L'utilisation des historiques d'appels pour pondérerune enquête téléphonique: une étude par simulation à partir de l'enquête Fecond

Stéphane Legleye[1,2,3], Nirin Razakamanana[1], Géraldine Charrance[1] et Hélène Juillard[1]

1. Institut national des études démographiques (INED), Paris (France)
2. Inserm, U669, Paris (France)
3. Univ Paris-Sud and Univ Paris Descartes, UMR-S0669, Paris (France)

Domaine : Traitement de la non-réponse

Résumé

Dans une enquête, malgré les efforts mis en place lors de la collecte,la non-réponse totale peut, créer du biaiset réduire la précision des estimateurs. Pour limiter le biais, il est courant de corriger l'échantillon par des méthodes de repondération. Ces traitements peuvent doc mobiliser des variables auxiliaires de la base de sondage, renseignées pour les répondants et non répondants, dans le cas de modélisations de la non-réponse totale, ou bien des données externes concernant la population cible, dans le cas d'un calage sur marges. Dans les enquêtes, outr. e les données de la base de sondage, sont recueillies des informations sur le processus de collecte lui-même, les dates et heures des tentatives de contact, les issues, etc. Ces « paradonnées » contiennent de l'information prédictive de la participation (par définition), et sont potentiellement très liées aux réponses à l'enquête, comme cela a été montré dans plusieurs contextes [1-3] et peuvent permette de limiter les biais résiduels à l'issue d'une post-stratification [1]. Elles reflètent en effet la disponibilité des personnes, ce qui s'avère lié à leurs caractéristiques professionnelles et familiales et plus généralement à leur mode de vie, leur bienveillance relative aux enquêtes, à l'intérêt public etc. [4, 5]. Ces éléments ne sont pas réductibles aux caractéristiques sociodémographiques généralement considérées pour les calages. C'est pourquoi, potentiellement, l'usage des paradonnées dans les traitements post-collecte peut également réduire la variance des estimateurs [6].

Lors de la conduite d'enquêtes téléphoniques aléatoires, il n'existe pas de base de sondage et donc pas de variables auxiliaires en dehors des paradonnées constituées par les historiques d'appels des numéros (dates, heureset issue des appels). Elles sont complexes à traiter et c'est pourquoi seul un calage sur marges est en général effectué.

Nous proposons d'évaluer l'utilité desparadonnées en comparant trois méthodes de traitement post-collecte d'un échantillon d'une enquête téléphonique : le calage direct, la correction de la non-réponse totale par groupes homogènes mobilisant les paradonnées, d'une part par modélisation logistique, d'autre part par classification, auxquelles un calage est ensuite appliqué. Ces trois méthodes sont comparées sur des données simulées extraites d'une enquête réelle, l'enquête Fecond, réalisée par téléphone en 2010 par l'INSERM et l'INED où 8638 individus âgés de 15-49 ans ont rempli un questionnaire (taux de participation de 45%). Cette base de répondants est considérée comme la population cible pour notre étude.

Les paradonnées utilisées consistent en 5 variables : le nombre total d'appels effectués et l'issue de chaque appel (non contact/répondeur, rendez-vous, passation partielle ou complète, refus –plusieurs refus pouvant être enregistrés). Trois mécanismes de non-réponse sont générés : un mécanisme complètement au hasard (Missingcompletely at random, MCAR), un mécanisme aléatoire ou ignorable (Missing at random, MAR) s'appuyant sur les variables sociodémographiques utilisées pour le calage direct, et un mécanisme non aléatoire ou non ignorable (Not missing at random, NMAR), liant la non-réponse totale à9 variables d'intérêt de l'enquête. Chaque mécanisme est appliqué avec un taux de réponse de 50% sur l'échantillon initial de répondants et simulé 1000 fois (tirages aléatoires sans remise). Biais, variance et erreur quadratique moyenne sont calculés sur cette base pour les trois méthodes de traitement proposées et pour 9 variables d'intérêt,relativement aux « vraies » valeurs observées sur l'échantillon de répondants initial.

Une macro SAS paramétrable est disponible pour répliquer l'analyse sur d'autres données.

Dans cette approche simple, les paradonnées ne prennent pas en compte le temps (la chronologie des appels), pas plus que la variété des plages horaires (jours de la semaine et moment de la journée où sont passés les appels). Cela sera l'objet d'une prochaine analyse.

Références

1. Blom, A.G. *Nonresponse Bias Adjustments: What Can Process Data Contribute?*Institute for Social and Economic Research, 2009.
2. Maitland, A., C.C. Cordero, and F. Kreuter, *An exploration into the use of paradata for nonresponse adjustment in a health survey*, in *JSM proceedings*. 2009, American Statistical Association: Alexandria, VA. p. 370-378.
3. Legleye, S., et al., *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**(3): p. 666-695.
4. Groves, R.M., E. Singer, and A. Corning, *Leverage-saliency theory of survey participation.* Public Opinion Quarterly, 2000(64): p. 299-308.
5. Singer, E., *Toward A Benefit-Cost Theory Of Survey Participation: Evidence, Further Tests, And Implications.* Journal Of Official Statistics, 2011. **27**(2): p. 379-392.
6. Little, R.J.A. and S. Vartivarian, *Does weighting for nonresponse increase the variance of survey means?*Survey methodology, 2005. **31**(2): p. 161–68.