Journées de méthodologie statistique de l'Insee 2022

PROBABILISTES OU DÉTERMINISTES, DES MÉTHODES D'APPARIEMENTS AU BANC D'ESSAI DU PROGRAMME RÉSIL

Olivier Haag (*), Heïdi Koumarianos(**) , Lucas Malherbe(**)

(*) Insee, Direction des statistiques démographiques et sociales

(**) Insee, Direction de la méthodologie et de la coordination statistique et internationale

olivier.haag@insee.fr, heidi.koumarianos@insee.fr, lucas.malherbe@insee.fr

Mots-clés : appariement, sources administratives, Fellegi et Sunter

Domaine concerné : Combinaison de sources, données administratives

Le programme de Répertoires Statistiques d'Individus et de Logements (RéSIL) vise à construire un système de répertoires statistiques d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses.

Dans ce contexte, les appariements seront fondamentaux non seulement pour la constitution des répertoires mais aussi parce que le système de répertoires servira d'ossature au système d'information de la DSDS. Il permettra en effet l'appariement avec d'autres sources : données d'enquêtes, données administratives, voire données privées, dans la mesure où elles incluraient un identifiant commun avec le répertoire considéré, soit directement soit par le biais d'une identification préalable.

Ainsi, dans le but définir l'offre d'identification proposée par Résil, il a été décidé de tester différentes méthodes d'appariement afin de choisir celle(s) qui semble(nt) la plus efficace non seulement en termes de qualité statistique mais aussi d'un point de vue performance informatique (essentiel compte tenu des volumes à traiter).

Par ailleurs, la production de répertoires s'accompagnera d'une mesure de leur qualité, et en particulier de leur couverture afin de dépasser la situation actuelle où l'on constate des écarts entre les sources fiscales ou sociales et le recensement sans pouvoir les expliquer ni les imputer à l'une ou l'autre de ces sources. Elle sera mise en œuvre par comparaison entre la population définie par Résil et les données du recensement en utilisant la méthodologie capture-recapture et le modèle DSE (Dual System Estimation). Dans le cas de la France, l'existence de collectes annuelles de type recensement est un atout très important pour mesurer la montée en qualité d'un système de répertoires en construction. Tous les ans, les enquêtes annuelles de recensement couvrent 5 millions de logements et 9.3 millions d'habitants, ce qui fournit un échantillon de taille considérable pour servir de base à une telle opération.

Dans ce contexte, cet article comparera les résultats de l'appariement des individus présents dans la source fiscale (Fichier Imposable des Personnes) et l'EAR 2019 obtenus par différentes méthodes.

La première partie de l'article s'attachera à présenter les deux sources utilisées pour cet appariement en pointant les défauts de qualité qui auront un impact sur la qualité finale des appariements. Ainsi, par exemple, on constate que sur les variables identifiantes utiles pour l'appariement (nom, prénom, dates et lieux de naissance, adresse), la source fiscale présente beaucoup moins d'anomalies (0,8 % des individus) que l'EAR (13,2 %). Pour cette dernière on note une moindre qualité pour les personnes répondant par questionnaire papier qui comportent en plus des erreurs liées à la saisie, du fait de la qualité de certaines écritures manuscrites.

La deuxième partie présentera les différentes méthodes mises en œuvre :

- Rapsodie : Cet outil développé par le pôle « Revenus Fiscaux et Sociaux » de Rennes met en œuvre une méthode d'appariement déterministe. Un zoom sera fait sur l'intérêt d'un examen visuel d'un échantillon de paires issues du processus d'appariement. Il permet non seulement de mesurer la qualité de l'appariement (estimations des taux de faux positifs et faux négatifs) mais aussi de proposer des règles de gestion qui peuvent permettre d'améliorer l'appariement final lorsque des erreurs fréquentes sont identifiées (inversion des noms et prénoms dans l'EAR par exemple) ;
- Relais : Cet outil développé par Istat met en œuvre la méthode d'appariement probabiliste de Fellegi et Sunter. Cet outil dispose d'une IHM permettant de paramétrer les appariements (choix des distances et des seuils par exemple) et pourrait donc être utile dans le cadre de Résil. Il présente toutefois à ce jour des problèmes de performances qui empêchent l'appariement de « gros fichiers ». Le test de cet outil s'est donc fait uniquement sur les départements 48 et 53.
- Packages R et Python mettant en œuvre des méthodes probabilistes de type Fellegi et Sunter ainsi que des méthodes déterministes faisant intervenir du *machine learning*. L'objectif de ce test est plutôt de mesurer les performances que la méthodologie en tant que telle.

La troisième partie aura pour objectifs de comparer les résultats obtenus par ces différentes méthodes non seulement en termes de taux d'appariement mais aussi en termes de représentativité de la population des individus appariés. Ainsi, par exemple, l'appariement réalisé par Rapsodie permet de retrouver 92,7 % des individus de l'EAR dans FIP avec un taux de faux-positif de l'ordre de 0,2 %. Toutefois, la population appariée présente un biais de couverture des moins de 20 ans et à un degré moindre des plus de 90 ans, ainsi qu'une sous-représentation des individus nés à l'étranger.

Bibliographie

[1]Patrick JABOT, Pierre-Éric TREYENS (JMS 2010) "Appariement d'enquêtes avec des données administratives sociales ou fiscales "

[2] Cibela and ali (2010) "From theory to practice: the software RELAIS as a solution for record linkage" 2010

[3] HARRON, Katie, DIBBEN, Chris, BOYD, James, HJERN, Anders, AZIMAEE, Mahmoud, BARRETO, Mauricio L et GOLDSTEIN, Harvey, 2017. Challenges in administrative data linkage for research. *Big Data & Society* [en ligne]. décembre 2017. Vol. 4, n° 2

[4] CALIFORNIA POLICY LAB, 2018. *Linking Administrative Data: Strategies and Methods*. California Policy Lab.

[5] JAMES DOIDGE, PETER CHRISTEN, et KATIE HARRON, 2020. *Quality assessment in data linkage*. 2020. ONS.

[6] Zhang, L.-C., Dunne, J. (2017). "Trimmed Dual System Estimation." In Capture Recapture Methods for the Social and Medical Sciences, publié par D. Bo¨hning, J. Bunge, et P.v.d. Heijden: 239–259. Chapman and Hall/CRC.