Comparaison de méthodes pour la correction de la non-réponse totale par repondération

Nicolas Sigler [1] & Benoît Buisson [1]

[1] *Insee, Direction régionale des Pays de la Loire, Sed, Pôle Ingénierie Statistique Entreprises, 105 rue des Français Libres, BP 67401, 44274 NANTES CEDEX 2,*
nicolas.sigler@insee.fr, benoît.buisson@insee.fr

Le pôle ingénierie statistique entreprises de l'Insee a en charge le redressement de nombreuses enquêtes thématiques entreprises. Pour ces enquêtes, la non-réponse totale est habituellement corrigée par repondération, après mise en évidence de sous-populations supposées homogènes, du point de vue de leur comportement de réponse. Pour construire ces "groupes de réponses homogènes" (GRH), on modélise la probabilité de répondre selon différentes variables auxiliaires (connues sur les répondants et sur les non répondants) telles que le secteur, la taille ...

Cette communication exposera les conclusions d'un mémoire réalisé au sein du pôle ISE par Émilie Dequidt dans le cadre de la Formation continue diplômante des attachés (FCDA) de l'ENSAI. Le mémoire s'appuie sur des simulations à partir d'une enquête réelle (l'enquête TIC 2011 relative aux technologies de l'information et de la communication et au commerce électronique) pour comparer deux méthodes de constitution des GRH : la *méthode des scores* et la *segmentation par arbres*.

La *méthode des scores* s'appuie sur le calcul d'une probabilité de réponse "prédite" par unité (calcul faisant suite à une régression logistique avec la PROC LOGISTIC de SAS) et sur le classement des unités échantillonnées selon la valeur croissante de cette probabilité de réponse. L'échantillon des répondants et des non-répondants est alors découpé en 25 GRH de même taille.

La *segmentation par arbres* est une technique inductive de classement. Partant de l'ensemble de l'échantillon, la méthode détermine à chaque étape la variable et les modalités qui séparent le mieux l'ensemble des unités, relativement au fait de répondre ou non. La procédure (effectuée selon l'algorithme CHAID, à l'aide de la macro treedisc du logiciel SAS) est itérative, et permet de construire un "arbre" dont les "feuilles" (ou noeuds terminaux) sont les différents GRH.

La comparaison entre les deux méthodes est effectuée sur une population fictive (reconstituée à partir des unités répondantes de l'enquête TIC 2011) dont sont extraits 1000 échantillons ; sur chacun, on génère de la non-réponse totale selon 21 scénarios distincts, combinaisons de divers mécanismes de réponse (aléatoire simple, secteur x taille, avec variable cachée ...) et de différents taux de non-réponse (10, 20 et 30 %). La non-réponse totale est alors corrigée selon les 2 méthodes envisagées, puis les estimateurs obtenus par chaque méthode et pour chaque scénario sont comparés (en calculant notamment le biais de Monte Carlo et l'erreur relative de Monte-Carlo) aux estimateurs cibles calculés sur la population entière, pour quelques variables d'intérêt de l'enquête.

Bibliographie

Caron, N. (2005). La correction de la non-réponse totale par repondération et par imputation, Document de travail de l'Insee, M0502.

Confais, J., et Nakache, J.-P. (2003). Statistique explicative appliquée, Paris, Éditions Technip.