N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

ÉVALUATION ET AJUSTEMENT POUR LA NON-RÉPONSE À L'ENQUÊTE SUR LA POPULATION ACTIVE DU CANADA

Asma Alavi et Jean-François Beaumont[1]

RÉSUMÉ

Même si l'on adopte un bon plan d'échantillonnage et que les intervieweurs s'efforcent par tous les moyens d'éviter la non-réponse, celle-ci est incontournable lors de la réalisation d'enquêtes à grande échelle. Quoiqu'il soit assez faible, le taux de non-réponse à l'Enquête sur la population active (EPA) est surveillé de près et de meilleurs moyens de corriger pour la non-réponse sont étudiés périodiquement. L'article porte sur ces deux aspects. Nous examinons certaines statistiques descriptives permettant d'évaluer la non-réponse à l'EPA, ainsi que des moyens d'améliorer la méthode courante d'ajustement pour la non-réponse. Afin de compenser les effets de cette dernière, on procède à la repondération des répondants de façon à tenir compte de la partie non répondante de l'échantillon. La méthode de repondération la plus courante consiste à créer des classes de pondération fondées sur les variables du plan de sondage et à calculer un facteur de correction pour la non-réponse. Notre approche s'appuie sur la modélisation par la régression logistique pour former des classes d'ajustement pour la non-réponse. Nous modélisons les ménages répondants d'après l'information sur la prise de contact, comme le nombre de tentatives en vue de prendre contact avec une unité d'échantillonnage et l'heure où elles ont eu lieu, conjuguée à l'information tirée du plan de sondage. Les résultats témoignent d'une amélioration significative par rapport à la méthode utilisée à l'heure actuelle pour l'EPA, à savoir la réduction du nombre de classes et un meilleur ajustement du modèle.

MOTS CLÉS : Classes d'ajustement pour la non-réponse, information sur la prise de contact, modélisation, régression logistique, repondération.

1. INTRODUCTION

La non-réponse et les méthodes visant à la compenser sont maintenant des éléments dont on tient compte régulièrement en échantillonnage. La situation idéale où toutes les unités échantillonnées répondent a fort peu de pertinence en pratique. L'importance grandissante des enquêtes par sondage a accru le fardeau de réponse et, donc, la prévalence de la non-réponse. Il y a non-réponse à une enquête quand, quelle qu'en soit la raison, une unité sélectionnée ne répond pas. Les méthodes habituelles d'estimation en présence de non-réponse donnent des résultats entachés d'un biais, car, en général, les non-répondants diffèrent des répondants en ce qui concerne les variables d'intérêt.

Le meilleur moyen de lutter contre la non-réponse consiste à faire, aux étapes de la conception et du développement de l'enquête, tous les efforts possibles en vue de l'éviter, comme la mise en place de programmes de suivi et de rappel. Toutefois, le coût de cette approche en ressources tant humaines que financières est élevé. L'autre possibilité consiste à appliquer des méthodes complexes et approfondies visant la collecte des données et l'estimation, comme le sous-échantillonnage des répondants et la randomisation de la réponse, qui rendent l'effet de la non-réponse négligeable. Toutefois, ces méthodes posent aussi des problèmes de ressources. Donc, le scénario utilisé le plus fréquemment consiste à traiter la non-réponse, une fois qu'elle a été observée, de façon à obtenir des estimateurs ne contenant pas un biais trop prononcé. Sous ce scénario, qui est appliqué après la phase de collecte des données, le but est d'obtenir un ensemble complet de données, puis d'utiliser des méthodes d'estimation typiques.

[1] Asma Alavi et Jean-François Beaumont, Statistique Canada, immeuble R.-H.-Coats, 16e étage, pré Tunney, 120 avenue Parkdale, Ottawa, Ontario, Canada, K1A 0T6, asma.alavi@statcan.ca, Jean-Francois.Beaumont@statcan.ca

Les deux grands types de non-réponse sont la non-réponse totale et la non-réponse partielle. Il y a non-réponse totale quand, pour diverses raisons, l'unité échantillonnée ne fournit pas de réponse. Donc, mise à part l'information tirée du plan de sondage et celle sur la prise de contact, on ne possède aucune donnée sur l'unité. Par non-réponse partielle, on entend les cas où des données manquent pour au moins une composante de l'enquête, mais non toutes, pour une unité particulière. Dans le présent article, nous n'examinons que la non-réponse totale.

2. CONTEXTE

L'Enquête sur la population active du Canada est une enquête mensuelle avec renouvellement de panel dans le cadre de laquelle des données sont recueillies chaque mois sur les occupants d'environ 54 000 logements. La population cible comprend tous les membres de la population civile non placés en établissement ayant au moins 15 ans qui vivent dans les dix provinces du Canada. Les logements sélectionnés sont retenus dans l'échantillon pendant six mois consécutifs. L'EPA est la seule source officielle de statistiques sur la population active, comme les taux nationaux et provinciaux de chômage. Les logements visés par l'EPA sont sélectionnés conformément à un plan d'échantillonnage stratifié à plusieurs degrés. Le premier degré d'échantillonnage consiste à sélectionner de petites régions géographiques, appelées grappes, à l'intérieur de chaque strate. Le deuxième degré d'échantillonnage consiste à sélectionner des logements à l'intérieur de chaque grappe échantillonnée. Les données de l'EPA sont recueillies pour tous les membres admissibles du ménage résidant dans les logements sélectionnés, en recourant à l'interview par procuration.

L'EPA se fait par interview assistée par ordinateur (IAO), aussi bien téléphonique (ITAO) que sur place (IPAO). L'information sur la prise de contact, comme le nombre de tentatives en vue de prendre contact avec les occupants d'un logement et le moment de la journée où ont lieu les tentatives, est enregistrée pour chaque logement. Une fois recueillies, les données sont vérifiées pour repérer toute discordance ou omission. Dans le cas de logements non répondants pour lesquels des données ont été recueillies lors d'un mois antérieur, on procède à l'imputation des enregistrements par la méthode de la valeur précédente ou par la méthode hot-deck. Puis, on procède à la repondération des logements répondants afin de compenser pour les autres logements non répondants. Cette repondération est fondée sur l'hypothèse selon laquelle les logements répondants et non répondants ont les mêmes caractéristiques à l'intérieur des classes de repondération ou d'ajustement pour la non-réponse.

À l'heure actuelle, la stratégie de repondération de l'EPA consiste à créer des classes d'ajustement pour la non-réponse, appelées secteurs de non-réponse, d'après l'information provenant du plan de sondage. Les variables du plan de sondage sont les strates de revenu élevé, les régions économiques d'assurance-emploi, le type de strate et le numéro de groupe de renouvellement du logement. Chaque strate de revenu élevé constitue un secteur de non-réponse, et le recoupement des trois autres variables du plan de sondage forme les autres secteurs de non-réponse. À l'heure actuelle, il existe environ 900 secteurs de non-réponse dans le cas de l'EPA. Pour chacun de ces secteurs, on détermine le taux de réponse observé pondéré, puis on divise le poids de sondage des logements répondants par ce taux. De cette manière, les poids des logements répondants sont ajustés à la hausse pour tenir compte des cas de non-réponse. Ensuite, les logements non répondants traités par repondération sont éliminés de l'échantillon.

Bien que la méthode actuelle de repondération semble satisfaisante en général, il est toujours possible de la perfectionner et de l'améliorer. Deux imperfections de la méthode sont la création d'un grand nombre de classes, comme nous le mentionnons plus haut, et la non-utilisation de l'information sur les prises de contact. L'un des inconvénients que présente un grand nombre de classes est l'augmentation de la probabilité qu'un faible taux de réponse à l'intérieur d'une classe produise une valeur élevée du poids associé aux logements répondants dans cette classe d'ajustement pour la non-réponse. L'application de poids élevés aux ménages pourrait accroître considérablement la variance d'échantillonnage des estimateurs. Pour le moment, les classes d'ajustement pour la non-réponse dont le facteur de correction est supérieur à 2 (taux de réponse inférieur à 50 %) sont fusionnées pour réduire la taille de ce facteur et contrôler la valeur des poids. Un autre problème que pose un grand nombre de classes est l'instabilité des résultats, puisqu'un petit changement dans la configuration des classes peut altérer

considérablement les poids. L'inconvénient qu'il y a à ne pas utiliser l'information sur les prises de contact est le gaspillage éventuel d'information étroitement associée au processus de réponse ou de non-réponse à l'enquête.

3. ÉVALUATION DE LA NON-RÉPONSE

L'évolution de la non-réponse au cours du temps a été observée de façon systématique dans le cadre du processus de surveillance de la qualité de l'EPA. Le taux de non-réponse à l'EPA est assez faible et dépend de divers facteurs. Une étude descriptive entreprise pour évaluer la non-réponse et les questions connexes a montré que le taux de non-réponse à l'EPA a augmenté avec le temps et que, même s'il est maintenant à la baisse, il est encore supérieur au taux de non-réponse historique. Il semble donc important d'élaborer une meilleure stratégie d'ajustement pour la non-réponse afin de réduire au minimum les effets de cette dernière sur les estimations d'enquête.

L'étude descriptive visait à examiner, entre autres, les moyens d'améliorer le taux de réponse à l'EPA. L'objectif du projet, qui consistait à optimiser le profil de réponse, était triple. Il s'agissait de déterminer le nombre optimal de tentatives de prise de contact, le meilleur moment de la journée et le meilleur jour de la semaine pour obtenir une réponse (interview complète ou partielle) auprès d'un ménage. Nous ne présentons ici que certains résultats de cette étude.

En général, le pourcentage de logements pour lesquels on ne dénombre qu'une seule tentative est d'environ 38 % et le pourcentage pour lesquels le nombre de tentatives est supérieur à 10 est d'environ 5 %. En outre, le nombre de tentatives peut aller jusqu'à 70. Presque 38 % des logements échantillonnés ont répondu au moment de la première et seule prise de contact. L'étude confirme le sentiment que le taux de réponse diminue à mesure que le nombre de tentatives augmente. Par exemple, pour 30 tentatives ou plus, le taux de réponse s'approche de 20 %, alors que pour une tentative ou plus, il est d'environ 95 %. Nous avons également examiné la composition du ménage et les caractéristiques démographiques des ménages pour lesquels le nombre de tentatives de prise de contact était inférieur ou égal à 10 et pour ceux pour lesquels il était supérieur à 10. Les variables étudiées sont la situation d'activité (occupé, chômeur et inactif), le sexe, l'âge et la taille du ménage. Dans l'ensemble, l'étude indique qu'un plus grand nombre de tentatives sont nécessaires pour les hommes, les petits ménages (une ou deux personnes), les personnes occupées et les personnes jeunes (de 15 à 55 ans).

Nous avons également étudié l'effet de la limitation des efforts en vue de prendre contact avec les occupants d'un logement à un nombre donné de tentatives. Nous avons comparé les estimations du nombre de personnes occupées, chômeuses et inactives et des taux de chômage calculés d'après les données originales de l'EPA et d'après les données de l'EPA avec cas de non-réponse ajoutés (à cause de la limitation des tentatives de prise de contact à un nombre déterminé). Nous avons pris les données de l'EPA et considéré tous les logements pour lesquels le nombre de tentatives de prise de contact était supérieur à 10 comme des cas de non-réponse, puis nous avons procédé au rajustement et au calage habituels pour tenir compte de la non-réponse afin d'obtenir les estimations souhaitées. Pour les grands domaines, comme les provinces, la différence entre les estimations habituelles de l'EPA (nombre de chômeurs, nombre de personnes occupées, taux de chômage, etc.) et les nouvelles estimations n'était pas significative. Mais, pour certains domaines plus petits, l'effet était assez important. En valeur absolue, l'écart relatif en pourcentage entre les taux de chômage produits par les deux méthodes allait jusqu'à 80 %.

L'étude descriptive a renforcé le sentiment que l'information provenant du processus d'interview est très utile pour expliquer le comportement de réponse des ménages occupant les logements sélectionnés.

4. APPROCHE DE MODÉLISATION

Quelques remaniements de l'EPA ont déjà eu lieu depuis l'adoption de la méthode utilisée à l'heure actuelle. Un remaniement de l'EPA est de rigueur après chaque recensement décennal de la population. Le remaniement entrepris après le dernier recensement, qui a eu lieu en 2001, est en cours. Cet exercice offre une bonne occasion de

redéfinir les méthodes de repondération utilisées pour l'EPA, puisque nous disposons maintenant d'information supplémentaire sur la collecte des données qui peut être utilisée comme information auxiliaire.

Nous proposons l'utilisation de classes de modélisation (voir, à titre de référence, Little 1986, Eltinge et Yansaneh, 1997 et Haziza et coll., 2001) pour déterminer les classes d'ajustement pour la non-réponse ou d'ajustement de la pondération. Puisque la variable de réponse est binaire (un logement répond ou non), nous utilisons la régression logistique. Au lieu de nous servir simplement de l'information tirée du plan de sondage tel que mentionné plus haut, nous ajoutons deux variables, à savoir le nombre de tentatives de prise de contact avec les occupants d'un logement et l'heure du début de la dernière tentative. Donc,

$$\log\left(\frac{\hat{p}(\mathbf{x})}{1-\hat{p}(\mathbf{x})}\right)=\hat{\boldsymbol{\beta}}'\mathbf{x}\,, \tag{4.1}$$

où $\hat{p}(\mathbf{x})$ est la moyenne conditionnelle estimée de la variable de réponse (probabilité de répondre ou propension à répondre) étant donné $\mathbf{x}$, le vecteur de variables auxiliaires. Nous donnons à cette approche le nom de modélisation en une étape.

Nous partons d'un modèle contenant six effets principaux, à savoir la province (10 catégories), l'heure du début de la dernière tentative (5 catégories), le nombre de tentatives (5 catégories), le type de strate (9 catégories), le groupe de renouvellement (6 catégories), le fait que la strate soit à revenu élevé ou non (2 catégories) et toutes les interactions de premier ordre (15). Nous utilisons SAS pour effectuer une régression pas à pas, ou régression stepwise, afin de choisir le modèle. Nous répétons le processus en utilisant les données de l'EPA recueillies pour plusieurs mois. Nous choisissons comme modèle final celui contenant les effets principaux et les interactions présentes et les plus importantes lors de divers mois. Le modèle final contient les cinq effets principaux (voir l'annexe), non compris la variable de strate de revenu élevé et quatre interactions (l'interaction entre le nombre de tentatives et le groupe de renouvellement, l'interaction entre le nombre de tentatives et l'heure à laquelle a débuté la dernière tentative, l'interaction entre le nombre de tentatives et la province et l'interaction entre la province et l'heure à laquelle a débuté la dernière tentative).

Après la sélection du modèle logistique final, nous calculons les probabilités de réponse estimées pour chaque logement. Nous appliquons PROC FASTCLUS dans SAS pour former des classes d'ajustement pour la non-réponse homogènes par rapport aux probabilités de réponse estimées. Nous conditionnons le processus de façon à ce que chaque classe contienne au moins 20 logements répondants. Puis, dans chaque classe, nous calculons un taux de réponse pondéré pour obtenir le facteur de correction pour la non-réponse. Nous éliminons de l'échantillon les logements non répondants et nous rajustons les poids des logements répondants en les multipliant par les facteurs de correction pour la non-réponse correspondants. Nous créons des classes d'ajustement pour la non-réponse permettant d'obtenir des résultats robustes à l'échec du modèle. Parallèlement, nous voulons retenir le pouvoir prédictif élevé du modèle original.

Nous étudions également l'effet de la modélisation distincte de la probabilité d'une prise de contact et de la probabilité de réponse, étant donné une prise de contact. Selon cette approche en deux étapes, la probabilité finale de non-réponse prend la forme du produit des probabilités calculées d'après les deux modèles distincts. La première étape est la régression logistique des prises de contact sur diverses variables auxiliaires comparables à celles définies pour la modélisation de la réponse en une étape, et la deuxième est la modélisation distincte de la réponse, étant donné une prise de contact. Ensuite nous comparons en détail les résultats de la modélisation en deux étapes à ceux de la modélisation en une étape. Bien que le modèle en deux étapes donne de meilleurs résultats en ce qui concerne le R^2 maximal rééchelonné (coefficient de détermination généralisé) de Cox-Snell (voir Cox et Snell, (1989)), le pouvoir prédictif, etc., les écarts entre les deux méthodes ne sont pas importants, particulièrement si l'on tient compte de la complexité de la modélisation en deux étapes. Par conséquent, nous ne considérerons ici que le modèle en une étape.

5. RÉSULTATS

À la présente section, nous comparons les résultats obtenus en utilisant la méthode courante d'ajustement pour la non-réponse à l'EPA à celle fondée sur la modélisation décrite à la section 4. Nous avons obtenu plusieurs mesures diagnostiques pour comparer la méthode courante (en supposant implicitement qu'il s'agit aussi d'une méthode de modélisation où les secteurs de non-réponse correspondent aux classes) à la méthode d'ajustement pour la non-réponse par modélisation proposée pour l'EPA. Ces mesures diagnostiques incluent le coefficient généralisé de détermination (R^2), le test de qualité de l'ajustement de Hosmer-Lemeshow et la distribution du facteur de correction pour la non-réponse. Nous obtenons la statistique de qualité de l'ajustement de Hosmer-Lemeshow en calculant la statistique du chi carré de Pearson à partir de la table 2×10 des fréquences observées et prévues, où 10 est le nombre de groupes utilisés dans SAS. Le tableau 1 présente les valeurs de R^2 pour les méthodes courante et par modélisation pour certains mois, ainsi que le facteur de correction pour la non-réponse le plus important.

Hosmer et Lemeshow (1980) ont proposé un test de qualité de l'ajustement pour des données binaires dans un cadre de régression logistique. Pour les données provenant de divers mois, nous avons testé la qualité de l'ajustement du modèle logistique final et constaté, dans 91 % des cas, que le modèle est assez bien ajusté. Par exemple, pour octobre 2001, la valeur de la statistique de test d'Hosmer-Lemeshow est de 4,7362 avec 8 degrés de liberté et une valeur p de 0,7854.

Tableau 1 : Diagnostics pour les méthodes courante et par modélisation

	R^2 de Cox-Snell		Facteur maximal de correction pour la non-réponse	
	Courante	Modélisation	Courante	Modélisation
Mars 2001	0,1242	0,2085	3,6667	2,8323
Juin 2001	0,1209	0,2145	1,8458	2,0850
Octobre 2001	0,1312	0,2355	1,6667	2,4301

La méthode d'ajustement pour la non-réponse à l'EPA par modélisation semble mieux expliquer le profil de non-réponse qui se dégage des données observées. Les variables auxiliaires fondées sur l'information sur les prises de contact sont fort significatives et améliorent le modèle. L'étude a montré que les méthodes courante et par modélisation produisent des estimations identiques des taux de réponse observés pour les catégories de variables géographiques et de variables de plan de sondage comme la province, le groupe de renouvellement et le type de strate. Par ailleurs, comme l'indiquent les figures 1 et 2, la méthode par modélisation produit des estimations des taux de réponse observés de haute précision pour les variables de prise de contact, contrairement à la méthode courante. La figure 1 montre ces différences pour les catégories de nombre de tentatives de prise de contact, allant du plus faible au plus grand (consulter l'annexe pour des précisions) et la figure 2 donne les résultats pour les provinces du Canada, d'est en ouest. Comme son nom l'indique, le taux de réponse observé est le taux de réponse observé pour diverses catégories d'une variable donnée, d'après les données recueillies pour juin 2001.

Nous avons produit les taux de chômage nationaux et provinciaux pour une série de mois au moyen des données de l'EPA en corrigeant les poids pour la non-réponse selon la méthode courante et la méthode par modélisation. Nous avons calculé les écarts relatifs entre les taux de chômage obtenus pour les deux méthodes d'ajustement pour la non-réponse, en prenant comme valeur de référence les taux de chômage obtenus par la méthode courante. Sur la période observée, les écarts relatifs varient de –1,4 % à 1,4 %. Donc, les deux méthodes produisent des taux de chômage dont les valeurs sont très rapprochées. Nous avons obtenu des résultats comparables après avoir calé les poids ajustés pour la non-réponse. Il en est vraisemblablement ainsi parce que le taux de non-réponse à l'EPA est faible. S'il était plus élevé, nous observerions sans doute des écarts plus importants.

Figure 1 : Taux de réponse observé comparativement aux probabilités de réponse estimées par les méthodes courante et par modélisation pour la variable de nombre de tentatives de prise de contact

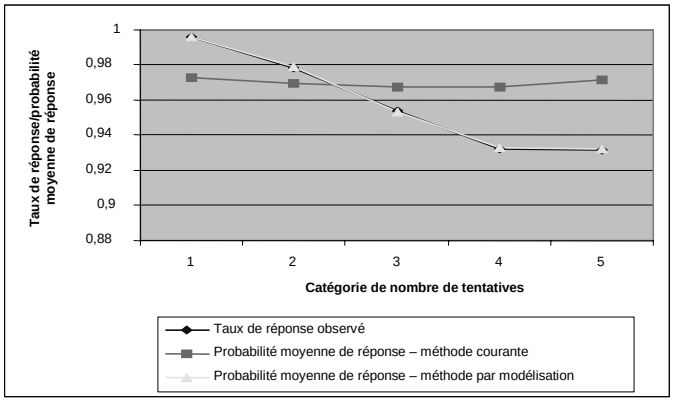
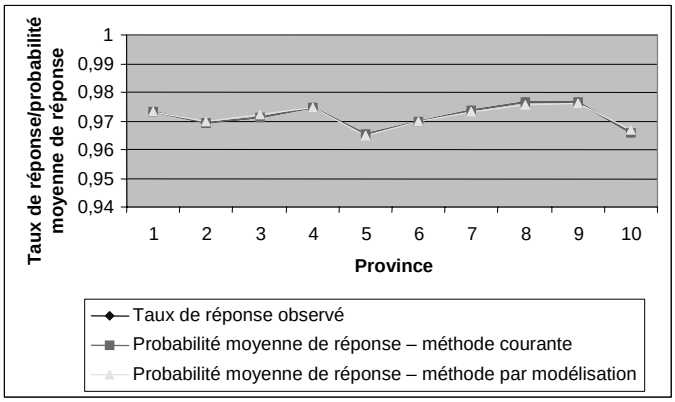


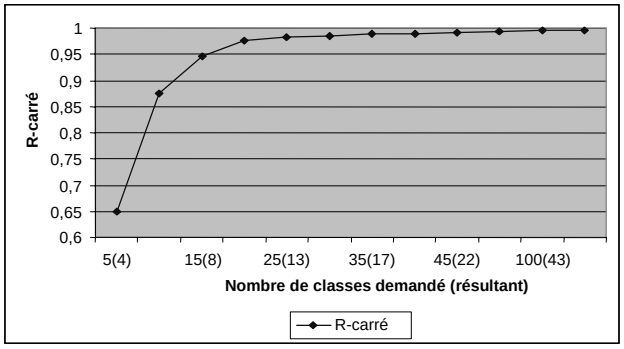
Figure 2 : Taux de réponse observé comparativement aux probabilités de réponse estimées par les méthodes courante et par modélisation pour les provinces



Comme nous l'avons mentionné plus haut, le nombre de classes résultant de l'application de la méthode courante d'ajustement pour la non-réponse est d'environ 900 par mois. Par contre, le nombre moyen de classes qui résulte de l'application de la méthode par modélisation est d'environ 50 par mois. La figure 3 montre l'effet de l'augmentation du nombre de classes d'ajustement pour la non-réponse sur $\tilde{R}^2$, c'est-à-dire le coefficient de détermination du modèle dans lequel la probabilité de réponse estimée pour chaque logement est la variable dépendante et la classe, la variable indépendante, pour un mois (juin 2001). Ce graphique illustre le degré d'homogénéité des classes d'ajustement pour la non-réponse. En ce qui concerne les probabilités de réponse estimées, nous utilisons le terme « demandé » pour le nombre de classes demandé dans le programme SAS et le terme « résultant » pour le nombre de classes produit par SAS. Il est évident que de 40 à 50 classes donnent une homogénéité suffisante.

Nous avons examiné la variabilité du facteur de correction de la non-réponse pour la repondération résultant de l'application des deux approches aux données de l'EPA recueillies pour divers mois. Par exemple, pour un mois, nous avons obtenu une variance plus faible pour le facteur calculé d'après la méthode courante que pour celui calculé d'après la méthode par modélisation (0,0064 comparativement à 0,0085), mais l'inverse pour l'étendue (3,05 comparativement à 1,39). Le profil est typique pour la gamme de mois considérés. Il n'est pas étonnant que la méthode par modélisation ait un meilleur pouvoir prédictif et produise des facteurs de correction pour la non-réponse dont la variabilité est forte, mais dont l'étendue est plus restreinte.

Figure 3 : R-carré entre la probabilité estimée de réponse et les classes d'ajustement pour la non-réponse fondées sur ces probabilités



Nous avons fait une analyse de variance en prenant la moyenne estimée des probabilités de réponse individuelles (probabilités de réponse estimées) comme variable dépendante et la situation d'activité des individus (employés, chômeurs ou inactifs), c'est-à-dire la variable présentant le plus d'intérêt dans l'EPA, comme variable explicative. Nous avons constaté que la corrélation des moyennes estimées des probabilités de réponse obtenues selon la méthode courante ou la méthode par modélisation n'est pas très forte. Nous avons calculé le coefficient de détermination $\tilde{R}^2$, qui est le carré du coefficient de corrélation linéaire, entre les probabilités estimées de réponse et la moyenne des probabilités de réponse estimées pour chaque niveau de situation d'activité. Par exemple, pour octobre 2001, les valeurs de $\tilde{R}^2$ sont respectivement 0,0000 et 0,0025 pour l'analyse fondée sur les méthodes courante et par modélisation. Le test F résultant de l'analyse de variance a une valeur p de 0,2048 pour la méthode courante et de 0,0001 pour la méthode par modélisation, ce qui donne à penser que la valeur moyenne des probabilités de réponse obtenues par la méthode de modélisation varie selon la catégorie d'activité.

Un autre diagnostic utilisé pour comparer les deux méthodes est fondé sur la mesure de la variation des poids. Ce test est décrit en détail dans Dufour, Gagnon, Morin, Renaud et Särndal (2001). Si nous définissons le poids initial comme étant le poids de sondage avant l'ajustement pour la non-réponse, le poids intermédiaire comme étant le poids après l'ajustement pour la non-réponse, mais avant le calage, et le poids final comme étant le poids après le calage, alors la mesure de la variation D est définie comme étant

$$D = R_{01} + R_{12} + R_{int} + G \tag{5.1}$$

où R_{01} mesure les variations individuelles des poids entre les ensembles initial et intermédiaire de poids, R_{12} mesure les variations individuelles des poids entre les ensembles intermédiaire et final de poids, R_{int} mesure l'interaction entre les deux types de variation et G mesure la variation du poids moyen entre les ensembles initial et final. Le tableau 2 présente la mesure de la variation pour trois mois. D'après les résultats empiriques, la méthode réduit d'autant plus le biais dû à la non-réponse que la valeur de D, et plus spécialement R_{01}, est grande. Comme le montre le tableau 2, les écarts entre les mesures de la variation selon les méthodes courante et par modélisation semblent très faibles, quoique la méthode par modélisation produise systématiquement des valeurs plus élevées de D et R_{01}.

Enfin, nous avons réalisé une étude en simulation pour comparer le biais de non-réponse et la variance pour les méthodes courante et par modélisation. Nous nous sommes servis des données sur les répondants d'un mois à l'EPA, en traitant les probabilités de réponse estimées au moyen du modèle en deux étapes comme si elles étaient les valeurs réelles, et nous avons généré la non-réponse par échantillonnage de Poisson. Nous avons répété le processus 100 fois pour obtenir 100 pseudo échantillons. Puis, nous avons appliqué les méthodes courante et par modélisation d'ajustement pour la non-réponse séparément à chaque pseudo échantillon et construit des classes

d'ajustement pour la non-réponse, procédé à la repondération et calculé les taux de chômage (sans calage, d'après les poids corrigés pour la non-réponse).

Tableau 2 : Mesure de la variation pour les méthodes courante et par modélisation

	Courante		Modélisation	
	D	R_{01}	D	R_{01}
Mars 2001	0,0734	0,0059	0,0818	0,0137
Juin 2001	0,0552	0,0013	0,0613	0,0076
Octobre 2001	0,0609	0,0010	0,0649	0,0050

Puis, nous avons calculé les estimations du biais relatif (BR) et la racine carrée relative de l'erreur quadratique moyenne (RREQM) pour les deux méthodes. Le biais relatif estimé est défini comme suit

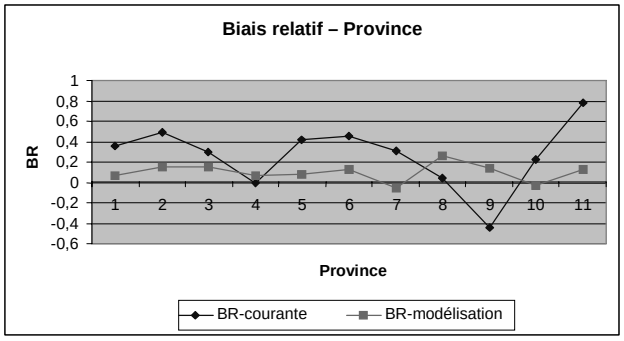
$$est(BR) = \left[\left(\frac{1}{100}\right)\sum_{i=1}^{100}\left(\hat{\theta}_i - \theta\right)\right]\frac{1}{\theta}\times 100\% \tag{5.2}$$

et la racine carrée relative de l'erreur quadratique moyenne comme étant

$$est(RREQM) = \sqrt{\frac{\sum_{i=1}^{100}\left(\hat{\theta}_i - \theta\right)^2}{100}}\times\left(\frac{1}{\theta}\right)\times 100\% \tag{5.3}$$

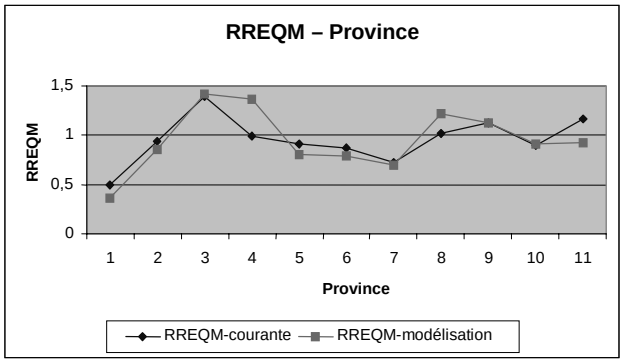
où $\hat{\theta}_i$ est l'estimation du taux de chômage pour un domaine donné après repondération pour le i^e pseudo échantillon et θ est le taux de chômage fondé sur le poids de sondage (avant l'ajustement pour la non-réponse) provenant de l'échantillon original de l'EPA.

Figure 4 : Comparaison du biais relatif des taux de chômage pour les méthodes courante et de modélisation, pour juin 2001



Les domaines considérés sont les provinces et les régions économiques d'assurance-emploi. Les figures 4 et 5 présentent les comparaisons pour les provinces. Il est évident que, pour la plupart des provinces, la méthode de modélisation réduit le biais de non-réponse. En outre, la RREQM est presque la même pour les deux méthodes, sauf pour quelques provinces pour lesquelles elle est plus élevée pour la méthode par modélisation. Ce profil indique que la variance du facteur de correction pour la non-réponse est plus forte pour la méthode par modélisation, puisque la variabilité des poids est plus grande. Nous avons également comparé le BR et la RREQM pour d'autres mois et observé les mêmes profils que ceux présentés aux figures 4 et 5.

Figure 5 : Comparaison de la RREQM des taux de chômage pour les méthodes courante et par modélisation, pour juin 2001



6. CONCLUSIONS

Nous comparons la méthode courante d'ajustement de la pondération pour la non-réponse totale à l'EPA à une nouvelle méthode fondée sur la modélisation de la réponse par la régression logistique.

Nous constatons que les méthodes courante et par modélisation de création de classes d'ajustement pour la non-réponse produisent des estimations du taux de réponse des ménages pour divers domaines ayant la même précision si les domaines sont fondés sur des variables géographiques ou des variables de plan de sondage. Par contre, la méthode par modélisation donne des estimations des taux de réponse observés de plus grande précision pour les domaines définis d'après des informations sur les prises de contact.

Diverses mesures diagnostiques témoignent de la supériorité générale de la méthode par modélisation relativement à la méthode courante. La prochaine étape consisterait à poursuivre l'amélioration du modèle de régression logistique utilisé pour créer les classes d'ajustement pour la non-réponse, en ajoutant un plus grand nombre d'informations sur les prises de contact et en précisant davantage les variables déjà incluses dans le modèle. En outre, la rédaction des spécifications de la méthode par modélisation est en cours.

REMERCIEMENTS

Les auteurs remercient Joanne Moloney et Zachary Pritchard, tous deux de Statistique Canada, pour leurs commentaires et suggestions constructifs qui les ont aidés à améliorer l'article.

ANNEXE

Suivent des renseignements détaillés sur les catégories des cinq effets principaux utilisés dans la méthode de modélisation en une étape.

Province = 1, Terre-Neuve
Province = 2, Île-du-Prince-Édouard
Province = 3, Nouvelle-Écosse
Province = 4, Nouveau-Brunswick
Province = 5, Québec
Province = 6, Ontario
Province = 7, Manitoba

Province = 8, Saskatchewan
Province = 9, Alberta
Province = 10, Colombie-Britannique

1 = Renouvellement de janvier ou de juillet
2 = Renouvellement de février ou d'août
3 = Renouvellement de mars ou de septembre
4 = Renouvellement d'avril et d'octobre
5 = Renouvellement de mai ou de novembre
6 = Renouvellement de juin ou de décembre

si type de strate = 0, alors type de strate = 1;
si 1 <= type de strate <= 9, alors type de strate = 2;
si 11 <= type de strate <= 19, alors type de strate = 3;
si 21 <= type de strate <= 29, alors type de strate = 4;
si 31 <= type de strate <= 39, alors type de strate = 5;
si 41<= type de strate <= 59, alors type de strate = 6;
si type de strate = 61, alors type de strate = 7;
si 65<= type de strate <= 98, alors type de strate = 8;
si type de strate = 99, alors type de strate = 9;

si n^{bre} de tentatives = 1, alors la catégorie de n^{bre} de tentatives = 1;
si n^{bre} de tentatives = 2, alors la catégorie de n^{bre} de tentatives = 2;
si 3 <= n^{bre} de tentatives <= 5, alors la catégorie de n^{bre} de tentatives = 3;
si 6 <= n^{bre} de tentatives <= 10, alors la catégorie de n^{bre} de tentatives = 4;
si 10 < n^{bre} de tentatives, alors la catégorie de n^{bre} de tentatives = 5;

si minuit <= heure de début de la dernière tentative < 11 h, alors heure = 1;
si 11 h <= heure de début de la dernière tentative < 14 h, alors heure = 2;
si 14 h <= heure de début de la dernière tentative < 17 h, alors heure = 3;
si 17 h pm <= heure de début de la dernière tentative < 19 h, alors heure = 4;
si 19 h <= heure de début de la dernière tentative < minuit, alors heure = 5

RÉFÉRENCES

Cox, D. R., et E. J. Snell (1989), *The Analysis of Binary Data*, Second Edition, London: Chapman and Hall.

Dufour, J., F. Gagnon., Y. Morin, M. Renaud, et C. E. Särndal (2001), "A Better Understanding of Weight Transformation through a Measure of Change", *Survey Methodology*, 27, 97-108.

Eltinge, J. L., et I. S. Yansaneh (1997), "Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey", *Survey Methodology*, 23, 33-40.

Haziza, D., O. Chow, C. Charbonnier, et J.-F. Beaumont (2001), "Construction of Imputation Cells for the Canadian Labour Force Survey", *2001 Proceedings of the Symposium*, Statistics Canada.

Hosmer, D. W., et S. Lemeshow (1980), "A Goodness-of-Fit Test for the Multiple Logistic regression Model", *Communications in Statistics*, A10, 1043-1069.

Hosmer, D. W., et S. Lemeshow (1989), *Applied Logistics Regression*, Wiley Series in Probability and Mathematics.

Lafortune, Y. (2002), “Diagnostic Analysis of Logistic Regression Models With survey Data”, SSMD-2002-006E, Methodology Branch Working Paper, Statistics Canada

Little, R. J. A. (1986), “Survey Nonresponse Adjustment for Estimate of Means”, *International Statistical Review*, 54, 139-157.

Statistics Canada. (1998), *Methodology of the Canadian Labour Force Survey*, Catalogue no. 71-526-XPB