

LE FRAGILE ÉQUILIBRE ENTRE LES BESOINS DES CLIENTS EN DONNÉES, LE FARDEAU DE RÉPONSE ET LA QUALITÉ DES DONNÉES: LE CAS DE L'ENQUÊTE SUR LES VÉHICULES AU CANADA

François Gagnon, Martin Beaulieu et Sébastien Landry¹

RÉSUMÉ

Trouver le juste équilibre entre le besoin des clients en données, le fardeau de réponse et la qualité des estimations est un défi de taille, notamment dans le cas d'enquêtes à coûts recouvrables comme l'Enquête sur les véhicules au Canada (EVC). Recueillir toute l'information souhaitée par les clients peut parfois mener à un fardeau de réponse inacceptable pour les répondants. Ceci peut résulter en de faibles taux de réponse et conséquemment en une piètre qualité des estimations, tant au niveau du biais que de la précision. La présentation traitera des moyens utilisés pour prévenir la non-réponse, des méthodes retenues pour la traiter efficacement lors de l'imputation et de la repondération ainsi que des méthodes utilisées pour mesurer les effets de la non-réponse sur le biais et la variance.

1. INTRODUCTION

Le méthodologiste d'enquête est fréquemment confronté au défi de répondre aux besoins toujours grandissants des clients en termes de données tout en s'assurant de maintenir un niveau de qualité acceptable. Répondre en particulier à tous les besoins en données des clients externes à Statistique Canada, qui défraient souvent la totalité des coûts de l'enquête, se traduit parfois par un lourd fardeau de réponse et peut entraîner une diminution du taux de réponse. Un faible taux de réponse peut ensuite mener à des estimations biaisées et moins précises. Trouver le juste équilibre entre les besoins des clients, le fardeau de réponse et la qualité des estimations est un défi de taille, notamment dans le cas d'enquêtes à coûts recouvrables comme l'Enquête sur les véhicules au Canada (EVC). Afin de relever ce défi, l'équipe de l'EVC a choisi de répondre à la plupart des besoins justifiés des clients; et ce, même s'il en découlait un fardeau de réponse élevé pour les répondants. En contrepartie, l'équipe de l'EVC tente de limiter les effets négatifs d'un fardeau de réponse élevé en utilisant des méthodes adéquates de prévention et de traitement de la non-réponse. Une brève description de l'EVC est présentée à la section 2 du présent article. Les pratiques courantes ainsi que la recherche en matière de prévention et de traitement de la non-réponse dans l'EVC font ensuite l'objet des sections 3 et 4. À la section 5, les méthodes étudiées et utilisées pour mesurer les effets de la non-réponse sur la qualité des données sont décrites. Enfin, nos conclusions sont formulées à la section 6.

2. L'ENQUÊTE SUR LES VÉHICULES AU CANADA

2.1 Plan de sondage de l'EVC

L'EVC a été développée en 1999 à la demande de Transports Canada afin de fournir des estimations trimestrielles de véhicules-kilomètres (nombre de kilomètres parcourus par les véhicules) et de passagers-kilomètres (somme des distances parcourues par les passagers) selon les caractéristiques des véhicules, l'âge et le sexe des conducteurs, l'origine et la destination des déplacements et le moment où le déplacement s'est effectué. La population cible de l'EVC comprend tous les véhicules routiers immatriculés au Canada, à l'exception du matériel spécial, des remorques, des motocyclettes et, depuis 2004, des autobus. La base de

¹ François Gagnon, Statistique Canada, Immeuble R.-H.-Coats – 17^{ième} étage, Pré Tunney, Ottawa, Ontario, Canada, K1A 0T6, Francois.Gagnon@statcan.ca, Martin Beaulieu, Statistique Canada, Immeuble R.-H.-Coats – 17^{ième} étage, Pré Tunney, Ottawa, Ontario, Canada, K1A 0T6, MartinJ.Beaulieu@statcan.ca, Sébastien Landry, Statistique Canada, Immeuble R.-H.-Coats – 17^{ième} étage, Pré Tunney, Ottawa, Ontario, Canada, K1A 0T6, Sebastien.Landry@statcan.ca.

sondage est composée des 13 fichiers provinciaux et territoriaux des véhicules routiers immatriculés au Canada. Chaque trimestre, un échantillon est tiré selon un plan de sondage à deux degrés. Au premier degré, un échantillon stratifié aléatoire simple de 5 375 véhicules pour les provinces et 2 700 véhicules pour les territoires est prélevé. La stratification est effectuée en utilisant la province ou le territoire ainsi que le type et l'âge du véhicule. Au deuxième degré, une grappe de jours consécutifs est choisie. Plus précisément, une date de départ est attribuée aléatoirement pour chacun des véhicules choisis dans l'échantillon; les propriétaires des véhicules doivent commencer à remplir le carnet de bord à la date de départ qui leur a été attribuée.

Dans le cas des provinces, la collecte des données s'effectue en deux temps. Tout d'abord, une interview téléphonique assistée par ordinateur (ITAO) est menée auprès des propriétaires des véhicules choisis. Cette interview sert à recueillir des renseignements généraux sur le véhicule et à demander au répondant s'il accepte de recevoir un carnet de bord. Le carnet de bord est ensuite envoyé par la poste afin de recueillir des renseignements détaillés sur l'utilisation du véhicule. Il comprend des questions sur le type de carrosserie du véhicule, les caractéristiques des déplacements (tels que la distance parcourue, l'objet et la durée de chaque déplacement, le nombre de passagers, les caractéristiques démographiques du conducteur et des passagers), ainsi que sur le carburant acheté. Dans le cas des véhicules de plus de 4 500 kg, on pose des questions supplémentaires sur la configuration du camion et sur le transport de marchandises dangereuses. Dans le cas des territoires, la collecte des données consiste simplement à l'envoi de cartes postales pour recueillir les lectures d'odomètre au début et à la fin du trimestre.

Suite à un traitement minutieux des données, les estimations sont produites à l'aide d'un estimateur par calage pour un plan à deux degrés. Les comptes à jour du nombre de véhicules routiers immatriculés au Canada sont utilisés pour effectuer une post-stratification au premier degré. Au deuxième degré, un calage sur le nombre de jours de travail et de congé du trimestre est effectué.

2.2 Remaniement de l'EVC

Un remaniement de l'EVC a récemment été effectué afin de répondre aux nouveaux besoins en données de nos clients. Dans le contexte de l'entrée en vigueur de l'Accord de Kyoto sur la réduction des gaz à effet de serre, les commanditaires de l'enquête, Transports Canada et, depuis 2004 Ressources Naturelles Canada, étaient intéressés à obtenir des estimations de la consommation de carburant des véhicules immatriculés au Canada. Le principal objectif de ce remaniement était donc d'ajouter, à compter de l'année de référence 2004, un carnet supplémentaire pour recueillir des informations beaucoup plus détaillées qu'avant sur les achats de carburant (quantité de carburant acheté, montant de l'achat, prix du litre, date et lecture de l'odomètre) afin de produire des estimations sur la consommation de carburant. Le remaniement a permis d'apporter d'autres modifications à l'enquête dans le but de mieux satisfaire les besoins des clients tout en compensant l'augmentation du fardeau de réponse associée à l'ajout du carnet sur les achats de carburant. Dans cette optique, les répondants doivent dorénavant déclarer 20 déplacements (un nouveau déplacement étant enregistré chaque fois que le conducteur du véhicule change ou encore qu'un passager entre ou sort du véhicule) au lieu de déclarer tous leurs déplacements sur une période de 7 jours comme c'était le cas auparavant. Suite à cette modification, le nombre de jours requis pour colliger 20 déplacements est maintenant d'environ 5 jours en moyenne. Le remaniement fut également une opportunité d'améliorer le carnet de bord en modifiant certaines questions, certains choix de réponse et en clarifiant les instructions. Des groupes de discussions ont permis de tester le carnet de bord remanié. De plus, une modification a été apportée à la population cible de l'EVC; puisque le taux de réponse et la qualité des données reçues étaient jugés trop faibles pour les autobus, ce type de véhicule est exclu de la population cible de l'EVC à compter de 2004.

3. PRÉVENTION DE LA NON-RÉPONSE

Les taux de réponse de l'EVC ne sont que de l'ordre de 50%. Le faible taux de réponse soulève évidemment des inquiétudes quant au risque que les estimations de l'EVC soient entachées d'un biais de non-réponse. Tout doit donc être mis en œuvre pour prévenir et traiter la non-réponse.

3.1 Qualité des informations sur la base de sondage

Une bonne qualité des informations de contact disponibles sur la base de sondage est un élément clé pour prévenir la non-réponse. Dans l'EVC, les adresses des propriétaires de véhicules échantillonnés sont mises à

jour à partir d'une base de données qui identifie les changements de propriété des véhicules. Ensuite, une recherche de numéros de téléphone est effectuée en consultant les annuaires téléphoniques informatisés.

3.2 Importance d'un contact téléphonique initial

Une étude réalisée par Wronski (2001) a démontré la nécessité de conserver un contact téléphonique initial afin de maintenir les taux de réponse de l'EVC à un niveau acceptable. Cette étude, réalisée en 2000, visait à comparer deux procédures de collecte. La première consistait en un simple envoi postal du carnet de bord (peu coûteux : environ 1\$ par envoi) alors que la deuxième prévoyait une ITAO initiale (très coûteuse : environ 40\$) avant l'envoi postal du carnet de bord. L'étude a démontré que dans le cas de la collecte effectuée par simple envoi postal, le taux de réponse était inférieur à la moitié du taux de réponse obtenu selon la procédure qui incluait une ITAO. Il était donc clair qu'il fallait conserver l'ITAO malgré les coûts élevés s'y rattachant.

3.3 Procédure de suivi

La procédure de suivi revêt une importance capitale dans une enquête comme l'EVC où le fardeau de réponse élevé peut inciter les propriétaires de véhicules à ne pas répondre. La procédure de suivi intense et relativement coûteuse - en raison des appels téléphoniques - qui est utilisée dans l'EVC est un autre moyen de prévenir la non-réponse. Le tableau 1 décrit les différentes étapes des procédures de collecte et de suivi de l'EVC.

Tableau 1 : Procédures de collecte et de suivi utilisées dans l'EVC

X : représente la semaine où le répondant doit commencer à remplir son carnet	<u>Identificateur de la semaine</u>
ITAO auprès des propriétaires des véhicules échantillonnés	X-2
Envoi postal du carnet de bord et du carnet sur les achats de carburant	X-1
Début du carnet de bord et du carnet sur les achats de carburant	X
1 ^{er} suivi téléphonique	X
Envoi d'une lettre de rappel	X+1
Envoi d'un questionnaire court (s'il y a lieu)	X+9
2 ^e suivi téléphonique (s'il y a lieu)	X+10
Envoi d'une carte postale pour recueillir la lecture de l'odomètre (s'il y a lieu)	X+12

3.4 Étude sur l'utilisation d'incitatifs

Toujours à la recherche de moyens pour réduire la non-réponse dans l'EVC, nous étudions présentement la possibilité d'utiliser des incitatifs pour augmenter les taux de réponse de l'enquête. Des études réalisées aux États-Unis ont démontré que dans certains cas l'utilisation d'incitatifs pouvait faire augmenter les taux de réponse de façon significative. L'utilisation d'incitatifs dans les enquêtes n'est cependant pas une pratique courante à Statistique Canada. Quelques rares études réalisées par l'agence (notamment par l'Enquête sur les voyageurs internationaux et l'Enquête sur les dépenses des ménages) n'ont pas été concluantes. Lors de groupes de discussions ayant eu lieu avant le développement de l'EVC, le sujet des incitatifs avait été abordé et une suggestion intéressante avait été apportée par les participants, soit celle de fournir un crayon pouvant s'accrocher au carnet de bord. Ils ont mentionné qu'ils laissaient le carnet dans leur voiture mais n'avaient pas toujours de crayon sous la main pour le compléter. Ils devaient alors le compléter plus tard en tentant de se souvenir de toutes les informations. Ceci peut entraîner des erreurs de réponse et même de la non-réponse totale (si trop de déplacements sont oubliés, le répondant peut décider de ne pas retourner le carnet). Nous procédons actuellement à un test sur l'utilisation d'incitatifs dans l'EVC. Les échantillons des deuxième et troisième trimestres de 2005 (5 375 véhicules par trimestre) seront divisés aléatoirement en trois groupes distincts d'environ 3 600 véhicules. Le premier groupe recevra un crayon à mine pouvant s'accrocher au carnet de bord, le deuxième recevra un porte-clés en plus du crayon, et finalement le troisième sera utilisé comme groupe témoin et ne recevra aucun incitatif. Le nom de l'enquête sera inscrit sur les crayons et les porte-clés. Les résultats seront analysés à l'aide du test exact de Fisher. Notre taille d'échantillon nous permettra de qualifier de statistiquement significative toute différence de plus de 2.2% entre les taux de réponse des trois groupes.

4. TRAITEMENT DE LA NON-RÉPONSE

4.1 Imputation

Diverses méthodes d'imputation sont utilisées dans l'EVC pour compenser la non-réponse partielle. Les renseignements de toutes les sources de données sont utilisés au maximum lors de l'imputation (l'ITAO, les questionnaires abrégés, les cartes postales, les carnets de bord et les carnets sur l'achat de carburant). La principale méthode d'imputation utilisée est l'imputation par plus proche voisin, tant au niveau des déplacements (afin d'imputer les renseignements manquants à l'aide d'un « déplacement donneur » parmi les autres déplacements du même véhicule) qu'au niveau des véhicules (en trouvant un « véhicule donneur »). Un exemple typique est l'imputation des déplacements d'un véhicule dont le propriétaire a accepté de répondre à l'ITAO initiale mais n'a jamais retourné le carnet de bord par la suite. Les informations recueillies lors de l'ITAO, notamment le nombre de kilomètres parcourus dans la semaine précédant l'interview, sont utilisées pour trouver un « véhicule donneur » et tous les déplacements de ce donneur sont attribués au « véhicule receveur ». L'imputation des variables sur les achats de carburant fait appel à des modèles de régression dans lesquels la consommation de carburant est expliquée par la distance parcourue et les caractéristiques du véhicule. L'imputation par déduction est également utilisée, un exemple typique étant l'utilisation des dates et lectures d'odomètre du carnet de bord pour imputer celles du carnet sur les achats de carburant (et vice-versa).

4.2 Repondération

Dans l'EVC, la repondération est utilisée pour compenser la non-réponse totale. La repondération vise essentiellement à augmenter les poids de sondage des répondants afin de compenser pour les non-répondants. On commence d'abord par diviser l'échantillon (répondants et non-répondants) en C classes s_1, s_2, \dots, s_C tel que

$\bigcup_{i=1}^C s_i = s$. Le poids des répondants après l'ajustement pour la non-réponse, w_i^* , pour l'unité i dans la classe c

est donné par : $w_i^* = \frac{w_i}{\hat{p}_c}$ où \hat{p}_c désigne le taux de réponse dans la classe c . L'estimateur par repondération

qui utilise C classes est donné par : $\hat{Y}_{RC} = \sum_{c=1}^C \hat{N}_c \bar{y}_{RC}$ où $\hat{N}_c = \sum_{i \in s_c} w_i$ et \bar{y}_{RC} désigne la moyenne des

répondants dans la classe c . Le biais de non-réponse de \hat{Y}_{RC} est donné par :

$$\text{Biais}(\hat{Y}_{RC}|s) = E_r(\hat{Y}_{RC} - \hat{Y}|s) = \sum_{c=1}^C \bar{p}_c^{-1} \sum_{i \in s_c} w_i (p_i - \bar{p}_c) (y_i - \bar{y}_c) \text{ où } \bar{p}_c = \frac{\sum_{i \in s_c} w_i p_i}{\sum_{i \in s_c} w_i} \text{ et } \bar{y}_c = \frac{\sum_{i \in s_c} w_i y_i}{\sum_{i \in s_c} w_i}.$$

Un des cas où l'expression du biais est égale à zéro est lorsque le mécanisme de réponse à l'intérieur des classes de repondération est uniforme, c'est-à-dire lorsque $p_i = \bar{p}_c \forall i \in s_c$, p_i étant la probabilité de répondre du véhicule i . L'objectif est donc de créer des classes qui soient homogènes par rapport à p_i afin d'éliminer ou de réduire le biais de non-réponse. Dans le cas de l'EVC les classes de repondération ont été créées à l'aide de la méthode des « scores », étudiée par Eltinge et Yansaneh (1997). La première étape de cette méthode, est de prédire les p_i pour toutes les unités de l'échantillon (répondants et non-répondants) à l'aide d'un modèle de régression logistique. Dans le cas de l'EVC, les variables explicatives du modèle sont les variables dichotomiques dérivées à partir des $m-1$ modalités de chacune des variables suivantes : « âge du véhicule » (2 groupes d'âge), « province/territoire » (13 modalités), « type de véhicule » (3 modalités) et « type de carburant » (3 modalités). La deuxième étape est d'ordonner les valeurs de \hat{p}_i en ordre croissant et d'utiliser l'analyse par grappe pour regrouper les véhicules ayant des \hat{p}_i similaires. Une fois les classes de repondération formées (entre 10 et 20 selon le trimestre, dans le cas de l'EVC), il s'agit simplement de procéder à la repondération dans chaque classe en calculant les poids ajustés pour la non-réponse. L'efficacité de la méthode des scores pour réduire le biais de non-réponse a notamment été démontrée par Haziza et coll. (2001).

5. MESURE DES EFFETS DE LA NON-RÉPONSE

5.1 Étude des non-répondants pour déterminer l'amplitude du biais de non-réponse

Malgré un traitement méticuleux de la non-réponse, les faibles taux de réponse observés dans l'EVC peuvent mener à des estimations entachées d'un biais de non-réponse si le mécanisme de réponse est non-négligeable, c'est-à-dire si la probabilité de répondre à l'enquête est corrélée avec une des variables à l'étude. Il est évidemment d'un grand intérêt de connaître la direction et l'amplitude du biais de non-réponse afin de le réduire ou de le corriger. À compter d'avril 2005, nous espérons entreprendre une étude des non-répondants afin d'estimer le biais de non-réponse, de recueillir les caractéristiques des non-répondants et de connaître les raisons pour lesquelles ils n'ont pas répondu à l'enquête. L'étude s'effectuera par téléphone. Afin d'estimer le biais, une question portera sur la distance parcourue par le véhicule au cours d'une période donnée. D'autres questions nous permettront de déterminer les caractéristiques des non-répondants afin d'améliorer les méthodes de traitement de la non-réponse. Les raisons pour lesquelles ils n'ont pas répondu à l'enquête feront également l'objet d'une question; ces raisons pourraient permettre d'améliorer le carnet de bord et/ou les procédures de collecte.

5.2 Estimation de la variance due à la non-réponse

Dans le but de renseigner les utilisateurs sur la précision réelle des estimations, la variance due à la non-réponse (incluant la variance due à l'imputation) est estimée à l'aide du Système pour l'estimation de la variance due à la non-réponse et à l'imputation de Statistique Canada, aussi connu sous le nom de SEVANI (Beaumont et Mitchell, 2002). Les coefficients de variation publiés pour l'EVC tiennent compte à la fois de la variance due à l'échantillonnage et de la variance due à la non-réponse.

6. CONCLUSION

Répondre aux besoins des clients en matière de données tout en maintenant la qualité des estimations à un niveau acceptable demeure un défi de taille, notamment dans les enquêtes à coûts recouvrables comme l'EVC. Il est primordial de poursuivre nos efforts en matière de prévention et de traitement de la non-réponse en plus de bien informer les utilisateurs des effets des faibles taux de réponse de l'EVC. À court terme, cela se traduira par la poursuite de l'étude sur l'utilisation d'incitatifs et par la réalisation de l'étude des non-répondants.

REMERCIEMENTS

Les auteurs aimeraient remercier Christian Nadeau, Isabelle Marchand et Julie Trépanier pour leurs commentaires, lesquels ont servi à améliorer la qualité du document.

RÉFÉRENCES

- Beaumont, J.-F. et Mitchell, C. (2002), « Système pour l'estimation de la variance due à la non-réponse et à l'imputation (SEVANI) », *Recueil du Symposium 2002 de Statistique Canada*, Statistique Canada.
- Eltinge, J. L., et Yansaneh, I.S. (1997), « Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey », *Techniques d'enquête*, 23, p. 37-45.
- Haziza, D., Charbonnier, C., Chow, S.Y. et Beaumont, J.-F. (2001), « Construction de cellules d'imputation pour l'enquête sur la population active du Canada », *Recueil du Symposium 2001 de Statistique Canada*, Statistique Canada.
- Wronski, A. (2001), « Recueillir des données sur l'utilisation de véhicules – l'expérience de l'Enquête sur les véhicules au Canada », *Recueil du Symposium 2001 de Statistique Canada*, Statistique Canada.