Journées de méthodologie statistique de l'Insee 2022

TRAJAM : UN PANEL ADMINISTRATIF REPONDÉRÉ POUR PRENDRE EN COMPTE L'ABSENCE DE CLEF D'APPARIEMENT

Marie BOREL, Cindy REIST, Meddy KACY (*)

(*)DARES, Sous-direction Suivi et évaluation des politiques de l'emploi et de la formation professionnelle, département insertion professionnelle

marie.borel@travail.gouv.fr

Mots-clés : Appariement, repondération, correction de la non-réponse.

Domaine concerné : **Combinaison de sources, Contrôle et redressement des données, *data editing***

Résumé

Cet article présente l'utilisation de technique de redressement (correction de la non-réponse et calage sur marge) dans le cadre d'un appariement de données administratives.

La base TRAJAM est un panel administratif qui permet de suivre les jeunes de seize à trente-cinq ans dans des dispositifs d'insertion professionnelle et en emploi. Il est issu d'un appariement de douze bases. Pour les onze bases qui concernent les dispositifs d'insertion professionnelle, l'appariement se fait par un rapprochement avec une distance de Jaro-Winkler des données d'état civil (parmi lesquelles nom, prénom, date de naissance, lieu de naissance, NIR, même si toutes ces données ne sont pas présentes pour chaque individu). Les données d'insertion sont ensuite appariées aux données d'emploi, issues pour TRAJAM du panel DADS « TOUS SALARIES ». Les données identifiantes n'étant pas disponibles à la DARES, l'appariement se fait sur un identifiant adossé au NIR. Or dans TRAJAM, 8% de l'échantillon n'a pas de NIR. Ces individus sont donc retirés du panel puisqu'il est impossible de suivre leur parcours en emploi. Cependant, les individus sans NIR ont des caractéristiques particulières qui pourraient influer sur leur insertion (il y a notamment parmi eux une surreprésentation des jeunes de nationalité étrangère). Cette sélection pourrait donc provoquer un biais dans les calculs de taux d'insertion.

Afin de redresser l'appariement de cet éventuel biais, l'absence de NIR est traitée avec des méthodes similaires à celles de correction de la non-réponse. La méthode utilisée pour créer les groupes de réponse homogène sous-tendant la correction de la non-réponse est la méthode par arbre de décision (CHAID). Les variables prises en compte pour la construction des groupes de réponses sont principalement la nationalité, ainsi que le nombre et la nature des tables dans lesquelles le jeune est observé. Les poids sont peu déformés par cette correction de la non-réponse : ainsi 95% des individus ont un rapport de poids entre le poids de correction de la non-réponse et leur poids initial compris entre 1 et 1,06.

Suite à cette correction de la non-réponse, un calage sur marge est effectué. Les marges sont tirées des bases de données brutes présentes dans TRAJAM, afin de rendre les résultats représentatifs de

l'ensemble des jeunes passés par un dispositif. Sept jeux de poids sont créés : un poids par année (soit six en tout) ainsi qu'un poids transversal pour les études longitudinales, c'est-à-dire les études de trajectoires. Il a été décidé de multiplier les poids plutôt que de n'en garder qu'un seul afin de ne pas soumettre celui-ci à un trop grand nombre de marges, ce qui aurait à terme fortement déformé les poids de chaque individu. Chaque poids annuel est calé sur le nombre de mois passés par les jeunes dans chacun des dispositifs l'année d'intérêt, ventilé par âge (16-17,18-25, 26-35) et par sexe. Le calage se fait sur le nombre de mois par an passés par les jeunes en dispositif plutôt que sur le nombre de jeunes en dispositif afin de prendre en compte la saisonnalité des entrées qui, pour certains dispositifs (par exemple, les contrats d'apprentissage), peut être très forte.

On obtient ainsi un panel TRAJAM représentatif de sa population d'intérêt, à savoir les jeunes de 16 à 35 ans passés par un dispositif d'insertion professionnelle, malgré la sélection opérée à cause de l'absence de clef d'appariement entre données des dispositifs et d'emploi pour certains individus.

Bibliographie

[1] La correction de la non-réponse par repondération, Thomas Deroyon, Insee, 2017
[2] Les techniques de sondage, Pascal Ardilly.