

# UTILISATION D'UN PLAN À BASE DUALE POUR CIBLER UNE POPULATION À FAIBLE REVENU À L'ENQUÊTE SUR LES DÉPENSES DES MÉNAGES

Christian Nadeau et Bruno Lapierre<sup>1</sup>

## RÉSUMÉ

L'Enquête sur les dépenses des ménages de Statistique Canada est une enquête annuelle dont le principal objectif est de produire des estimations de dépenses aux échelles provinciale et nationale pour l'ensemble de la population. Lors de l'enquête de 2003, le Ministère des finances du Québec a financé l'ajout d'un échantillon supplémentaire afin d'améliorer la qualité de certaines estimations de dépenses pour une population à faible revenu de la province de Québec. Un plan de sondage à base duale a été développé pour la sélection de l'échantillon de cette province afin de rencontrer cet objectif additionnel. Dans cet article, nous décrivons le plan de sondage développé dans le cadre de ce projet de même que la méthodologie de pondération utilisée et nous présentons une évaluation qui permet de juger de l'efficacité de ce plan.

## 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 reposait en grande partie sur des estimations de moyennes de dépenses pour une population d'intérêt qui, selon le Recensement de 2001, ne représente que 2,5% de l'ensemble des ménages du Québec. Celle-ci est constituée des 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.

Afin de refaire cette étude à partir d'estimations de dépenses à jour plus précises que celles habituellement disponibles, le MFQ a financé l'ajout d'un échantillon supplémentaire à l'Enquête sur les dépenses des ménages (EDM) de 2003 de Statistique Canada. L'EDM est une enquête annuelle qui vise à produire des estimations de dépenses des ménages aux échelles provinciale et nationale pour l'ensemble des ménages. Le plan de sondage actuel ne permet pas de cibler les ménages de la population d'intérêt à laquelle s'intéresse le MFQ. Pour l'EDM 2003, un plan de sondage à base duale a été mis sur pied afin de répondre à l'objectif du MFQ tout en limitant la taille de l'échantillon supplémentaire.

Nous décrivons d'abord les différents éléments du plan de sondage à la section 2 et la méthodologie de pondération à la section 3. Une évaluation sommaire est ensuite présentée à la section 4.

## 2. PLAN DE SONDRAGE

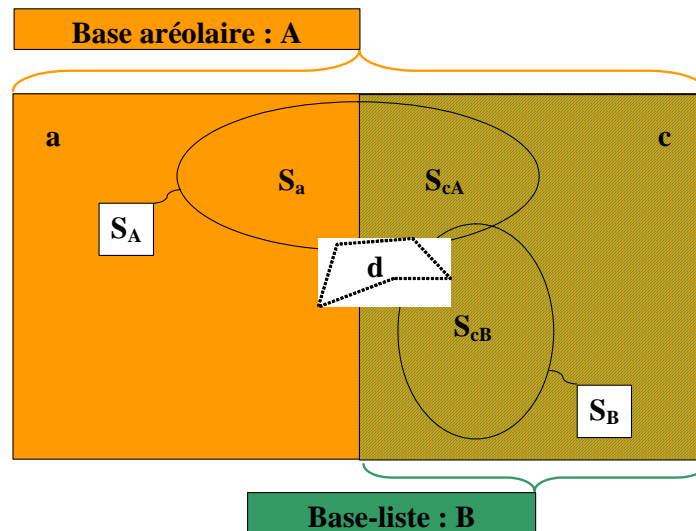
La taille de l'échantillon régulier de l'EDM pour le Québec est de 3 110 logements pour l'enquête de 2003. Cet échantillon est sélectionné à partir d'une base aréolaire selon un plan stratifié à plusieurs degrés (Arsenault et Tremblay, 2001). Comme cette base ne permet pas de cibler la population d'intérêt du MFQ, un plan de sondage à base duale a été élaboré afin d'améliorer la précision des estimations pour la population d'intérêt. On trouvera une description plus détaillée du plan de sondage dans Lapierre, Nadeau, Tremblay et Gaudet (2004).

---

<sup>1</sup> Christian Nadeau, Statistique Canada, Immeuble R.H. Coats, 16<sup>ième</sup> étage, Ottawa, Canada, K1A 0T6, Christian.Nadeau@Statcan.ca; Bruno Lapierre, Statistique Canada, Édifice Principal, pièce 2500, Ottawa, Canada, K1A 0T6, Bruno.Lapierre@Statcan.ca.

La figure 1 illustre l'utilisation d'une telle base duale. La base aréolaire **A** de laquelle est tiré l'échantillon régulier  $S_A$  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  $S_B$ . Dans le cas général d'utilisation de base duale, on partitionne la population en trois parties: la partie couverte par la base **A** seulement ( $a=A \cap B^c$ ), la partie couverte par la base **B** seulement ( $b=A^c \cap B$ ) et la partie commune aux deux bases ( $c=A \cap B$ ). Dans le cas présent où  $B \subseteq A$ , on note que  $b=\emptyset$  et que  $c=B$ . On partitionne également l'échantillon  $S_A$  pour obtenir  $S_a=S_A \cap a$  et  $S_{cA}=S_A \cap c$  et on note que  $S_B=S_{cB}$ . Le domaine défini par la population d'intérêt du MFQ sera noté par **d**.

**Figure 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-liste est construite à partir du Registre des adresses de Statistique Canada de façon à favoriser le suréchantillonnage des ménages de la population d'intérêt. Elle est constituée des logements situés dans les aires géographiques présentant les plus fortes prévalences<sup>2</sup> selon le Recensement de 2001. Elle couvre environ 50% de l'ensemble des ménages du Québec. Lors du Recensement de 2001, la prévalence sur la base-liste était de 4,4% et la couverture de la population d'intérêt était de 88%.

La base-liste est stratifiée selon trois niveaux. Les deux premiers niveaux de stratification sont de nature géographique et permettent de former douze strates. Le troisième niveau de stratification consiste à diviser chacune des douze strates de deuxième niveau en deux afin de regrouper les aires géographiques présentant les plus fortes prévalences dans la première de ces deux strates. Les détails relatifs à l'identification de la borne de stratification sont présentés dans Nadeau, Lapierre, Tremblay et Gaudet (2005). Cette borne est déterminée de façon à augmenter le gain d'efficacité obtenu à l'aide d'une répartition de l'échantillon proportionnelle à  $N_h \sqrt{P_h}$  dans les strates  $h$  de troisième niveau. Sous certaines hypothèses, une telle répartition correspond à la répartition de Neyman pour l'estimation d'une moyenne de domaine (Kalton et Anderson, 1986).

L'échantillon supplémentaire est d'abord réparti entre les strates du premier niveau proportionnellement au nombre de logements occupés lors du Recensement de 2001. L'échantillon de chaque strate de premier niveau est réparti entre les strates  $h$  de deuxième niveau proportionnellement 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. Cette méthode de répartition est également utilisée pour répartir l'échantillon des strates de deuxième niveau entre celles du troisième niveau.

Un plan d'échantillonnage à deux degrés est utilisé pour la sélection des logements à l'intérieur de chacune des strates afin d'éviter une trop grande dispersion de l'échantillon. Les aires géographiques utilisées pour la construction de la base-liste servent d'unités primaires d'échantillonnage (UPÉ) et sont sélectionnées avec une probabilité proportionnelle au nombre de logements occupés sur la base-liste selon le Recensement de 2001.

<sup>2</sup> La prévalence est définie par  $P=M/N$  où  $M$  représente la taille de la population d'intérêt et  $N$  représente la taille de l'ensemble de la population.

Dépendamment du niveau d'urbanisation de la strate, on sélectionne trois ou cinq logements à l'intérieur de chacune des UPÉ choisies au premier degré d'échantillonnage selon un plan d'échantillonnage systématique. La taille de l'échantillon supplémentaire est de 2 211 logements. Un seul logement a été identifié comme appartenant aux deux échantillons ( $S_{cA} \cap S_{cB}$ ).

### 3. MÉTHODOLOGIE DE PONDÉRATION

La pondération vise à produire un ensemble unique de poids de sondage utilisé pour estimer différentes caractéristiques, notamment des totaux et des moyennes de dépenses à l'échelle de la province et des moyennes de dépenses à l'intérieur du domaine d'intérêt du MFQ. Les poids peuvent être exprimés comme le produit de six composantes obtenues lors d'autant d'étapes de pondération. Celles-ci consistent au calcul des poids de sondage théoriques, à la suppression des unités hors du champs de l'enquête, à l'ajustement pour la non-réponse, à l'intégration des deux échantillons, à l'ajustement pour les observations influentes et au calage aux marges. Une description détaillée de la méthodologie de pondération est présentée dans Nadeau et Lapierre (2005). Nous traiterons ici brièvement de quatre de ces étapes.

*Calcul des poids de sondage théoriques:* Cette étape est effectuée indépendamment pour l'échantillon régulier et l'échantillon supplémentaire. Elle consiste à assigner à chaque logement échantillonné l'inverse de sa probabilité d'inclusion dans l'échantillon auquel il appartient.

*Ajustement pour compenser la non-réponse totale:* Afin de compenser la perte de représentativité causée par la non-réponse, on gonfle les poids de sondage théoriques des ménages répondants par un facteur égal à l'inverse du taux de réponse pondéré à l'intérieur de groupes de réponse homogènes (GRH). Les GRH sont préalablement formés à l'aide de l'information auxiliaire disponible sur chaque base de sondage. Comme celle-ci diffère selon la base de sondage, la formation des GRH est effectuée indépendamment à l'intérieur de chacun des échantillons. Les GRH utilisés pour l'échantillon régulier correspondent à des regroupements de strates qui tiennent compte de la situation géographique, du niveau d'urbanisation et de l'identification des strates à forte proportion de ménages à haut revenu. Les GRH de l'échantillon supplémentaire sont formés à l'aide d'une version améliorée de l'algorithme CHAID (Chi-Square Automatic Interaction Detection) disponible dans le logiciel Knowledge Seeker (ANGOSS Software, 1995). L'information auxiliaire utilisée tient compte de la situation géographique de l'UPÉ, de certaines caractéristiques des personnes, des ménages et des logements à l'intérieur de l'UPÉ selon le Recensement de 2001.

*Intégration des deux échantillons:* Avant cette étape, chaque échantillon est traité indépendamment. Étant donné l'utilisation d'un plan à base duale où la base-liste est incluse dans la base aréolaire ( $B \subseteq A$ ), les estimateurs de totaux prendront la forme générale  $\hat{Y} = \alpha \hat{Y}_{c,B} + (1 - \alpha) \hat{Y}_{c,A} + \hat{Y}_a$  où  $0 \leq \alpha \leq 1$ . L'intégration des deux échantillons de façon à produire un ensemble de poids cohérent avec cet estimateur consiste d'abord à déterminer si les logements de l'échantillon régulier  $S_A$  appartiennent à  $S_a$ , la partie non couverte par la base-liste  $B$ , ou à  $S_{cA}$ , la partie couverte par  $B$ . Cette identification s'effectue par appariement entre  $B$  et  $S_A$  à l'aide des adresses et des identificateurs géographiques. Le facteur d'intégration prendra la valeur  $\alpha$  pour les ménages répondants de l'échantillon supplémentaire  $S_B$ , la valeur  $1 - \alpha$  pour les ménages répondants de  $S_{cA}$  et la valeur 1 pour ceux de  $S_a$ . La valeur de  $\alpha$  est calculée de façon à favoriser la réduction des coefficients de variation pour l'estimation de la taille du domaine, et par le fait même pour les différentes estimations de moyennes de dépenses pour ce même domaine, selon la formule suivante:

$$\alpha = \frac{\frac{(N_A - n_A)(N_{d \cap c, A})(N_A - N_{d \cap c, A})}{n_A(N_A - 1)}}{\frac{(N_A - n_A)(N_{d \cap c, A})(N_A - N_{d \cap c, A})}{n_A(N_A - 1)} + \frac{\text{deff}(\hat{N}_{d \cap c, B})(N_B - n_B)(N_{d \cap c, B})(N_B - N_{d \cap c, B})}{\text{deff}(\hat{N}_{d \cap c, A})n_B(N_B - 1)}}.$$

Les tailles d'échantillon  $n_A$  et  $n_B$  correspondent respectivement au nombre de ménages répondants dans l'échantillon régulier et dans l'échantillon supplémentaire.  $N_A$  et  $N_B$  correspondent respectivement au nombre de ménages dans la population et au nombre de ménages sur la base-liste, alors que  $N_{d \cap c, A}$  et  $N_{d \cap c, B}$  sont égaux et correspondent au nombre de ménages du domaine d'intérêt dans la partie commune aux deux bases selon le Recensement de 2001. De façon générale, l'effet de plan dépend essentiellement de quatre facteurs: la

stratification, la répartition, la méthode d'échantillonnage dans les strates et l'estimateur. Le ratio  $deff(\hat{N}_{d \cap c, B}) / deff(\hat{N}_{d \cap c, A})$  est calculé sous l'hypothèse que seules les différences de stratification et de répartition ont une incidence sur l'efficacité relative de l'échantillon supplémentaire par rapport à celle de l'échantillon régulier.

*Calage aux marges:* Une étape de calage aux marges est effectuée de façon à réduire la variance des estimations et à respecter certains totaux connus de la population. Les effectifs de population selon un certain nombre de catégories d'âge-sexe et de salaires et traitements de même que le nombre de ménages selon leur taille et leur composition sont utilisés lors du calage (Arsenault, Gaudet, Nadeau et Tremblay, 2001). Le calage est effectué selon la méthode de pondération intégrée introduite par Lemaître et Dufour (1987) et mène à l'obtention de facteurs de calage pour tous les ménages répondants.

#### 4. ÉVALUATION DU PLAN DE SONDAGE

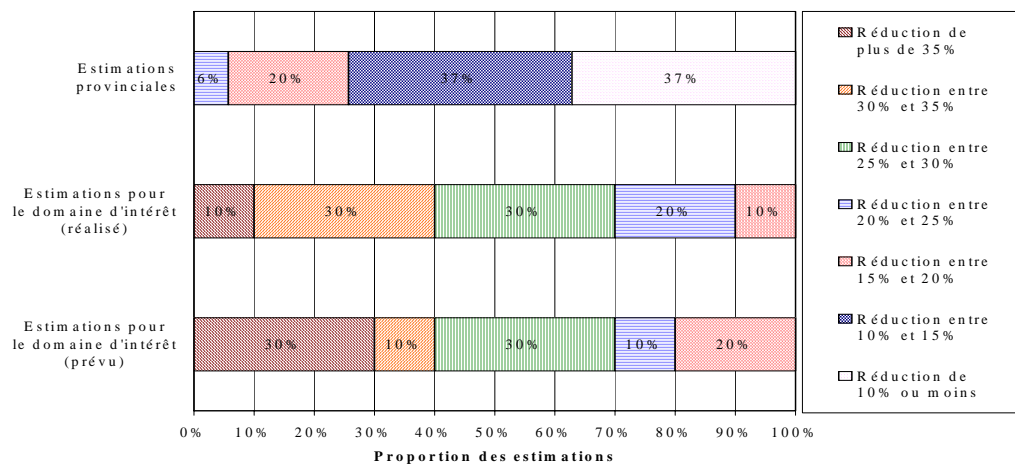
Lors du développement du plan de sondage, des coefficients de variation espérés ont été calculés sous différentes hypothèses, dont celle que les prévalences observées dans les différentes strates de la base-liste lors du Recensement de 2001 demeureraient les mêmes (Nadeau, Lapierre, Tremblay et Gaudet, 2005). Les résultats présentés au tableau 1 démontrent que cette hypothèse était trop optimiste. Tel que prévu, on constate que la prévalence observée lors de l'enquête est plus élevée dans les strates présentant les plus hautes prévalences lors du Recensement de 2001 que dans celles présentant les plus faibles prévalences. L'écart est toutefois moindre que prévu.

**Tableau 1: Prévalence non-pondérée sur la base-liste selon le troisième niveau de stratification**

Type de strate	Prévalence espérée	Prévalence observée
Haute prévalence	7,4%	4,0%
Faible prévalence	2,5%	2,0%
Total	5,0%	3,1%

Malgré ces observations, on constate à la figure 2 que la réduction des coefficients de variation engendrée par l'ajout de l'échantillon supplémentaire n'est que légèrement inférieure à celle prévue dans le cas des estimations pour le domaine d'intérêt. En effet, la réduction relative des coefficients de variation engendrée par l'ajout de l'échantillon supplémentaire est supérieure à 35% pour une seule des dix variables auxquelles s'intéresse le MFQ comparativement à une prévision de trois à l'étape de développement. Par contre, elle est supérieure à 20% pour neuf variables comparativement à une prévision de huit lors du développement. Les réductions relatives de coefficients de variation observées sont généralement supérieures à la réduction moyenne de 24% à laquelle on se serait attendu suite à une augmentation de la taille de l'échantillon régulier de 2 211 logements sans en changer le plan de sondage.

**Figure 2: Réduction relative des coefficients de variation suite à l'ajout de l'échantillon supplémentaire**



On observe également que l'amélioration de la précision des estimations de dépenses pour le domaine d'intérêt est supérieure à celle obtenue pour les estimations de dépenses provinciales. En effet la réduction relative des coefficients de variation est toujours supérieure à 15% pour les estimations de dépenses pour le domaine d'intérêt alors qu'une telle réduction n'est observée que pour 26% des estimations provinciales considérées. Il est à noter que les estimations provinciales considérées correspondent aux trente-deux principales catégories de dépenses de l'EDM et à trois variables de revenus et ne coïncident pas avec les dix variables de dépenses considérées pour le domaine d'intérêt.

Bien que la prévalence sur la base-liste soit inférieure à celle supposée lors de l'élaboration du plan, l'impact de l'ajout de l'échantillon supplémentaire selon le plan adopté est très similaire à celui envisagé. La figure 2 confirme que le plan adopté privilégie l'amélioration des estimations pour le domaine d'intérêt par rapport aux estimations provinciales contrairement à ce qui aurait été obtenu par une simple augmentation de la taille de l'échantillon régulier. Une évaluation plus complète est présentée dans Nadeau et Lapierre (2005).

## REMERCIEMENTS

Les auteurs remercient François Brisebois, François Gagnon et Johanne Tremblay pour leurs commentaires constructifs qui ont permis d'améliorer ce texte.

## RÉFÉRENCES

- ANGOSS Software (1995), « Knowledge Seeker IV for Windows – Users's Guide », ANGOSS Software International Limited.
- Arsenault, S., Gaudet, J., Nadeau, C. et Tremblay, J. (2001), « Introduction of a New Calibration Strategy for the Survey of Household Spending », *Proceedings of the Annual Meeting of the American Statistical Association*.
- 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-01003 au catalogue, octobre 2001.
- 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. et Anderson, D.W. (1986), « Sampling Rare Populations », *Journal of the Royal Statistical Society*, 129, pp.65-82.
- Lapierre, B., Nadeau, C., Tremblay, J. et Gaudet, J. (2004), « 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 », *Recueil du Symposium 2004 de Statistique Canada*.
- Lemaître, G. et Dufour, J. (1987), « Une méthode intégrée de pondération des personnes et des familles », *Techniques d'enquête*, Vol.13, pp. 211-220.
- Nadeau, C. et Lapierre, B. (2005), « Utilisation d'une base duale pour cibler une population à faible revenu: Stratégie de pondération et évaluation du plan de sondage », Document de travail de la Division des méthodes d'enquêtes auprès des ménages, Statistique Canada.
- 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.