Composante du produit nº 11-522-X au catalogue de Statistique Canada
La série des symposiums internationaux de Statistique Canada : recueil

Article

Symposium 2008 :
Collecte des données : défis, réalisations et nouvelles orientations

Utilisation de paradonnées dans la gestion de la non-réponse de l'Enquête sur la dynamique du travail et du revenu

par Wisner Jocelyn, Owen Phillips, Beatrice Baribeau et Amélie Lévesque

2009

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2008 de Statistique Canada
Collecte des données : défis, réalisations et nouvelles orientations

Utilisation de paradonnées dans la gestion de la non-réponse de l'Enquête sur la dynamique du travail et du revenu

Wisner Jocelyn, Owen Phillips, Beatrice Baribeau et Amélie Lévesque[1]

Résumé

Au cours des dernières années, l'utilisation des paradonnées a pris de plus en plus d'importance dans le cadre de la gestion des activités de collecte à Statistique Canada. Une attention particulière a été accordée aux enquêtes sociales menées par téléphone, comme l'Enquête sur la dynamique du travail et du revenu (EDTR). Lors des dernières activités de collecte de l'EDTR, une limite de 40 tentatives d'appel a été instaurée. Des examens des fichiers de l'historique des transactions Blaise de l'EDTR ont été entrepris afin d'évaluer l'incidence de la limite des tentatives d'appel. Tandis que l'objectif de la première étude était de réunir les renseignements nécessaires à l'établissement de la limite des tentatives d'appel, la seconde étude portait sur la nature de la non-réponse dans le contexte de la limite de 40 tentatives.

L'utilisation des paradonnées comme information auxiliaire pour étudier et expliquer la non-réponse a aussi été examinée. Des modèles d'ajustement pour la non-réponse utilisant différentes variables de paradonnées recueillies à l'étape de la collecte ont été comparés aux modèles actuels basés sur de l'information auxiliaire tirée de l'Enquête sur la population active.

Mots clés : Gestion de la collecte, non-réponse, paradonnées.

1. Introduction

Au cours des dernières années, l'utilisation de données sur le processus de collecte, ou paradonnées,[2] a pris de plus en plus d'importance dans le cadre de la gestion des activités de collecte à Statistique Canada. Pour répondre aux besoins en renseignements à jour et appropriés sur les coûts et la qualité de la collecte de données, Statistique Canada a mis au point un ensemble de processus et d'outils regroupés sous le titre de « gestion active de la collecte » (GAC) (Hunter et Carbonneau, 2005; Laflamme, Maydan et Miller, 2008). Une attention particulière a été accordée aux enquêtes sociales menées par téléphone, comme l'Enquête sur la dynamique du travail et du revenu (EDTR).

Le principal objectif de la communication est d'illustrer l'utilisation des paradonnées dans l'EDTR et d'examiner la possibilité d'y avoir davantage recours dans le contexte de l'ajustement pour la non-réponse.

Le document se présente de la façon suivante. La section 2 donne un aperçu de l'EDTR et de certains concepts utilisés dans le document, y compris la GAC appliquée à l'EDTR. Les études sur la limite des tentatives d'appel et les résultats connexes sont résumés à la section 3. À la section 4, les modèles d'ajustement pour la non-réponse utilisant des paradonnées sont décrits et les résultats préliminaires sont exposés. Enfin, le lecteur trouvera à la section 5 le mot de la fin et un aperçu des travaux futurs.

2. Présentation de l'EDTR

L'EDTR est une enquête longitudinale mesurant les changements dans le bien-être économique des Canadiens et les facteurs pouvant influer sur ces changements. D'une façon plus précise, elle suit une cohorte longitudinale pendant

[1]Wisner Jocelyn, Owen Phillips, Beatrice Baribeau et Amélie Lévesque, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.
[2] Pour une définition plus complète des paradonnées, prière de consulter Scheuren (2005).

six ans, recueillant chaque année des données sur le travail, le revenu et la situation familiale. Un panel ou une cohorte est un sous-échantillon d'environ 17 000 ménages répondants de l'Enquête sur la population active (EPA). Un nouveau panel est introduit tous les trois ans.

Tous les membres des ménages de l'EPA sont soumis aux observations longitudinales. Comme les données de l'enquête servent également à produire des estimations transversales du revenu familial et personnel, des cohabitants non longitudinaux sont interviewés. Les interviews sont menées en utilisant la technique des interviews téléphoniques assistées par ordinateur (ITAO). Tandis que des renseignements démographiques sont recueillis pour tous les membres du ménage, les questions sur le travail et le revenu ne sont posées qu'aux personnes de 16 ans et plus. Toutes les questions portent sur les activités pendant l'année civile (l'année de référence) précédant la collecte des données. Pour réduire le fardeau de réponse, les réponses par personne interposée sont autorisées, et les répondants peuvent donner à Statistique Canada la permission de consulter leurs déclarations de revenus plutôt que de répondre aux questions sur le revenu.

Les taux de réponse sont calculés chaque année. Le taux de réponse transversal est défini au niveau du ménage, et un ménage répondant est un ménage où au moins un membre répond aux questions sur le travail ou aux questions sur le revenu ainsi qu'à une partie minimale du questionnaire. D'autre part, le taux de réponse longitudinal est défini au niveau individuel et ne repose que sur l'échantillon longitudinal, malgré la disponibilité de renseignements longitudinaux concernant certains cohabitants.

Les taux de réponse sont calculés chaque année. Le taux de réponse transversal est défini au niveau du ménage, et un ménage répondant est un ménage où au moins un membre répond aux questions sur le travail ou aux questions sur le revenu ainsi qu'à une partie minimale du questionnaire. D'autre part, le taux de réponse longitudinal est défini au niveau individuel et ne repose que sur l'échantillon longitudinal.

2.1 Gestion active de la collecte des données de l'EDTR

Les responsables de l'EDTR ont mis sur pied un groupe chargé de la gestion active de la collecte (GAC) des données. Le groupe est composé de représentants des divisions responsables de la collecte des données, de l'élaboration des applications d'ITAO, de l'élaboration du contenu spécialisé et du traitement des données ainsi que des méthodes d'enquête. Dans les phases précédant et suivant la collecte, les membres du groupe examinent les améliorations pouvant être apportées à la prochaine phase de la collecte, et ils voient à ce que les changements proposés soient intégrés et mis à l'essai afin de réduire au minimum l'incidence des éventuels problèmes.

Entre autres choses, le groupe de la GAC est responsable de l'établissement et de la mise en œuvre des nouveaux outils visant à mieux gérer la collecte des données de l'EDTR dans un contexte d'ITAO ainsi que d'évaluer l'efficacité de ces outils. Voici des exemples d'initiatives prises par le groupe.

- Les *tranches de temps* sont des parties de la semaine établies en fonction de l'heure du jour et du jour de la semaine. Les tentatives d'appel sont réparties dans différentes tranches de temps afin d'optimiser la probabilité de contacter un ménage en particulier. Les tranches de temps sont définies par le bureau central avant la collecte, mais elles peuvent être modifiées pendant la collecte. Les *groupes de tranches de temps* sont établis en fonction des caractéristiques démographiques du ménage. Ils sont utilisés pour déterminer la répartition des tentatives d'appel dans les tranches de temps.
- Les *groupes Z* sont utilisés pour tenter de s'assurer que des taux de réponse appropriés sont obtenus concernant des domaines d'intérêt particuliers. Si un taux de réponse concernant un domaine particulier est jugé faible, le groupe z correspondant est activé, ce qui place sur la liste des ménages à contacter en priorité ceux qui possèdent la caractéristique d'intérêt.
- L'*ordonnanceur d'appels* fait partie du système de gestion des applications d'ITAO. Il travaille en arrière-plan et attribue les cas selon la tranche de temps, la répartition des tranches de temps, le groupe z et d'autres critères (interviews prévues, par exemple).

L'objectif de ces outils est de tirer le meilleur parti des tentatives d'appel. La gestion efficace des activités de collecte de données revêt de plus en plus d'importance, particulièrement dans le contexte de la limite imposée au nombre d'appels.

3. Limite des appels

Afin de réduire le fardeau de réponse et les coûts de la collecte, un nombre maximal de tentatives d'appel a été fixé en 2007 dans le cadre des enquêtes-ménages. Ce nombre a été fixé à 25 en général et à 40 pour l'EDTR en particulier. À l'été 2007, deux examens des fichiers de l'historique des transactions Blaise de l'EDTR ont été entrepris afin d'évaluer l'incidence de la limite des tentatives d'appel. Les données des trois plus récentes années d'enquête ont alors été utilisées.

La première étude (Lévesque et Poulin, 2007) a été réalisée sous l'impulsion du groupe de la GAC. Les auteurs cherchaient avant tout à évaluer l'incidence de l'établissement du nombre maximal de tentatives d'appel à 25. Ils ont constaté que, si la limite de 25 tentatives avait été adoptée en 2007, le taux de réponse transversal aurait été légèrement supérieur à 73 % plutôt que de se situer à 77 % avec une limite de 40 tentatives. De surcroît, le pourcentage de répondants longitudinaux perdus sur la période de trois ans aurait pu atteindre 21 % si la limite de 25 appels avait été adoptée en 2005 et 17 % si la limite de 40 appels avait été mise en œuvre en 2005. Comme des épisodes de non-réponse cyclique sont permis dans l'EDTR, où les ménages exigeant un grand effort ne sont généralement pas les mêmes d'une année à l'autre, ces estimations présentent le pire des scénarios et il est possible qu'il ne reflète pas la réalité. Cela dit, il n'est pas déraisonnable de croire que les efforts supplémentaires déployés pour entrer en contact avec un ménage dans une année facilitent le contact les années suivantes.

Les ménages exigeant un grand effort possèdent également des caractéristiques démographiques et socioéconomiques différentes des ménages exigeant un effort moindre, ce qui entraîne une possibilité de biais dans les estimations de certains domaines. Assez naturellement, la proportion de ménages exigeant plus de 25 tentatives est plus élevée chez les ménages composés uniquement de jeunes adultes célibataires. Les autres domaines affichant la même tendance sont les ménages d'immigrants, les ménages des provinces de l'Ouest et les ménages formés de cinq personnes ou plus. De plus, des différences sur le plan du revenu ont été observées chez les ménages formés d'une seule personne : le revenu médian des ménages exigeant un grand effort était supérieur à celui des ménages exigeant un effort moindre par une marge de 7 000 dollars.

Les auteurs de l'étude ont également examiné les caractéristiques démographiques des ménages exigeant un grand effort, les taux de prise de contact, la répartition des appels selon l'heure du jour de même que la séquence des appels afin d'établir la façon de tirer le meilleur parti possible des tentatives d'appel. Ils ont constaté ceci :

- la probabilité de contacter le répondant était plus élevée dans la soirée, et ce même si un plus grand nombre de tentatives étaient faites le jour. D'autres chercheurs ont observé ce phénomène dans le contexte d'autres enquêtes utilisant la technique d'ITAO (Laflamme, 2008);
- en 2007, un pourcentage élevé (68 %) des cas pour lesquels le nombre limite de tentatives avait été atteint étaient terminés aux deux tiers de la période de collecte;
- des tentatives n'avaient pas été faites dans toutes les tranches de temps concernant certains non-répondants pour lesquels la limite avait été atteinte. L'étude de ces cas a fait ressortir une incidence plus élevée de lignes occupées que les autres cas;
- au moins une tentative de prise de contact dans toutes les tranches de temps avait été faite pour seulement 55 % des non-répondants;
- on a plus souvent recours à l'élément *Parcourir* dans le cas des non-répondants et dans les régions où les taux de non-réponse sont élevés. Cette fonction permet à l'intervieweur de sélectionner des cas et de contourner ainsi la répartition établie par l'ordonnanceur d'appels.

Dans leur rapport final, les auteurs ont formulé un certain nombre de suggestions pour améliorer la gestion des activités de collecte de données pour 2008. Entre autres choses, ils ont recommandé que les responsables de l'EDTR :

- maintiennent une limite de 40 tentatives d'appel pour 2008 et qu'ils confient les cas ayant atteint de 32 à 35 tentatives à un groupe spécial afin de répartir les autres tentatives sur ce qui reste de la période de collecte;
- augmentent le nombre de tentatives faites en soirée et en fin de semaine;

- définissent les groupes de tranches de temps de façon qu'une tentative soit faite dans toutes les tranches de temps pour chaque cas;
- établissent de meilleures lignes directrices quant à l'utilisation de l'élément *Parcourir*, même si celle-ci est justifiée dans de nombreux cas.

La seconde étude (Chapman, 2007a et 2007b) portait sur la nature de la non-réponse dans le contexte de la limite de 40 tentatives. Les résultats de la collecte ont été évalués pour l'échantillon complet et au niveau des panels. Pour la période de collecte des données de 2007, nous avons assisté à une chute des taux de réponse, des taux de prise de contact et des taux de refus pour les deux panels (4 et 5). Le taux de refus suit la tendance observée pendant la durée de vie de ces panels et d'autres panels, mais la diminution des taux de réponse[3] et des taux de prise de contact ne va pas dans le même sens que les augmentations généralement observées pour ce qui est des premiers panels. Même si la hausse de la non-réponse en 2007 ne peut pas être entièrement attribuée à la limite du nombre de tentatives d'appel, environ 4 % de tous les ménages n'ont pas répondu parce qu'ils avaient tout simplement atteint la limite des tentatives d'appel.

Dans la prochaine section, les auteurs examinent de nouveau certains résultats et changements proposés en tenant compte des résultats de la collecte de 2008.

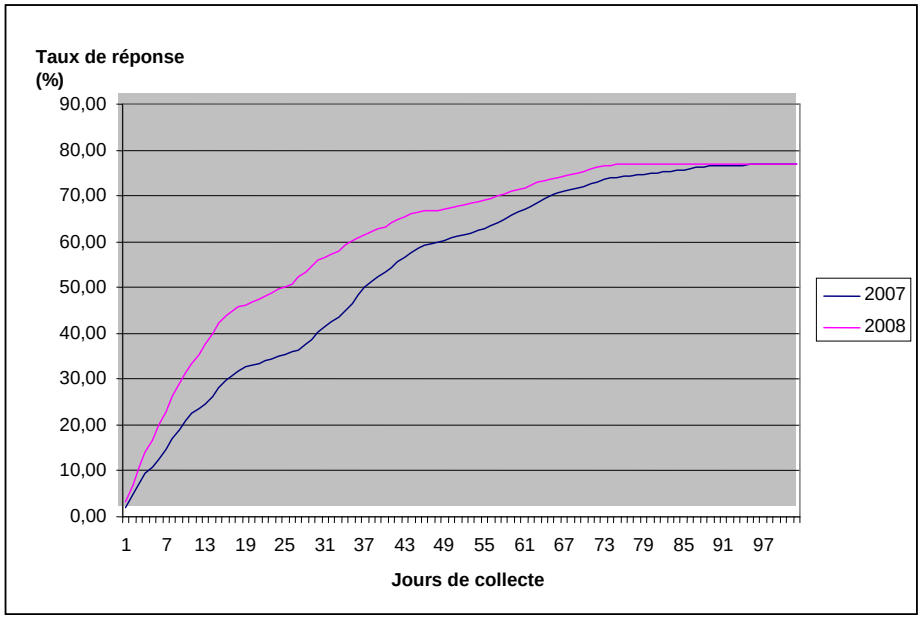
3.1 Résultats de la collecte de 2008

Les changements suggérés n'ont pas tous été adoptés dans la période de collecte de 2008, pour des raisons d'ordre opérationnel et budgétaire. Les propositions d'amélioration visant notamment à répartir les tentatives d'appel dans les groupes de tranches de temps pendant toute la période de collecte ainsi qu'à s'assurer que des tentatives sont faites dans toutes les tranches de temps n'ont pas été mise en œuvre. Cependant, les résultats préliminaires de la collecte indiquaient une grande amélioration au chapitre des taux de réponse par rapport à 2007 (figure 3.1-1). En règle générale, le nombre de tentatives faites en soirée et en fin de semaine a augmenté, et l'utilisation de l'élément *Parcourir* a considérablement diminué pour se situer à des niveaux semblables dans toutes les régions. Toutefois, le pourcentage de cas exigeant plus de 25 tentatives a augmenté et, bien que le pourcentage de cas ayant atteint le maximum permis de tentatives mais sans que des tentatives aient été faites dans toutes les tranches de temps ait diminué (2,7 % contre 5,3 % en 2007), le pourcentage de cas ayant atteint le maximum permis a augmenté, et le maximum a été atteint alors qu'il restait deux semaines de collecte dans 54 % de ces cas. En outre, le nombre moyen d'appels faits avant d'établir un premier contact a augmenté dans toutes les régions. Au fur et à mesure que la collecte avançait, l'écart entre 2008 et 2007 se rétrécissait et, à la fin de la collecte,[4] seule une augmentation marginale du taux de réponse final était enregistrée en 2008. La figure 3.1-2, présentée plus loin, montre la ventilation de la non-réponse en pourcentage de tout l'échantillon pour les années 2003 à 2008. Comme un nouveau panel est mis en place tous les trois ans, il est indiqué d'examiner la non-réponse en fonction de cycles de trois ans. À la figure 3.1-2, la catégorie du maximum atteint n'existe que pour les années 2007 et 2008. Il s'ensuit que les comparaisons entre les années sont faussées. La raison en est que la catégorie du maximum atteint comprend des non-répondants et des répondants potentiels.

[3] Certains ménages ne sont pas retournés sur le terrain en raison de certaines caractéristiques de non-réponse pendant un certain nombre d'années consécutives (deux années consécutives de refus, par exemple). Le fichier de l'échantillon est donc « épuré » pendant la durée de vie d'un panel, et les taux de réponse s'améliorent généralement. C'est ce que l'on appelle la *règle des deux ans*.

[4] La période de collecte de 2008 a duré 80 jours. Celle de 2007 a été prolongée à 102 jours.

Figure 3.1-1
Taux de réponse de 2008 et de 2007



Pour que les données soient comparables et pour jeter un certain éclairage sur la non-réponse générée par la limite des appels, une limite de 40 tentatives d'appel est simulée pour les années 2003 à 2006 (figure 3.1-3). Le groupe « pas de contact » semble avoir ressenti la plus forte incidence de cette limite. En fait, pour les années 2003 à 2006, les ménages qui ont nécessité plus de 40 tentatives d'appel étaient surtout des cas de répondants et de personnes avec lesquelles aucun contact n'avait été établi. Pendant la période de quatre ans, le pourcentage de cas de répondants a invariablement diminué, mais il a été compensé par une hausse du pourcentage des cas « pas de contact ».

L'examen de la figure 3.1-3 nous apprend que le pourcentage de refus a diminué sur un cycle de trois ans. Cette constatation n'est peut-être pas si surprenante toutefois : un certain nombre de ménages qui ont refusé de répondre ne sont plus contactés (voir la note à la page précédente concernant la *règle des deux ans*). Le taux de refus était plus élevé en 2008 qu'en 2005, ce qui peut indiquer que la limite des tentatives d'appel a restreint la possibilité de convertir les cas de refus, mais rien de tel n'a été observé pour 2007 et 2004. En outre, la baisse générale du nombre de cas de refus jumelée à la hausse du nombre de cas « pas de contact » et « autre non-réponse »[5] pourrait également indiquer une transition du simple refus vers le refus par évitement.

Nous avons également assisté à une augmentation du pourcentage de cas « hors du champ de l'enquête » (« hors du champ de l'enquête » n'est pas synonyme de « non-réponse »). Nous ne savons pas avec certitude si cela est révélateur d'un phénomène démographique ou si cela indique simplement un meilleur dépistage des ménages hors du champ de l'enquête.

La figure 3.1-3 montre également un renversement de la tendance au cours des trois dernières années en ce qui concerne le dépistage (ou les cas non résolus). Cette situation peut être expliquée en grande partie par une modification à la règle des deux ans en 2007 et en 2008, en vertu de laquelle les ménages qui ont fait l'objet d'un dépistage pendant deux années consécutives ont été contactés à nouveau. Ce changement, qui s'est traduit par un plus grand nombre de refus que de répondants et qui n'a permis qu'une petite hausse des taux de réponse, sera revu pour la collecte de 2009. La figure 3.1-4 montre que, en l'absence de cas supplémentaires, le pourcentage des cas de

[5] Les cas « autre non-réponse » comprennent les rendez-vous manqués, le chevauchement avec d'autres enquêtes, les interviews non réalisées en raison d'un obstacle linguistique ou de circonstances inhabituelles ou spéciales ou l'absence du ou des répondants longitudinaux pour la durée de l'enquête.

dépistage demeure relativement stable dans les trois dernières années, une légère amélioration étant enregistrée en 2008, peut-être en raison des meilleurs renseignements de dépistage.

Figure 3.1-2
Non-réponse à l'EDTR – 2003 à 2008

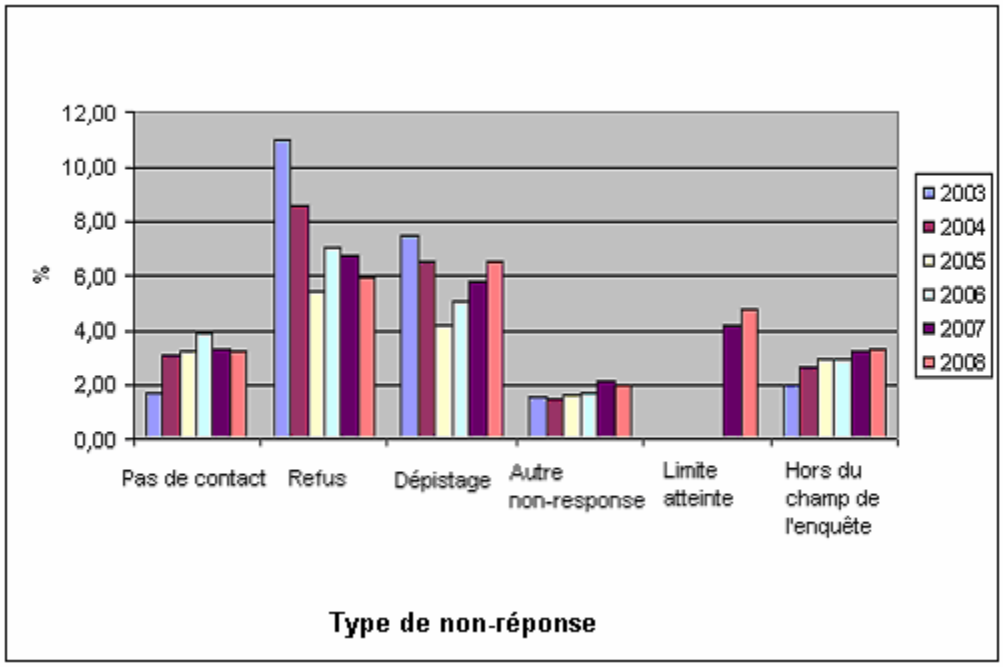
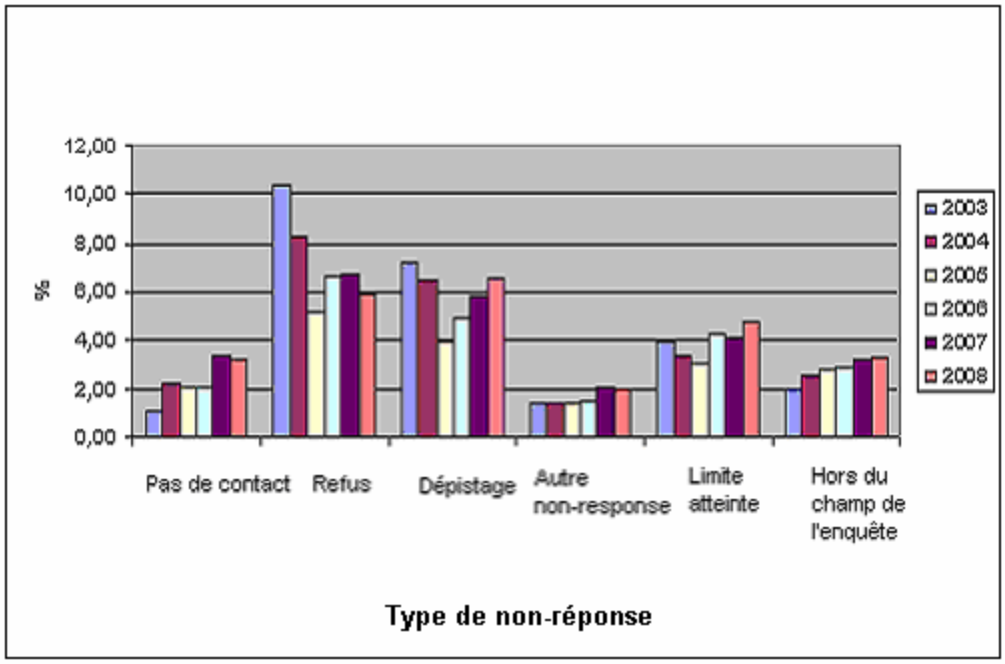


Figure 3.1-3
Non-réponse à l'EDTR, selon le type en supposant une limite de 40 appels chaque année



Pour avoir une idée de la façon dont les tendances observées relatives à la non-réponse pourraient influer sur notre capacité à faire des inférences concernant certaines sous-populations, examinons le pourcentage de personnes longitudinales associées aux ménages non-répondants de 2008, selon l'âge, présenté à la figure 3.1-5. Les refus

semblent survenir également dans tous les groupes d'âge. Cependant, les jeunes adultes semblent être plus souvent associés aux catégories « pas de contact », « limite atteinte » et « dépistage » que les autres membres de l'échantillon, et ils forment le groupe le plus touché par la non-réponse. Même si cette tendance n'indique pas la présence d'un biais, elle n'en révèle pas moins une préoccupation potentiellement croissante. Dans la prochaine section, nous examinons l'utilisation des paradonnées afin d'établir un modèle d'ajustement pour la non-réponse.

Figure 3.1-4
Dépistage dans le contexte de l'EDTR, ajusté pour la règle de deux ans

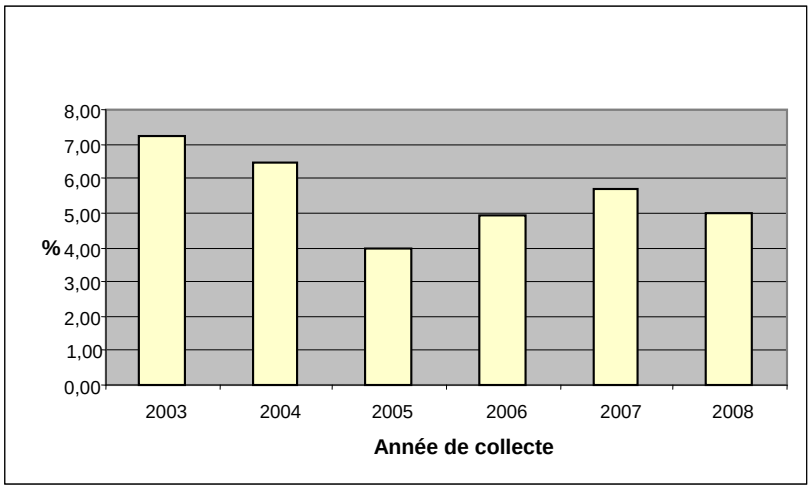
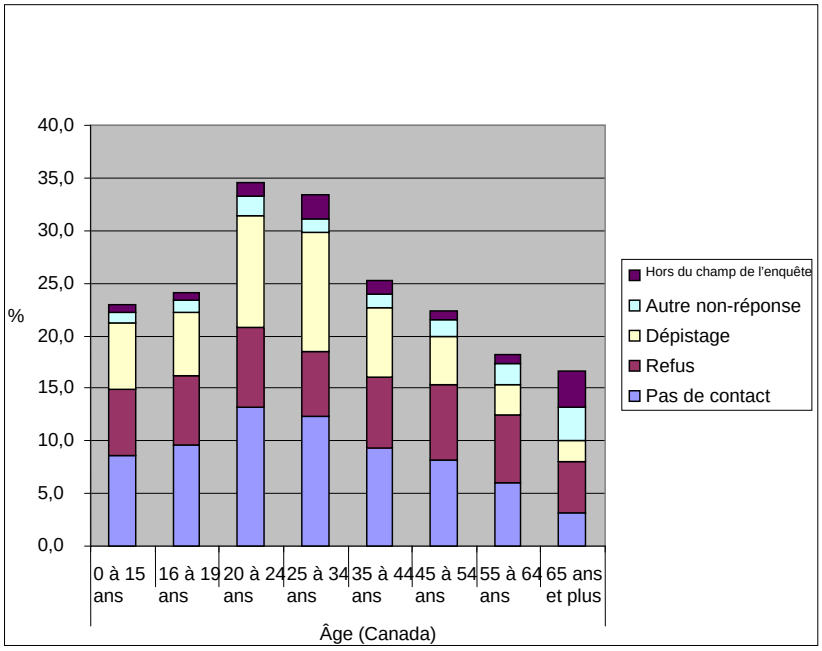


Figure 3.1-5
Non-réponse à l'EDTR de 2008, selon l'âge des répondants longitudinaux



4. Utilisation des paradonnées dans l'ajustement pour la non-réponse

Comme nous l'avons mentionné à la section 2 ci-dessus, l'échantillon de l'EDTR est un sous-échantillon des répondants à l'EPA du Canada. Nous disposons donc des caractéristiques initiales de base concernant ces

répondants (données démographiques et autres), qui sont actuellement utilisés pour modéliser la non-réponse dans le contexte de l'EDTR. L'inconvénient de ce processus est que, au fur et à mesure que le panel vieillit, certains renseignements initiaux deviennent de plus en plus désuets.

Au cours des dernières années, une mine de renseignements obtenus à l'étape de la collecte ont été rendus disponibles par entremise des fichiers de l'historique des transactions Blaise. Ces données sont plus actuelles et peuvent être étroitement corrélées aux variables de l'enquête qui nous intéressent. Dans les paragraphes qui suivent, nous proposons d'utiliser les données obtenues pendant la collecte (les paradonnées) pour améliorer le processus de modélisation de la non-réponse à l'EDTR. La présente partie se présente de la façon suivante. Aux sections 4.1 à 4.3, nous décrivons la méthode actuelle d'ajustement pour la non-réponse dans le cadre de l'EDTR. La section 4.4 donne une brève description d'une étude réalisée pour évaluer l'utilisation des paradonnées dans la modélisation de la non-réponse. Enfin, la section 4.5 présente un résumé des résultats.

4.1 Pondération et ajustement pour la non-réponse

L'ajustement pour la non-réponse fait partie du processus de pondération de l'EDTR. Chaque année, différents poids sont produits : un ensemble de poids longitudinaux par panel, un ensemble de poids longitudinaux pour les panels combinés et deux ensembles de poids transversaux. L'ajustement pour la non-réponse est effectué au niveau longitudinal et séparément pour chaque panel. Deux ajustements différents pour la non-réponse sont effectués : un pour la pondération longitudinale et un autre pour chaque panel puisque les répondants peuvent également représenter des non-répondants. Ces poids forment la base du schéma de pondération transversale. En combinant les panels, nous obtenons l'échantillon transversal, et une méthode de partage des poids est utilisée au sein du ménage pour calculer les poids concernant les cohabitants nouveaux dans le ménage. L'ajustement pour la non-réponse constitue la première étape du processus de pondération. Elle est toujours réalisée de façon indépendante et pour chaque panel.

4.2 Méthode actuelle concernant la non-réponse à l'EDTR

Les unités longitudinales sont classées soit comme des répondants, soit comme des non-répondants, soit comme des unités inadmissibles. Une unité inadmissible serait une personne qui a déménagé à l'extérieur du pays, qui a déménagé dans un établissement ou qui est décédée. Les unités inadmissibles ne sont pas prises en compte dans l'exercice de modélisation de la non-réponse. De même, les non-répondants faisant partie d'un ménage répondant ne sont pas pris en compte, même s'ils sont assimilés à des répondants dans le processus de modélisation de la non-réponse.

Les variables suivantes recueillies dans le contexte de l'EPA sont utilisées dans le processus de modélisation de la non-réponse : âge du répondant, état matrimonial, catégorie de revenu, statut d'emploi, taille du ménage, incapacité, etc. Il s'agit de variables binaires dans la plupart des cas, qui décrivent la personne au moment de la sélection. La technique utilisée pour la modélisation de la non-réponse s'appelle « modélisation par segmentation », et elle est fondée sur l'algorithme de détection de l'interaction automatique du chi carré. Dans le cadre de cette méthode, les variables les plus significatives expliquant la réponse à l'enquête sont choisies. On y arrive en comparant les valeurs du chi carré de Pearson entre les variables et en retenant les variables produisant les valeurs les plus élevées. Dès qu'une variable est sélectionnée, l'ensemble des personnes est divisé en deux groupes selon la variable choisie. La même procédure est appliquée à chaque sous-ensemble et aux sous-ensembles des sous-ensembles. Aucun groupe n'est créé si les variables ne sont pas corrélées à la réponse ou si le groupe obtenu est trop petit. Ce processus produit clairement un arbre, et il est stoppé dès que le groupe se situant dans les branches du bas ne peut plus être divisé. Nous obtenons alors des groupes de réponse homogènes (GRH).

Le taux de réponse pondéré (TRP) est évalué pour chaque GRH. Le poids du répondant est multiplié par l'inverse du TRP, tandis que le poids du non-répondant est fixé à zéro.

4.3 Méthode d'ajustement pour la non-réponse

L'actuelle méthode concernant la non-réponse utilise des données réunies au début du cycle de vie du panel, Il aurait aussi été possible d'utiliser la dernière vague de la collecte de données afin de modéliser la non-réponse pour la vague courante. Une autre méthode étroitement liée à la seconde consiste à introduire des paradonnées dans le modèle actuel.

Comme on peut le voir dans Watson et Wooden (2006), nous tenons compte de variables qui influent sur la prise de contact et de variables qui mesurent la probabilité que, une fois contactée, la personne se prêtera à une interview. D'un point de vue pratique, nous tiendrons compte des variables comme le changement d'adresse depuis la dernière vague, le nombre de tentatives faites pour joindre le répondant, le nombre total de contacts effectués, le nombre total d'appels effectués, la tranche horaire pendant laquelle le cas a été joint (jour et heure) ainsi que des variables démographiques de l'EPA et des variables du modèle actuel, comme l'âge, le sexe, l'état matrimonial, la taille du ménage, la province de résidence, la présence de jeunes enfants, l'emploi, le statut de propriétaire ou de locataire, le type de logement, etc.

L'équation ci-dessous montre le modèle de régression logistique utilisé. Ce modèle, ainsi que l'option de régression pas à pas (Beaumont, 2005), a été utilisé pour incorporer les variables les plus influentes dans le modèle :

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \sum_{k=1}^{t} \beta_k x_{ik}$$

où θ_i est la probabilité estimée de réponse pour l'unité, tandis que i et β_k sont des paramètres de régression à estimer et des variables auxiliaires respectivement. Le même modèle de régression logistique convenait à trois ensembles différents de variables, que nous appelons « modèle I », « modèle II » et « modèle III ». Dans le modèle I, seules les variables de l'EPA ont été utilisées. Seules les variables des paradonnées ont été utilisées dans le modèle II et, enfin, les variables des paradonnées et de l'EPA ont été regroupées dans le modèle III. La procédure Cluster du SAS ainsi que la procédure Tree ont alors été utilisées pour former les GRH.

4.4 Évaluation des modèles

Diverses statistiques ont été calculées au moyen de méthodes comme les tests chi carré de Pearson et le test de Hosmer-Lemeshow pour évaluer la validité des modèles. Pour mesurer l'incidence du modèle de non-réponse sur les estimations, nous formulons l'hypothèse que, suivant en cela Singh et coll. (1995), l'estimation de la première vague constitue l'estimation repère. Puis, uniquement à l'aide des unités longitudinales de l'enquête et pour chaque source de données (modèle I, modèle II et modèle III), nous comparons les estimations pondérées de chaque vague successive, calculées à l'aide du poids de la vague actuelle et des valeurs y de la première vague, à celles de la première vague, comme le montre la formule ci-dessous. La quantité obtenue est appelée « pseudo-biais relatif » pour des raisons évidentes :

$$pseudo\_rel\_bias = \left(\frac{\hat{Y}_{1,S}^{(wi)} - \hat{Y}_{1,LFS}^{(w1)}}{\hat{Y}_{1,LFS}^{(w1)}}\right) * 100$$

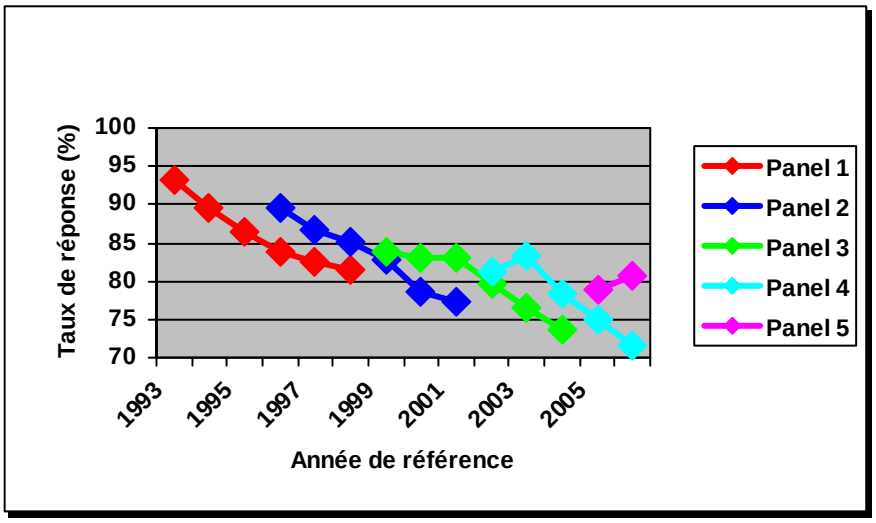
où $\hat{Y}_{1,S}^{(wi)}$ est une estimation pour une variable d'intérêt particulière (revenu, salaire, loyer, etc.) et une source ***S*** donnée (EPA, paradonnées, combinaison des deux) utilisant le poids de la vague i (wi) et des *valeurs y* de la vague 1.

4.5 Données utilisées

Les données concernant le panel 4 de l'EDTR ont été utilisées puisqu'il a été mis en place pour la première fois en 2002 et que la dernière vague de collecte de données concernant ce panel remonte au début de 2008. La taille de

l'échantillon du panel 4 est passée d'environ 34 000 à la vague 1 à 30 000 à la vague 5. La figure 4.5-1 ci-dessous montre les taux de réponse longitudinaux au fil des ans. Nous pouvons constater que les taux de réponse non seulement diminuent pendant la durée de vie du panel mais également au fil des ans.

Figure 4.5-1
Taux de réponse longitudinal par vague et par panel



4.6 Résultats

Les résultats concernant le revenu total et les loyers versés sont présentés aux figures 4.6-1 et 4.6-2 ci-dessous. Le pseudo-biais est indiqué pour les vagues 2 à 5. Nous pouvons voir que les sources combinées et les paradonnées affichent de façon constante des pseudo-biais plus petits. En ce qui concerne les variables retenues, le modèle de non-réponse utilisant les variables des paradonnées semble être plus près des estimations de la première vague (ce qui est en un certain sens la valeur la plus rapprochée du paramètre de la population à estimer que nous puissions obtenir).

Figure 4.6-1
Pseudo-biais relatif par source de données : revenu total versés

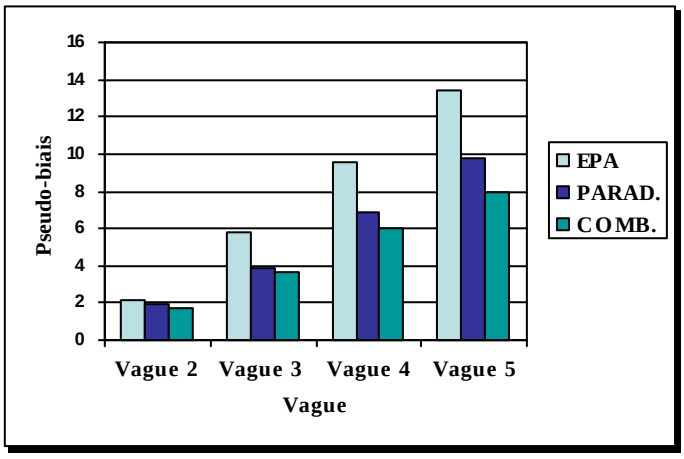
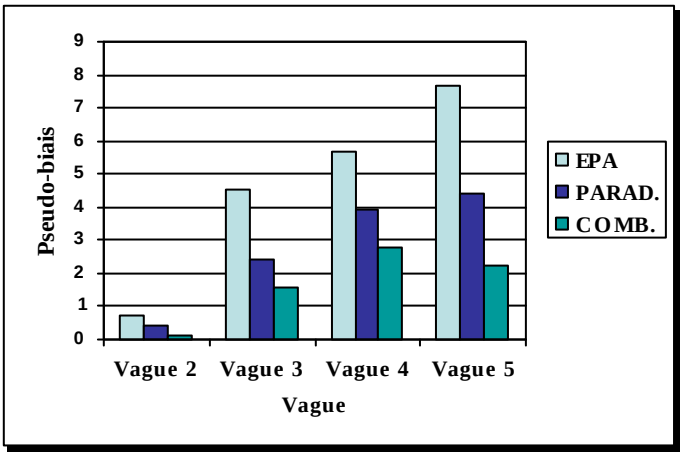


Figure 4.6-2:
Pseudo-biais relatif par source de données : loyers



5. Conclusion et travaux futurs

L'examen des paradonnées a conduit à des améliorations de la gestion active des activités de collecte de l'EDTR et, éventuellement, de la modélisation de la non-réponse. En fonction des résultats de la collecte de 2008, nous devons continuer à approfondir notre compréhension du processus de collecte ainsi que notre capacité à nous y adapter afin d'utiliser plus efficacement les ressources consacrées à la collecte.

De nouvelles mesures, comme des sources de dépistage supplémentaire, ont été prises en 2008, et elles seront évaluées sous peu. En outre, nous commencerons à recueillir en 2009 les numéros de téléphone cellulaire et les adresses de courriel, deux éléments d'information non susceptible de changer à l'occasion d'un déménagement. Pour ce qui est de la non-réponse, l'étude actuelle sera élargie de façon à englober d'autres variables de l'EDTR et à évaluer l'incidence potentielle de la nouvelle méthode sur les estimations (longitudinales et transversales). D'autres modèles seront examinés pour différents types de non-réponse (répondant non dépisté, refus catégorique, etc.).

Cet automne, Statistique Canada mènera une nouvelle enquête pilote, l'enquête Vivre au Canada, afin d'étudier l'influence réciproque de l'état de santé, de la formation et de la dissolution des familles, de la dynamique du travail, de l'évolution du capital humain et social ainsi que des effets géographiques pour mieux comprendre les « parcours de vie ». L'objectif est d'élargir le contenu de l'actuelle EDTR et de suivre le panel longitudinal pendant une période indéfinie. Dans ce contexte, la compréhension et l'atténuation des effets de la non-réponse revêtiront une importance majeure.

Bibliographie

Beaumont, J-F. (2005). L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids. *Techniques d'enquête*, 31, 249-254.

Chapman, B. (2007a). BTH File Analysis of SLID Collection 2003-2007 by Panel, document interne, Statistique Canada.

Chapman, B. (2007b). BTH File Analysis of SLID Collection 2005-2007, document interne, Statistique Canada.

Hunter, L. et Carbonneau, J.-F. (2005). Une méthode de collecte de données d'enquête axée sur la gestion active, *Recueil: Symposium 2005, Défis méthodologiques reliés aux besoins futures d'information*. Statistique Canada

Laflamme, F., Maydan, M. et Miller, A. (2008), Using Paradata to Actively Manage Data Collection. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Laflamme, F. (2008), Understanding Survey Data Collection through the Analysis of Paradata at Statistics Canada. *Proceedings of the American Association for Public Opinion Research Section on Survey Research Methods*, American Statistical Association.

Lévesque, A. et Poulin, J. (2007). Collection Activities Management for the Survey of Labour and Income Dynamics: 2005 to 2007 Evaluations, document interne, Statistique Canada.

Scheuren, F. (2005). Les paradonnées de la conception à la réalisation, *Recueil: Symposium 2005, Défis méthodologiques reliés aux besoins futures d'information*. Statistique Canada

Singh, A.C., Wu, S., et Boyer, R. (1995). Longitudinal Survey Nonresponse Adjustment by Weight Calibration for Estimation of Gross Flows. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 396-401.

Watson, N. et Wooden, M. (2006). Modelling Longitudinal Survey Response: The Experience of the HILDA Survey, Hilda Project Discussion Paper Series no.2/06, Australie.