Journées de méthodologie statistique de l'Insee 2022

UTILISATION DES PROBABILITÉS D'INCLUSION EXACTES POUR LE SONDAGE INDIRECT EN POPULATION ASYMÉTRIQUE

Henri Bodet (*), Arnaud Fizzala (**)

(*) Insee, Pôle Ingénierie Statistique d'Enquête
(**) Insee, Division Sondages

henri.bodet@insee.fr, arnaud.fizzal@insee.fr

Mots-clés : Enquêtes auprès des entreprises, sondage indirect, partage des poids.

Domaine concerné : Théorie des sondages aval, pondération et repondération, Intégration de données, appariement et fusion de sources.

Résumé

La méthode généralisée de partage des poids (MGPP) mise au point par Lavallée [1] est la solution habituelle pour les situations relevant des sondages indirects. Sa principale motivation est qu'elle fournit une solution lorsque l'on n'est pas capable de calculer les probabilités d'inclusion. Son application pose toutefois des problèmes lorsque les poids sont très dispersés et peuvent prendre des valeurs proches de l'unité (voire égales à l'unité).

Plusieurs solutions ont été proposées par Lavallée et Labelle-Blanchet [2] et l'une d'elles - la version pondérée de la MGPP - est à présent mise en application dans l'enquête sectorielle annuelle de l'Insee [3].

Cependant, beaucoup d'enquêtes auprès des entreprises menées à l'Insee présentent deux caractéristiques : les poids sont très dispersés (ils varient typiquement de 1 à 60) et le plan de sondage est très simple (il s'agit d'un sondage stratifié à un seul degré).

Dans ce cadre, nous proposons d'approfondir l'approche, mentionnée par Lavallée et Labelle-Blanchet [2], qui consiste à utiliser comme poids l'inverse de la probabilité d'inclusion théorique.

Nous détaillons le calcul des probabilités d'inclusion d'ordre 1 et 2. Cela permet notamment d'estimer la variance de l'estimateur obtenu à l'aide des résultats classiques de la théorie des sondages (que l'on trouve en [5] ou [6]). La méthode permet également de traiter le cas de la réunion de deux bases de sondage ayant des unités en commun – lorsqu'on ne peut identifier les doublons qu'au moment de la collecte.

Enfin nous donnons le résultat de simulations que nous avons menées sur des populations de type "entreprises".

Sur la base de ces simulations, nous concluons que cette méthode reposant sur un calcul exact des probabilités d'inclusion des unités finales aboutit -lorsque les poids initiaux sont dispersés - à des estimateurs plus précis que ceux reposant sur la MGPP, même lorsque les liens sont pondérés par une

variable auxiliaire. Dans ce dernier cas, il faut intégrer l'information provenant de cette variable auxiliaire dans le processus d'estimation, par un calage par exemple. En revanche, en menant nos simulations sur une population de type « ménages » (variables moins dispersées et taux de sondage homogènes), nous obtenons des résultats plus précis avec la MGPP qu'avec le calcul des probabilités d'inclusion exactes.

Bibliographie

[1] Pierre Lavallée. « Indirect sampling » Springer Series in Statistics, 2007.
[2] Pierre Lavallée et Sébastien Labelle-Blanchet. « Le sondage indirect appliqué aux populations asymétriques » Techniques d'enquête, Vol. 39, No 1, pp. 207-241, juin 2013.
[3] Arnaud Fizzala. « La gestion par partage des poids des changements de contour des entreprises dans l'Enquête Sectorielle Annuelle ». Acte des Journées de Méthodologie Statistique de l'Insee 2018.
[4] Ronan Le Gleut et Thomas Merly-Alpa. « L'impact du profilage sur la refonte du plan de sondage des Enquêtes Sectorielles Annuelles ». Acte des Journées de Méthodologie Statistique del'Insee 2018.
[5] Camilia Coga Cours de sondages dispensé à l'université de Besançon. http ://-goga.perso.math.cnrs.fr/
[6] Pascal Ardilly Les techniques de Sondage. Editions TECHNIP, 2006.