

---

## **ENJEUX STATISTIQUES DU SYSTÈME D'INFORMATION INSERJEUNES SUR L'INSERTION PROFESSIONNELLE**

Nathalie CARON, Loïc MIDY (\*)

(\*) DEPP, Ministère de l'Éducation, de la Jeunesse et des Sports

[nathalie.caron@education.gouv.fr](mailto:nathalie.caron@education.gouv.fr)

[loic.midy@insee.fr](mailto:loic.midy@insee.fr)

**Mots-clés** : appariement sur identifiants indirects, sources administratives, mesure de l'insertion, modèles économétriques multiniveaux.

**Domaine concerné** : Appariements et fusion statistiques de données administratives (*record linkage*, appariement « exact » de fichiers), économétrie appliquée (étude de cas)

---

### **Résumé**

La Direction de l'Évaluation, de la Prospective et de la Performance (Depp) réalise depuis le début des années 90 deux enquêtes d'insertion annuelles permettant de suivre l'entrée dans la vie active des jeunes sortant d'apprentissage ou de voie professionnelle scolaire. Mais celles-ci ne permettent pas de publier des statistiques au niveau établissement, centre de formation d'apprentis et lycée professionnel, comme requis par la loi pour la liberté de choisir son avenir professionnel de 2018. Afin de répondre à ce besoin, la Depp et la Direction de l'Animation de la Recherche, des Etudes et des Statistiques (Dares) ont construit un nouveau dispositif, InseJeunes, basé sur l'appariement d'une vingtaine de sources administratives principalement sur identifiants indirects. En particulier, les appariements des bases des jeunes sortants du système éducatif avec la source Mouvement de Main d'Œuvre (MMO) de la Dares basée sur la Déclaration Sociale Nominative (DSN) permettent de savoir s'ils ont un emploi salarié en France. Le premier enjeu statistique de ce dispositif consiste à définir une méthodologie d'appariements robuste et rapide qui ne nécessite pas de reprise manuelle par des gestionnaires. Dans InseJeunes, chaque processus d'appariement comporte cinq étapes (normalisation des données, indexation, calcul de similarités, classification supervisée et évaluation de la qualité) et est réalisé via un outil d'appariement spécifique développé par la Depp. Le second enjeu statistique consiste à accompagner la publication des taux d'emploi par établissement par un indicateur de « valeur ajoutée » permettant de mesurer l'apport propre de l'établissement. En effet, une simple comparaison des taux d'emploi calculés entre établissements ne suffit pas car l'insertion dépend en partie de facteurs extérieurs à l'établissement : profil des jeunes, type et niveau de formation, spécialité de formation et marché du travail local. La notion de valeur ajoutée est déjà utilisée par la Depp lors de la diffusion chaque année des taux de réussite au baccalauréat par lycée (indicateurs de valeur ajoutée des lycées – IVAL-). Etant donné la mobilisation de

nombreuses données à caractère personnel (ex : noms, prénoms, identifiant national propre à chaque élève, étudiant ou apprenti qui est spécifique à la sphère éducation – INE-), l’aspect « sécurité » ainsi que la traçabilité complète des traitements réalisés sont également des enjeux importants pour le dispositif InserJeunes.

## **Bibliographie**

- [1] Christen, P., 2012, data matching concepts and technique for record linkage, entity resolution and duplicate detection, Springer
- [2] Givord P., Guillermin, M., 2016, « Les modèles multiniveaux », document « Méthodologie Statistique », Insee
- [3] Midy, L., 2021, « un outil d’appariement sur identifiants indirects : l’exemple du système d’information sur l’insertion des jeunes », Courrier des statistiques N6 – juillet 2021, Insee
- [4] Midy, L., Deschamps, G., 2021, « InserJeunes – calcul de valeur ajoutée », document de travail série « Méthodes » N°2021-M01, DEPP-MENJS