Estimations communales exploitant les données de l'enquête « Famille et logements » et du recensement : une opération périlleuse

Pascal Ardilly, Insee, Département des méthodes statistiques
pascal.ardilly@insee.fr

Domaine concerné : sondages; estimation sur petits domaines.

L'enquête Famille-Logements est une enquête associée au recensement de 2011. Elle s'est déroulée en métropole dans près de 1500 communes de toutes tailles, recueillant des informations très diverses autour des liens entre les cohabitants du logement, autour des caractéristiques sociodémographiques des membres du ménage, de la filiation, du mode de garde des jeunes enfants, de la multi résidence des adultes et des enfants, etc. Assise sur un très gros échantillon national (360 000 répondants), elle devrait produire une information de qualité satisfaisante au niveau régional.
Afin d'organiser une restitution de données « personnalisées » qui avait été promise aux communes participantes, il a fallu produire des estimations communales à partir de cette enquête, en s'appuyant sur les données du recensement. L'opération est évidemment particulièrement ambitieuse - et peu banale - parce que les tailles d'échantillon communales, à l'exception peut-être d'une poignée de très grandes communes, sont largement insuffisantes pour produire des estimations acceptables à partir des méthodes classiques. Aussi, sans autre alternative, l'Insee a utilisé une approche par modèle, dite « petits domaines », pour construire une quarantaine de dénombrements, lesquels seront intégrés dans une fiche de restitution aux communes parmi d'autres données provenant directement du recensement.

Sur le principe, une variable d'intérêt ayant été fixée, il s'agit d'abord de trouver des variables explicatives disponibles à la fois dans l'enquête et dans une source « exhaustive » qui, en la circonstance, est l'enquête annuelle de recensement 2011. Une fois ces variables isolées - par des techniques de régression classiques - on forme des sous-populations en combinant leurs modalités et on estime, sous-population par sous-population, la probabilité pour que l'unité statistique impliquée (adulte, enfant, famille,…) vérifie la condition que traduit la variable d'intérêt. Cette probabilité a la propriété essentielle d'être estimée à partir d'un échantillon de « grande taille » - ou du moins dont la taille est bien supérieure à celle de l'échantillon répondant dans la commune. Par exemple, pour estimer le nombre de personnes pacsées, on explique le fait d'être pacsé par le sexe, l'âge, le diplôme, la géographie, et on calcule - pour chaque zone géographique distinguée - la probabilité qu'une personne quelconque de sexe donné, dans une tranche d'âge donnée, possédant un niveau de diplôme donné, soit pacsée. Pour terminer, on multiplie les différentes probabilités par les effectifs communaux recensés (au sens du recensement cumulé sur 5 ans) dans chaque sous-population.

Cette méthode, finalement assez simple à expliquer aux utilisateurs, a l'avantage de limiter la variance d'échantillonnage, puisque les termes aléatoires que sont les probabilités associées aux différentes sous-populations sont construits à partir d'échantillons de tailles généralement substantielles. En contrepartie, on doit accepter un biais dont le modèle est directement responsable : les estimations sont ainsi fortement modèle-dépendantes et cela renvoie à la question récurrente de la place que doit accorder la statistique publique à la modélisation. De plus, comme les estimations sont soumises à l'examen des élus locaux, qui connaissent bien

leur commune, il y a un vrai risque de décalage avec la réalité du terrain et cela rend l'opération « politiquement » périlleuse.
Cette méthodologie n'est applicable que parce qu'il existe une source d'information jugée fiable au niveau communal et suffisamment explicative des variables d'intérêt : c'est le recensement de 2009. Néanmoins, elle n'est pas sans poser des problèmes de diverses natures (absence d'exhaustivité en grande commune et dans les petites communes lorsqu'on doit utiliser l'échantillon complémentaire, erreurs de mesure qui contribuent au biais d'estimation final, décalage temporel à gérer avec l'EFL). *In fine*, il est possible d'apprécier le biais d'estimation pour chaque variable, ce biais étant la principale source de critiques à l'encontre de cette approche. On souligne, pour terminer, que le biais est finalement corrigé par un système de double calage assurant simultanément une cohérence, au niveau national avec l'estimation nationale EFL, et au niveau communal avec les données du recensement.