Appariements de données individuelles sur traits d'identité

Lucas Malherbe

Dans un contexte établi de taux de réponse aux enquêtes en baisse, il est naturel de chercher d'autres moyens d'obtenir de l'information. En parallèle, de plus en plus de sources de données administratives sont disponibles et offrent une opportunité d'enrichir l'information provenant des enquêtes, voire de la remplacer dans certains cas. Par exemple, utiliser l'information présente dans une source de données complémentaire pour retirer certaines questions d'une enquête permet d'alléger le fardeau de réponse. En outre, les sources administratives couvrent généralement un champ assez large, voire sont exhaustives, et certaines présentent des données d'excellente qualité. Mais ces opportunités s'accompagnent également de difficultés puisque les diverses sources appariées ne disposent en général pas d'identifiant commun. Dans ce contexte, l'appariement sur traits d'identité (nom, prénom, date de naissance, etc.) prend tout son sens et s'impose comme l'unique solution.

Apparier des données sur traits d'identité est cependant une tâche ardue, car les données à apparier présentent en général des erreurs ou des valeurs manquantes. S'il est possible de développer une solution simple en combinant quelques règles de façon assez rapide, cette approche ne donnera pas souvent les meilleurs résultats.
Plusieurs méthodes existent, mais aucune ne s'est imposée comme la solution de choix pour tous les cas. Les problèmes d'appariements dépendent en effet beaucoup des données à apparier, et en particulier de leur volume et de leur qualité. Cet atelier fournira des éléments pour choisir une méthode d'appariement en fonction du cas rencontré.

L'atelier sera articulé de la façon suivante :

- Une discussion des enjeux et des méthodes associées aux appariements sur traits d'identité ;
- Une présentation des principaux outils existants, de leurs avantages et inconvénients et des cas d'usage associés ;
- La démonstration de deux cas d'usage d'appariements :
    - un appariement à usage unique de quelques dizaines de milliers d'individus, avec des contraintes temporelles fortes sur le temps de développement mais une certaine latitude sur la qualité du résultat,
    - un appariement de données volumineuses (plusieurs millions de lignes), devant éventuellement s'insérer dans un processus de production ;
- Un ensemble de conseils pratiques et d'erreurs à éviter.

Note biographique

Lucas Malherbe travaille à l'Insee, où il mène des expérimentations autour des nouvelles sources de données et des méthodes de data science relatives aux productions statistiques du système de statistique publique. Il est membre d'un groupe sur les sources administratives, au sein duquel il mène un investissement méthodologique sur les appariements. Il est l'auteur d'un document de travail méthodologique sur les appariements de données individuelles et participe aux réflexions sur les outils d'appariements utilisés à l'Insee.