2001, le questionnaire faisait mention de la possibilité d'utiliser les données du Recensement à des fins de sélection d'échantillon pour d'autres enquêtes de SC. Une telle possibilité a été exploitée récemment dans une enquête de SC (Duggan, Neusy et Bélanger, 2003) et peut donc être envisagée dans le cas de la sélection d'un échantillon à l'EDM.

On désire créer une liste de logements ayant de fortes chances d'abriter des ménages de la population d'intérêt à la collecte de l'EDM 2003, en utilisant les données des ménages habitant ces logements lors du Recensement de 2001. Afin d'identifier les caractéristiques des ménages permettant le mieux d'identifier les logements pouvant abriter des ménages de la population d'intérêt trois années plus tard, des tableaux croisant des caractéristiques des ménages à l'année "0" et le fait d'appartenir ou non à la population d'intérêt à l'année "3" ont été construits à partir des données de l'Enquête sur la dynamique du travail et du revenu (EDTR) (se référer à Lavigne et Michaud, 1998 pour plus de détails sur l'EDTR)[3]. Les résultats observés ont mené à l'étude d'une base-liste formée des logements abritant un ménage de la « population d'intérêt élargie »[4] lors du Recensement de 2001 pour l'approche post-censitaire. En appliquant aux données du Recensement de 2001 les proportions appropriées observées dans les tableaux croisés faits à partir des données de l'EDTR, on estime que la prévalence à l'intérieur de cette base serait de 8,3% et la couverture de la population d'intérêt serait de 3,0% pour l'année de référence 2003.

4.2 Ciblage d'aires géographiques à forte concentration

Les ménages de la population d'intérêt possèdent certaines caractéristiques communes qui portent à croire qu'ils habiteront en plus forte concentration certains secteurs géographiques. Parmi ces caractéristiques, notons leur faible revenu et le fait d'être locataire. On désire créer une base-liste formée des ménages qui se trouvent dans des secteurs géographiques à forte concentration de ménages de la population d'intérêt. L'identification de telles aires géographiques peut être effectuée à l'aide des données du Recensement de 2001. Celles-ci contiennent de l'information sur le revenu, le travail, les caractéristiques sociodémographiques des ménages et la géographie. Les données sur le revenu du Recensement réfèrent à l'année 2000 et sont disponibles pour un ménage sur cinq. Les unités géographiques choisies doivent être suffisamment petites pour permettent de cibler avec précision les secteurs où se concentrent les membres de la population d'intérêt. Elles doivent également être suffisamment grandes pour assurer une certaine fiabilité de la prévalence estimée.

Trois bases-listes à concentration géographique sont considérées ici et leurs caractéristiques sont présentées au tableau 4.1. Elles sont toutes trois formées des logements situés dans les aires de diffusion (AD) présentant les plus fortes prévalences parmi celles couvertes par le RA selon les données du Recensement de 2001. Une AD est une petite unité géographique relativement stable formée d'un ou de plusieurs îlots avoisinants et regroupant de 400 à 700 habitants (environ 250 ménages en moyenne). Il s'agit de la plus petite région géographique normalisée pour laquelle toutes les données du Recensement sont diffusées. L'ensemble du territoire du Québec est divisé en 12 153 AD.

[3]L'EDTR permet de recueillir des données sur le travail et le revenu pour les mêmes personnes sur une période de six années consécutives. Il a donc été possible d'analyser les caractéristiques des ménages sur une période de 3 ans. Les données utilisées ici portent sur la période de 1997 à 2000.

[4]Tel que vu à la section 2, la population d'intérêt est constituée des 10% de ménages ayant le plus bas revenu dans la sous-population des ménages d'un seul gagne-pain et dont au moins 50% des revenus avant impôts proviennent de la rémunération. On entend par « population d'intérêt élargie » les 20% de ménages ayant le plus bas revenu dans la même sous-population.

Tableau 4.1 : Concentration géographique : Caractéristiques des bases-listes à l'étude

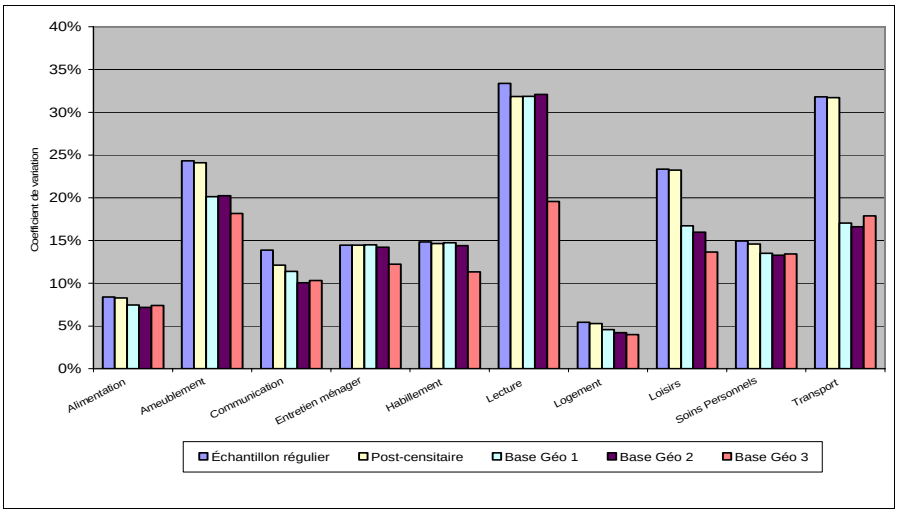
Options de bases-listes à concentration géographique	Nombre de AD sur la base-liste	Prévalence de la base-liste	Couverture de la population d'intérêt	Couverture de la population totale
Base géographique 1	1 083	9,5%	38%	10%
Base géographique 2	2 125	7,3%	59%	20%
Base géographique 3	5 060	4,4%	88%	50%

Ces bases diffèrent entre elles en terme de la proportion de la population totale qu'elles couvrent. L'augmentation de la couverture de la population s'obtient par l'ajout d'AD avec des concentrations de plus en plus faibles. Elle résulte en une diminution de la prévalence globale sur la base et en une augmentation de la couverture de la population d'intérêt. On note également que, de par leur construction, la base géographique 1 est incluse dans la base géographique 2 qui à son tour est incluse dans la base géographique 3. Pour choisir une base efficace, on cherche donc le meilleur compromis entre la prévalence et la couverture de la population d'intérêt.

4.3 Évaluation des options et choix de la base-liste

Afin d'évaluer les quatre options de bases duales retenues, soient l'approche post-censitaire et les trois bases géographiques, on calcule des CV attendus pour les estimations des moyennes des dix variables de dépenses pour la population d'intérêt. Les CV calculés en combinant l'échantillon régulier de l'EDM et un échantillon supplémentaire de 2 000 logements sont présentés à la figure 4.1. Les CV des estimations obtenus à partir de l'échantillon régulier de l'EDM uniquement sont également présentés sur ce graphique à titre comparatif. La méthodologie utilisée pour le calcul des CV attendus est décrite en annexe. Des détails additionnels sont fournis dans Nadeau et coll. (2005).

Figure 4.1 : Coefficients de variation attendus pour les estimations de moyennes de dépenses pour la population d'intérêt selon différentes options de bases-listes



On constate d'abord que l'ajout d'un échantillon de 2 000 logements en utilisant l'option post-censitaire ne permet qu'une très faible amélioration des CV par rapport à l'échantillon régulier. La base géographique 3, couvrant la plus grande proportion de la population d'intérêt, procure les meilleurs CV pour la majorité des variables. Elle est particulièrement efficace pour les variables ayant des CV élevés avec l'échantillon régulier. En terme de CV, une bonne couverture de la population d'intérêt par la base semble donc plus importante qu'une meilleure prévalence. On constate que l'utilisation de la base géographique 3 est la seule option qui permet de satisfaire les critères de

qualité établis dans l'entente entre SC et le MFQ, c'est-à-dire des CV inférieurs à 15% pour les estimations d'au moins sept des dix variables. La base-liste formée des logements des AD ayant les plus fortes prévalences et couvrant 50% de l'ensemble des ménages du Québec est donc celle retenue pour sélectionner l'échantillon supplémentaire.

5. CHOIX DES MÉTHODES DE STRATIFICATION DE LA BASE DE SONDAGE, DE RÉPARTITION ET DE SÉLECTION DE L'ÉCHANTILLON

Les résultats présentés jusqu'à maintenant ont mené au choix d'une base de sondage pour la sélection de l'échantillon supplémentaire. La seconde étape consiste à déterminer les méthodes de stratification de la base, de répartition de l'échantillon et de sélection des logements qui permettent de réduire les CV pour les estimations de moyennes pour la population d'intérêt.

5.1 Stratification de la base de sondage

Bien que l'objectif principal soit d'améliorer la qualité des estimations pour la population d'intérêt à l'échelle provinciale, la possibilité d'utiliser les données à un autre niveau lors de l'analyse a été évoquée par le MFQ en cours de projet. Pour ce faire, l'ensemble du Québec a été divisé en trois strates, soit la région métropolitaine de recensement (RMR) de Montréal, la RMR de Québec, et le reste du Québec. On désirait ainsi s'assurer d'obtenir une taille d'échantillon minimale dans chacune de ces régions. Comme on le verra plus loin, le scénario de répartition retenu ne tient finalement pas compte d'un tel objectif.

Un deuxième niveau de stratification est créé afin de former des strates homogènes quant aux moyennes des dix variables de dépenses dans la population d'intérêt. L'étude des données des EDM de 1997 à 2001 démontre que les moyennes de dépenses en *Transport* diffèrent sensiblement en fonction de la géographie. Par conséquent, les trois strates de premier niveau ont été divisées en douze strates de deuxième niveau. Pour les RMR de Montréal et de Québec, les régions économiques de recensement ont été utilisées comme deuxième niveau de stratification. Pour le reste du Québec, les quatre autres RMR forment chacune une strate distincte, une strate est formée en regroupant toutes les agglomérations de recensement qui ne sont pas des RMR, et une autre en regroupant la portion du Québec hors agglomérations de recensement.

Afin de tirer profit de la répartition optimale présentée à la section 3, un troisième niveau de stratification est introduit. On désire diviser chacune des douze strates de deuxième niveau en deux de façon à augmenter les gains d'efficacité obtenus à l'aide de cette méthode de répartition. Pour ce faire, on partitionne les AD de façon à former deux strates, la première étant composée des AD présentant les plus fortes prévalences et la deuxième étant formée des AD présentant les plus faibles prévalences. Sous les hypothèses présentées à la section 3, Kalton et Anderson (1986) montrent que la variance de l'estimateur de la moyenne est approximativement proportionnelle à $\left(\frac{A_1^{\,2}}{M_1\sqrt{P_1}} + \frac{A_2^{\,2}}{M_2\sqrt{P_2}} \right)\left(N_1\sqrt{P_1} + N_2\sqrt{P_2} \right)$. En utilisant l'information sur les AD quant au nombre de ménages dans la population totale et dans la population d'intérêt lors du Recensement de 2001, chaque strate de deuxième niveau est partitionnée en deux ensembles d'AD afin de minimiser cette quantité et améliorer le gain d'efficacité. On obtient ainsi vingt-quatre strates finales. L'étude de Monte Carlo décrite à la section 5.3 confirme que l'utilisation du troisième niveau de stratification permet une réduction de la variance par rapport au plan à deux niveaux de stratification.

5.2 Répartition de l'échantillon

L'échantillon supplémentaire est d'abord réparti entre les strates du premier niveau proportionnellement au nombre de logements dans chacune des trois strates. Cette option représente un cas particulier d'une famille de méthodes de répartition de puissance évaluées afin d'identifier un compromis visant à satisfaire des objectifs provincial et infra-provincial. Bien que l'objectif infra-provincial ait été abandonné en cours de route, il était avantageux, d'un point de vue opérationnel, d'utiliser une des méthodes évaluées pour la suite des travaux. De plus, étant donné des

prévalences très similaires pour les trois strates de premier niveau, une répartition optimale aurait été très similaire à celle obtenue ainsi.

L'échantillon de chaque strate de premier niveau est réparti entre les strates de deuxième niveau selon la méthode de répartition optimale présentée à la section 3, où la taille de l'échantillon attribuée à chaque strate est proportionnelle à $N_h\sqrt{P_h}$. On alloue ainsi une plus grande proportion de l'échantillon aux strates présentant les plus hautes prévalences que ne l'aurait fait une répartition proportionnelle à la taille. La répartition optimale est également utilisée pour répartir l'échantillon attribué aux strates de deuxième niveau entre celles du troisième niveau puisque celles-ci ont été construites spécifiquement pour que ce type de répartition soit efficace.

5.3 Sélection des logements

La méthode d'échantillonnage à deux degrés est utilisée pour la sélection des logements à l'intérieur de chacune des strates afin d'éviter une trop grande dispersion de l'échantillon. Les unités primaires d'échantillonnage (UPÉ) correspondent aux AD et sont sélectionnées avec une probabilité proportionnelle à leur taille, c'est-à-dire au nombre de logements sur la base-liste. Un certain nombre de logements sont ensuite sélectionnés à l'intérieur de chacune des AD choisies à l'aide d'un échantillonnage systématique. Les logements sont préalablement ordonnés selon une séquence assurant une certaine proximité géographique aux logements consécutifs sur le fichier.

Une étude de Monte Carlo a été effectuée pour mesurer l'efficacité de deux options pour sélectionner les logements au deuxième degré. Une première option consiste à choisir cinq logements dans chacune des AD sélectionnées, soit le rendement moyen des UPÉ dans le plan de sondage de l'échantillon régulier de l'EDM. La deuxième option consiste à choisir trois logements par AD dans les vingt strates relatives à des RMR et cinq logements par AD dans les quatre autres strates. Cette option permettrait de sélectionner davantage d'UPÉ dans les RMR, tout en limitant la dispersion de l'échantillon dans les régions urbaines de petite taille et dans les régions rurales.

La simulation a été effectuée à partir des données du Recensement de 2001. Elle consiste à sélectionner 1 000 échantillons puis à calculer la variance des estimations pour les ménages de la population d'intérêt, et ce pour les deux plans d'échantillonnage évalués. Le Recensement ne comportant pas de variables de dépenses autre que la variable *coût du loyer*, trois variables de revenu ont également été utilisées afin de comparer la variabilité des estimations obtenues selon les différents plans. Tel que mentionné à la section 5.1, un plan limité aux deux premiers niveaux de stratification a aussi été étudié, de sorte à évaluer l'efficacité du troisième niveau de stratification. Finalement, la variance selon un plan d'échantillonnage aléatoire simple a été calculée pour permettre de dériver les effets de plan des différentes options.

Contrairement à ce qui était attendu, l'étude de Monte Carlo montre que la sélection de trois logements pour les AD des strates à l'intérieur des RMR (et donc l'augmentation du nombre d'UPÉ sélectionnées) a peu d'effet sur la qualité des estimations des variables étudiées pour la population d'intérêt. Malgré que les résultats pour les deux options soient très similaires et puisque la sélection d'un plus grand nombre d'UPÉ entraîne généralement une réduction de la variance, la sélection de trois logements par UPÉ a été retenue.

Cette simulation indique que l'effet de plan obtenu lorsque ce plan est utilisé est inférieur à l'effet de plan observé lorsque l'échantillon régulier est utilisé. L'impact de ce gain en efficacité par rapport aux résultats présentés en 4.3 est présenté dans Nadeau et coll. (2005).

La taille de l'échantillon supplémentaire a finalement été fixée à 2 211 logements suite à une réévaluation des contraintes liées à la collecte une fois sa répartition géographique connue. Un appariement entre cet échantillon et l'échantillon régulier de l'EDM a permis de minimiser les risques de chevauchement de ceux-ci.

6. CONCLUSION

Le MFQ désire améliorer la précision des estimations de dépenses moyennes provenant de l'EDM pour les ménages du premier décile de revenu parmi les ménages d'un seul gagne-pain pour lesquels au moins 50% des revenus proviennent de la rémunération. Une augmentation de la taille de l'échantillon régulier de l'EDM de plus de 4 000 logements est jugée nécessaire à l'atteinte de l'objectif du MFQ si aucune modification n'est apportée au plan de sondage de l'EDM. Un plan de sondage visant à mieux cibler la population d'intérêt a été élaboré de sorte à réduire la taille de l'échantillon supplémentaire requise pour atteindre cet objectif.

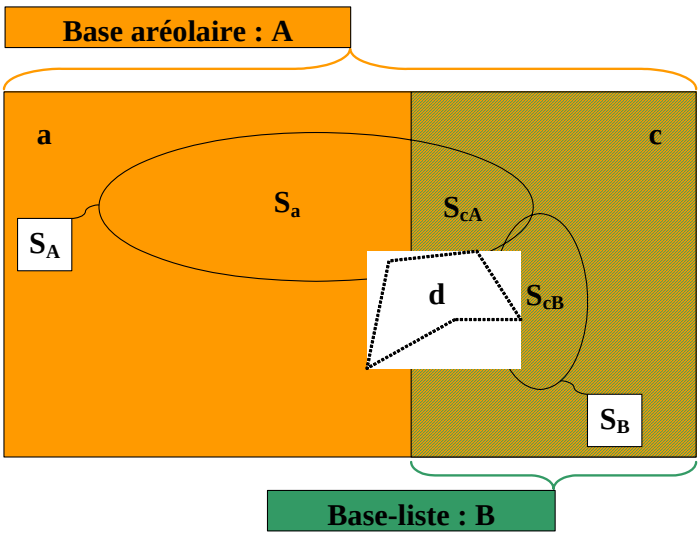
L'échantillon supplémentaire du MFQ a été sélectionné d'une liste de logements selon un plan stratifié à deux degrés d'échantillonnage. La base-liste utilisée est constituée des logements inscrits au RA dans les AD présentant les plus fortes prévalences lors du Recensement de 2001 de sorte à couvrir 50% des ménages du Québec. Elle est stratifiée selon trois niveaux, les deux premiers niveaux étant de nature géographique alors que le troisième permet de tenir compte de la concentration des ménages de la population d'intérêt. La répartition de l'échantillon entre les strates du premier niveau est proportionnelle au nombre de logements dans ces strates alors que la répartition pour les niveaux de stratification subséquents tient compte du nombre de logements et de la concentration des ménages de la population d'intérêt, et ce de façon à réduire au maximum la variance. À l'intérieur de chaque strate, un nombre déterminé d'AD ont été sélectionnées avec une probabilité proportionnelle au nombre de logements. Dans chacune des AD sélectionnées, on utilise l'échantillonnage systématique pour sélectionner trois logements dans les RMR et cinq ailleurs. La taille de l'échantillon supplémentaire a été déterminée de sorte à satisfaire les critères de qualité établis dans l'entente entre SC et le MFQ tout en minimisant l'impact sur l'EDM. Celle-ci a été fixée à 2 211 logements.

L'efficacité réelle du plan adopté est présentement évaluée et les résultats pourront être présentés suite à la diffusion de l'EDM 2003.

ANNEXE – MÉTHODOLOGIE DE CALCUL DES COEFFICIENTS DE VARIATION ATTENDUS

On désire évaluer et comparer la précision des estimations de moyenne pour le domaine d'intérêt obtenues en sélectionnant l'échantillon de l'EDM à partir de chacune des quatre bases duales présentées à la section 4. La figure A.1 illustre l'utilisation de ces bases duales.

Figure A.1 : Base duale constituée d'une base aréolaire couvrant l'ensemble de la population et d'une base-liste ne couvrant qu'une partie de la population



La base aréolaire **A** de laquelle est tiré l'échantillon régulier $\mathbf{S_A}$ couvre l'ensemble de la population du Québec. Une base-liste **B** qui ne couvre qu'une partie de la population du Québec est utilisée pour tirer l'échantillon supplémentaire $\mathbf{S_B}$. On peut subdiviser la base aréolaire en deux parties, celle non couverte par la base-liste ($\mathbf{a=A\cap B^c}$) et celle couverte par la base-liste ($\mathbf{c=A\cap B}$) et on subdivise de la même façon l'échantillon $\mathbf{S_A}$ pour obtenir $\mathbf{S_a=S_A\cap a}$ et $\mathbf{S_{cA}=S_A\cap c}$. Notons que puisque $\mathbf{B\subseteq A}$, alors $\mathbf{c=B}$ et $\mathbf{S_{cB}=S_B}$. Le domaine défini par la population d'intérêt du MFQ sera noté par **d**.

La notation utilisée pour les estimateurs et les paramètres de population est telle qu'un premier indice est utilisé pour désigner la population ou le domaine selon le cas. Dans le cas des estimateurs, un deuxième indice désigne la base de sondage utilisée alors qu'un exposant est utilisé pour indiquer le type d'estimateur lorsque nécessaire (« r » pour un estimateur par régression et « π » pour un estimateur d'Horvitz-Thompson). De façon générale, $\hat{\bar{Y}}$ représente un estimateur de la moyenne d'une population ou d'un domaine $\bar{Y}$, $\hat{Y}$ un estimateur de total alors que $\hat{N}$ représente un estimateur de N, la taille d'une population ou d'un domaine selon le cas.

On suppose d'abord l'utilisation de l'estimateur $\hat{\bar{Y}}_{d,\,AB}=\hat{Y}_{d,\,AB}\big/\hat{N}_{d,\,AB}$ où $\hat{Y}_{d,\,AB}=\hat{Y}^{r}_{d,\,A}-\alpha\left(\hat{Y}^{r}_{d\cap c,\,A}-\hat{Y}^{\pi}_{d\cap c,\,B}\right)$, $\hat{N}_{d,\,AB}=\hat{N}^{r}_{d,\,A}-\alpha\left(\hat{N}^{r}_{d\cap c,\,A}-\hat{N}^{\pi}_{d\cap c,\,B}\right)$, et où $0<\alpha<1$. En utilisant la linéarisation de Taylor, on obtient la formule de variance approximative suivante:

$$V(\hat{\bar{Y}}_{d,\,AB})\approx\frac{1}{N_d{}^2}\left[V(\hat{Y}_{d,\,AB})+\bar{Y}_d{}^2V(\hat{N}_{d,\,AB})-2\bar{Y}_d\,C(\hat{Y}_{d,\,AB},\hat{N}_{d,\,AB})\right] \tag{A.1}$$

où $V(\hat{Y}_{d,\,AB})=V(\hat{Y}^{r}_{d,\,A})+\alpha^2(V(\hat{Y}^{r}_{d\cap c,\,A})+V(\hat{Y}^{\pi}_{d\cap c,\,B}))-2\,\alpha\,C(\hat{Y}^{r}_{d,\,A},\hat{Y}^{r}_{d\cap c,\,A})$, (A.2)

et $V(\hat{N}_{d,\,AB})=V(\hat{N}^{r}_{d,\,A})+\alpha^2(V(\hat{N}^{r}_{d\cap c,\,A})+V(\hat{N}^{\pi}_{d\cap c,\,B}))-2\alpha\,C(\hat{N}^{r}_{d,\,A},\hat{N}^{r}_{d\cap c,\,A})$. (A.3)

Les domaines **c** définis par les bases-listes considérées n'étant pas identifiables à partir des données de l'EDM, on définit, pour chacune des bases, un domaine proxy **c'**, identifiable à partir des données de l'EDM, pour lequel la couverture de la population d'intérêt et l'incidence sont semblables à ceux observés pour le domaine **c** ($N_{d\cap c}/N_d\approx N_{d\cap c'}/N_d$ et $N_{d\cap c}/N_c\approx N_{d\cap c'}/N_{c'}$). On suppose que $S^2_{d\cap c}=S^2_{d\cap c'}=S^2_d$, $\bar{Y}_{d\cap c}=\bar{Y}_{d\cap c'}=\bar{Y}_d$ et que les effets de plan sont tels que $deff\,(\hat{Y}^{r}_{d\cap c,\,A})=deff\,(\hat{Y}^{r}_{d\cap c',\,A})$, $deff\,(\hat{Y}^{\pi}_{d\cap c,\,B})=deff\,(\hat{Y}^{\pi}_{d\cap c',\,A})$, $deff\,(\hat{N}^{r}_{d\cap c,\,A})=deff\,(\hat{N}^{r}_{d\cap c',\,A})$ et que $deff\,(\hat{N}^{\pi}_{d\cap c,\,B})=deff\,(\hat{N}^{\pi}_{d\cap c',\,A})$.

Sous ces hypothèses et en utilisant les formules de variance d'estimateurs d'Horvitz-Thompson d'une taille et d'un total pour un domaine selon un plan aléatoire simple on obtient :

$$V(\hat{Y}^{r}_{d\cap c,\,A})=V(\hat{Y}^{r}_{d\cap c',\,A})\frac{\left(N_{d\cap c}-1\right)S^2_d+N_{d\cap c}\left(1-N_{d\cap c}/N_A\right)\bar{Y}_d}{\left(N_{d\cap c'}-1\right)S^2_d+N_{d\cap c'}\left(1-N_{d\cap c'}/N_A\right)\bar{Y}_d}, \tag{A.4}$$

$$V(\hat{Y}^{\pi}_{d\cap c,\,B})=V(\hat{Y}^{\pi}_{d\cap c',\,A})\left(\frac{N_B{}^2}{N_A{}^2}\right)\left(\frac{1-n_B/N_B}{1-n_A/N_A}\right)\left(\frac{N_A-1}{N_B-1}\right)\left(\frac{\left(N_{d\cap c}-1\right)S^2_d+N_{d\cap c}\left(1-N_{d\cap c}/N_B\right)\bar{Y}_d}{\left(N_{d\cap c'}-1\right)S^2_d+N_{d\cap c'}\left(1-N_{d\cap c'}/N_A\right)\bar{Y}_d}\right), \tag{A.5}$$

$$V(\hat{N}^{r}_{d\cap c,\,A})=V(\hat{N}^{r}_{d\cap c',\,A})\left(\frac{N_{d\cap c}}{N_{d\cap c'}}\right)\left(\frac{N_A-N_{d\cap c}}{N_A-N_{d\cap c'}}\right)\text{ et} \tag{A.6}$$

$$V(\hat{N}^{\pi}_{d\cap c,\,B}) = V(\hat{N}^{\pi}_{d\cap c',\,A})\left(\frac{N_B - n_B}{N_A - n_A}\right)\left(\frac{N_A - 1}{N_B - 1}\right)\left(\frac{n_A}{n_B}\right)\left(\frac{N_{d\cap c}}{N_{d\cap c'}}\right)\left(\frac{N_B - N_{d\cap c}}{N_A - N_{d\cap c'}}\right). \quad \text{(A.7)}$$

On pose également :

$$C(\hat{Y}^{r}_{d,\,A}, \hat{Y}^{r}_{d\cap c,\,A}) = C(\hat{Y}^{r}_{d,\,A}, \hat{Y}^{r}_{d\cap c',\,A})\sqrt{\frac{\left(N_{d\cap c} - 1\right)S_d^2 + N_{d\cap c}\left(1 - N_{d\cap c}/N_A\right)\overline{Y}_d}{\left(N_{d\cap c'} - 1\right)S_d^2 + N_{d\cap c'}\left(1 - N_{d\cap c'}/N_A\right)\overline{Y}_d}}, \quad \text{(A.8)}$$

$$C(\hat{N}^{r}_{d,\,A}, \hat{N}^{r}_{d\cap c,\,A}) = C(\hat{N}^{r}_{d,\,A}, \hat{N}^{r}_{d\cap c',\,A})\sqrt{\left(\frac{N_{d\cap c}}{N_{d\cap c'}}\right)\left(\frac{N_A - N_{d\cap c}}{N_A - N_{d\cap c'}}\right)} \text{ et} \quad \text{(A.9)}$$

$$C(\hat{Y}_{d,\,AB}, \hat{N}_{d,\,AB}) = C(\hat{Y}^{r}_{d,\,A}, \hat{N}^{r}_{d,\,A})\sqrt{\left(\frac{V\left(\hat{Y}_{d,\,AB}\right)}{V\left(\hat{Y}^{r}_{d,\,A}\right)}\right)\left(\frac{V\left(\hat{N}_{d,\,AB}\right)}{V\left(\hat{N}^{r}_{d,\,A}\right)}\right)}. \quad \text{(A.10)}$$

En substituant A.4, A.5 et A.8 dans A.2, A.6, A.7 et A.9 dans A.3, A.2 et A.3 dans A.10, et finalement A.2, A.3 et A.10 dans A.1, on obtient une expression de la variance approximative de $\hat{\overline{Y}}_{d,AB}$. Les variances et les covariances des différents estimateurs pour la base aréolaire **A** ainsi que $\overline{Y}_d$ et S_d^2 sont remplacées par leurs valeurs estimées à partir des données de l'EDM 2000 et de l'EDM 2001 et les tailles de population et de domaines sont remplacées par leurs valeurs estimées à l'aide des données du Recensement de 2001 afin de calculer une variance approximative. Une valeur de $\alpha = 0{,}7$ a été utilisée pour l'évaluation, celle-ci représentant un compromis satisfaisant pour les dix variables dont on devait estimer la moyenne, et ce pour chacune des bases évaluées. Les résultats présentés à la figure 4.1 correspondent au maximum des CV obtenus en utilisant les données des EDM de 2000 et 2001 pour la substitution.

REMERCIEMENTS

Les auteurs remercient Guy Laflamme et Michel Latouche pour leurs commentaires constructifs qui ont permis d'améliorer cet article.

RÉFÉRENCES

Arsenault, S. et Tremblay, J. (2001), "Méthodologie de l'Enquête sur les dépenses des ménages", Statistique Canada, Division de la statistique du revenu, No. 62F0026MIF-2001003 au catalogue.

Duggan, J., Neusy, E. et Bélanger, Y. (2003), "Sample Design Issues in a Large-scale Multiple frame National Survey: The Canadian Component of the Adult Literacy and Life-skills Survey", *Proceedings of the Survey Methods Section*, SSC Annual Meeting.

Fugère, D. et Lanctôt, P. (1985), "Méthodologie de détermination des seuils de revenu minimum au Québec", Ministère de la Main d'œuvre et de la Sécurité du revenu.

Kalton, G. (2001), "Practical Methods for Sampling Rare and Mobile Populations", *Proceedings of the Annual Meeting of the American Statistical Association.*

Kalton, G. et Anderson, D. W. (1986), "Sampling Rare Populations", *Journal of the Royal Statistical Society*, 149, pp. 65-82.

Lavigne, M. et Michaud, S. (1998), "Aspects généraux de l'Enquête sur la dynamique du travail et du revenu", Statistique Canada, Division de la statistique du revenu, No. 75F0026MIF-1998005 au catalogue.

Nadeau, C., Lapierre, B., Tremblay, J. et Gaudet, J. (2005), "Plan de sondage pour l'ajout d'un échantillon supplémentaire à l'Enquête sur les dépenses des ménages de 2003 pour le Ministère des Finances du Québec", Document de travail de la Division des méthodes d'enquêtes auprès des ménages, Statistique Canada.

Swain, L., Drew, J. D., Lafrance, B. et Lance, K. (1992), "La création d'un registre d'adresses résidentielles pour améliorer la couverture du recensement du Canada de 1991", *Techniques d'enquêtes*, 18, pp.139-155.