VARIABLES AUXILIAIRES ET PARADONNÉES POUR AMÉLIORER LE REDRESSEMENT D'UNE ENQUÊTE TÉLÉPHONIQUE : LE CAS DE CAMME-TH

Stéphane LEGLEYE (*), Bénédicte MORDIER, Amandine NOUGARET, Marie CLERC, François BECK (**), Maxime LEVESQUE (***)

(*) Insee, Division Recueil et traitement de l'information
(**) Insee, Département de la Démographie
(***) Insee, Département de la Conjoncture

stephane.legleye@insee.fr
benedicte.mordier@insee.fr

Mots-clés : paradonnées, variables auxiliaires, redressement, téléphone, Camme

Résumé

Le redressement des enquêtes peut se faire en une ou deux étapes, le ce dernier étant généralement plus efficace [1]. Le redressement en une étape est généralement un calage, mobilisant des variables socio-démographiques renseignées par les répondants et dont les totaux sont connus dans la population cible. Lorsqu'il est fait en deux étapes, la première consiste en une correction de la non-réponse totale (CNRT) reposant sur une modélisation de la réponse à l'enquête. Pour cela, on mobilise généralement les variables auxiliaires de la base de sondage qui sont les plus corrélées à la réponse à l'enquête et à ses variables cibles. Les variables socio-démographiques sont des variables de choix, mais les paradonnées sont potentiellement intéressantes. Ce sont les variables décrivant le processus de collecte et notamment les efforts entrepris pour contacter les personnes sélectionnées [2]. Leur grand avantage est qu'elles se trouvent renseignées pour les répondants comme pour les non-répondants et dans de nombreuses enquêtes, elles s'avèrent liées aux comportements étudiés. Ce lien se maintient souvent dans des modèles multivariés contrôlant les variables de calage, faisant dès lors craindre des biais résultant de leur ignorance dans le redressement en une étape [2-8].
Dans cette étude, nous évaluerons l'effet d'un redressement en deux étapes plutôt qu'une sur les estimations de plusieurs variables cibles d'une enquête téléphonique. Nous envisagerons trois CNRT mobilisant exclusivement ou en combinaison, variables auxiliaires et paradonnées.

Les données utilisées sont celles de l'expérimentation menée en 2017 sur l'enquête mensuelle de conjoncture auprès des ménages (Camme). Camme est une enquête téléphonique menée par l'Insee qui quantifie la confiance des ménages dans la situation économique à l'aide d'indicateurs d'opinions et de perception de la conjoncture. Le tirage des ménages est fait dans les fichiers de la Taxe d'habitation : les numéros de téléphone mobilisés sont soit ceux retrouvés dans un annuaire inversé à l'aide des coordonnées du ménage fiscal (protocole de l'enquête Camme courante), soit ceux fournis par les contribuables à l'administration fiscale (protocole mis en place pour l'expérimentation de 2017). La mobilisation des coordonnées issues des sources fiscales, en complément des données annuaire, assure une couverture téléphonique de 70% de l'échantillon initial. Camme est un panel rotatif sur trois mois : nous retiendrons ici uniquement les premières interrogations, soit 3 210 questionnaires complets sur les mois de mai à août 2017.

Les variables cibles de l'enquête sont les dix « soldes » d'opinions classiques de Camme (variables ordonnées à trois modalités). Les variables auxiliaires sont : l'âge de la personne de référence, la tranche de revenu fiscal du ménage, le type de logement et le statut d'occupation, la taille de l'unité urbaine, la région de résidence et les types de coordonnées présentes dans les fichiers fiscaux (téléphone ou mobile et adresse courriel).
Les paradonnées sont : envoi d'un courrier, envoi d'un courrier de relance pour les impossibles à joindre ; date, heure, jour de la semaine et issues détaillées des appels ne sont connus que pour les 5 premiers appels, mais le nombre total de tentatives d'appels est toujours connu. Nous retiendrons ici

le nombre total d'appels (compris entre 1 et 20) ainsi que le nombre total de raccroche/refus pour ce ménage sur les 5 premiers appels.

La sélection des variables pour les modèles de CNRT est faite en retenant celles qui apparaissent les plus liées aux dix variables cibles (parmi les répondants) et à la participation à l'enquête.
Les variables auxiliaires les plus liées aux variables cibles sont : le revenu fiscal, le statut d'occupation du logement et la nature de ce dernier, l'âge de la personne de référence, la résidence en IdF et le types de coordonnées dans le fichier fiscal. Les paradonnées téléphoniques se révèlent très peu corrélées aux variables cibles ; en revanche, certaines sont associées au revenu fiscal de la base de sondage (le décile de revenus sont liés à l'envoi d'un courrier de relance et au nombre de raccroches/refus : r=-0.07). Les courriers sont chacun liés à 2 variables cibles.
Les variables auxiliaires les plus liées à la participation à l'enquête sont les mêmes que précédemment.

Chaque modèle de CNRT se fait par régression logistique et constitution de Groupes homogènes par la méthode des quantiles. Les corrélations des pondérations finales (post-calage) avec les variables cibles sont faibles (comprises entre 0.01 et 0.10, 5 sur 30 seulement étant supérieures à 0.05), particulièrement dans le modèle n'incluant que les paradonnées. Le coefficient de variation des poids vaut 37 pour le calage direct, 42, 75 et 77 pour les pondérations en deux étapes. Aucune estimation des variables cibles ne présente d'intervalle de confiance disjoint de ceux de l'estimation avec le calage direct.

Nous discutons la portée de ces résultats pour Camme mais aussi pour les enquêtes téléphoniques à génération aléatoire de numéros pour lesquelles les liens observés entre paradonnées et revenu fiscal pourraient avoir une utilité.

Bibliographie

[1] Haziza D, Lesage E: A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics 2016;32:129-145.
[2] Kreuter F (ed): Improving surveys with paradata, New York, John Wiley & Sons, 2013.
[3] Beck F, Legleye S, Peretti-Watel P: Le recours au téléphone dans les enquêtes en population générale sur les drogues: Journées de Méthodologie Statistique,. Paris, Insee, 2002,
[4] Legleye S, Charrance G, Razafindratsima N, Bohet A, Bajos N, Moreau C: Improving survey participation: cost effectiveness of call-backs to refusals and increased call attempts in a national telephone survey in France. Public Opinion Quarterly 2013;77:666-695.
[5] Legleye S, Razakamanana N, Charrance G, Juillard H: L'utilisation des historiques d'appels pour redresser une enquête téléphonique : une étude par simulation à partir de l'enquête Fecond XIIème Journées de méthodologie statistique de l'INSEE. Paris, France, 2015,
[6] Maitland A, Cordero CC, Kreuter F: An exploration into the use of paradata for nonresponse adjustment in a health survey; JSM proceedings. Alexandria, VA, American Statistical Association, 2009, pp 370-378.
[7] Legleye S, Razakamanana N, Charrance G, Juillard H: Is it worth using paradata to correct for total non-response in telephone survey? A simulation study based on real data: ESRA. Reykjavik, 2015,
[8] Blom AG: Nonresponse Bias Adjustments: What Can Process Data Contribute? Institute for Social and Economic Research.