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

Lise LEFEBVRE[1]

Division « Méthodes et traitements des recensements », DSDS, Insee.

En France, environ 9 millions de personnes sont recensées chaque année, toutes catégories de population confondues. Ces catégories distinguent dans une population de près de 65 millions d'habitants : la population des ménages (97,7 % au RP2011), celle des habitations mobiles ou sans abri (0,2 %) et celle des individus en communautés (2,1 %) à laquelle s'intéresse cette étude.

Le recensement des communautés permet de dénombrer exhaustivement - à raison d'un cinquième par an sur 5 ans - la population habitant en communauté, et de la caractériser. Le redressement de la non-réponse totale et partielle est concomitant et analogue dans le principe avec celui des ménages ; il s'effectue par hot-deck séquentiel, une méthode qui impute chaque valeur manquante par la modalité de la variable prise par le répondant précédent de même catégorie lorsque l'on parcourt le lot de saisie au sens de l'identifiant du recensement et de critères statistiques propres à chaque question. Cette méthode permet de respecter les contraintes du recensement, notamment en limitant les temps de traitement.

Cette méthode robuste se révèle globalement satisfaisante pour les ménages, dont la non-réponse est diffuse et de faible ampleur (2,3 % de logements non répondants). Toutefois, elle comporte certaines faiblesses, particulièrement visibles pour les individus en communautés. En raison de conditions de collecte plus difficile (intermédiaire d'un gestionnaire, population peu disponible), le phénomène de non-réponse y est souvent bien plus concentré. Par ailleurs, la population de chaque communauté est assez spécifique et ne correspond pas forcément à celle d'une communauté voisine. Ces particularités ont pour effet d'amplifier deux inconvénients majeurs du hot-deck séquentiel :

- La méthode restreint drastiquement le champ des donneurs, provoquant des distorsions dans la distribution des variables statistiques. En effet, dans le cas de la non-réponse en bloc d'une communauté, les réponses du dernier individu répondant d'une structure voisine seront imputées à tous les non-répondants de la communauté, jusqu'à ce qu'un autre répondant soit rencontré.
- Le modèle d'imputation repose sur l'hypothèse pas toujours adaptée que l'ensemble des réponses d'un individu ainsi que son comportement de réponse sont similaires à ceux de l'individu précédent. Cette mauvaise spécification du modèle d'imputation est susceptible de provoquer des biais d'imputation.

L'objectif de cette étude est d'améliorer la méthode de redressement afin d'éviter ces deux inconvénients. Dans une vision à court terme, nous proposerons des alternatives à la méthode actuelle de hot-deck séquentiel, en restant dans la logique des traitements standards du RP. L'introduction d'un nombre maximal de fois d'utilisation d'un donneur dans le processus d'imputation pourrait être envisagée. Elle permettrait d'éviter les donneurs multiples et serait également l'occasion d'enrichir la redéfinir les critères statistiques servant à la spécification du modèle d'imputation.

Ensuite, nous ferons des propositions concernant la meilleure méthode de redressement à mettre en place en cas de refonte à moyen terme du recensement qui changerait plus en profondeur la logique actuelle : nous testerons un hot-deck métrique à partir d'une mesure de similarité basée sur le V de Cramer et un hot-deck par classe doublement robuste. Enfin, nous envisagerons un traitement de la non-réponse totale par repondération.

[1] lise.lefebvre@insee.fr