Amélioration du redressement de la non-réponse des communautés dans le recensement

Lise REYNAERT

En France, le recensement repose désormais sur une collecte annuelle qui concerne successivement toutes les communes au cours d'une période de cinq ans. Le cumul des cinq enquêtes les plus récentes conduisent aux résultats du recensement, datés du milieu du cycle quinquennal considéré. Ces résultats établis à un niveau communal, et même infracommunal pour certains, sont essentiels au fonctionnement de la société et, à ce titre, sont attendus avec une extrême exigence sur l'exactitude des chiffres à tout échelon géographique.

Chaque année, environ 9 millions de personnes sont recensées, toutes catégories de population confondues. Ces catégories distinguent : la population des ménages, celle des habitations mobiles ou des sans-abris et celle des individus des communautés à laquelle s'intéresse ce mémoire.

Le recensement des communautés permet de dénombrer, exhaustivement sur un cycle de cinq ans, la population habitant en communauté et de la caractériser. Le redressement de la non-réponse de la collecte annuelle s'effectue par hot-deck séquentiel ; cette méthode impute chaque valeur manquante par la modalité de la variable prise par le répondant précédent « le plus proche » lorsque l'on parcourt le lot de saisie, les individus étant répartis dans une vingtaine de lots de saisie. Le « plus proche » donneur est choisi parmi les résidents des communautés, selon des contraintes propres à chaque question. Cette méthode limite les temps de traitement et permet ainsi de respecter la forte contrainte de délai à laquelle le recensement est soumis, de par son cadre législatif.

La population en communauté présente toutefois deux particularités qui engendrent quelques cas de redressements aberrants, certes limités mais regrettables au vu de la portée des résultats du recensement à un niveau fin : la première particularité est le lien entre le comportement de réponse et l'ordre du lot de saisie, occasionnant des redressements massifs de non-répondants successifs par un unique donneur ; la seconde est l'hétérogénéité des populations en communauté, entraînant le redressement de non-répondants par des donneurs très différents bien que proches dans le lot de saisie.

L'objectif de ce mémoire est de proposer des améliorations de la méthode de redressement afin d'éviter les inconvénients de la méthode actuelle. Dans une vision à court terme, nous proposons des modifications du hot-deck séquentiel actuel, en restant dans la logique des traitements standards du RP et de ses contraintes de délai. Ensuite, nous considérons des méthodes de redressement alternatives qui pourraient être mises en place en cas de refonte à moyen terme des applications du recensement.

Afin de quantifier les performances des différentes propositions et méthodes testées, nous avons défini quatre critères qui reflètent nos attentes quant à la qualité de la méthode de redressement :

- Nous souhaitons que la méthode résolve les cas de redressements aberrants engendrés par la méthode actuelle.
- La distorsion de la distribution des variables à un niveau communal doit être la plus faible possible en cas de non-réponse massive et concentrée.
- Le nombre maximal de dons par individu répondant doit être réduit par rapport à la méthode actuelle.
- Enfin, la méthode doit avoir une bonne qualité prédictive en cas de non-réponse diffuse.

Ces critères nous ont permis d'appréhender et de comparer la qualité des différentes propositions et méthodes testées.

A moindre coût en matière de production courante et d'évolution des processus, nous pouvons améliorer sensiblement la méthode actuelle de deux manières :

- d'une part en regroupant l'ensemble des individus en communauté dans un unique lot de saisie. Cette proposition limiterait les différences au regard des variables liées au lieu d'habitat entre des individus consécutifs dans le lot de saisie. Cela permet d'éviter certains cas de redressements aberrants liés à des variables géographiques.
- d'autre part en redéfinissant les contraintes d'imputation, et notamment en ajoutant la sous-catégorie de communauté qui permet de former des sous-populations homogènes (gendarmes, étudiants...). Cette proposition permet de résoudre les redressements aberrants de la méthode actuelle et de diminuer la distorsion de la distribution des variables.

Une troisième proposition, plus difficile à mettre en place à court terme, consiste à ajouter un aléa d'imputation dans le hot-deck séquentiel. En pratique, il s'agit pour chaque non-répondant de tirer son donneur parmi les 3 (ou plus) répondants qui précèdent dans le fichier. Cette proposition permet de limiter quelque peu les redressements massifs de communautés entières par un unique individu.

Dans une vision à plus long terme, nous avons testé et comparé les performances du hot-deck par classe, du hot-deck métrique à partir d'une mesure de similarité basée sur le V de Cramer et la repondération. Parmi ces méthodes, celle par hot-deck métrique permet d'obtenir les meilleurs résultats au regard de nos critères de qualité : elle permet notamment de résoudre les cas de redressements maladroits et de réduire la distorsion des variables en cas de non-réponse massive et concentrée. De plus, elle limite les dons multiples et montre une bonne qualité prédictive en cas de non-réponse diffuse et de faible ampleur.

Le hot-deck par classe donne de bons résultats : il permet de résoudre les cas de redressements aberrants, de limiter les dons multiples et de réduire la distorsion des variables par rapport au hot-deck séquentiel. En revanche, sa qualité prédictive en cas de non-réponse diffuse est inférieure à celle du hot-deck séquentiel.

Enfin, la méthode de redressement par repondération n'est pas envisageable car elle engendre une perte d'information à un niveau communal. Cette perte d'information n'est pas admissible dans la mesure où les populations légales et les résultats du recensement sont attendus pour refléter fidèlement la population de chaque commune.