Journées de méthodologie statistique de l'Insee 2022

Méthodologie des appariements de données individuelles

Lucas MALHERBE (*)

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

lucas.malherbe@insee.fr

Mots-clés : appariement, enrichissement de données

Domaines Appariements, Données administratives

Résumé

Avec de plus en plus de sources de données individuelles existantes, l'information disponible pour le statisticien public devient conséquente. Cependant, toutes ces sources ne sont pas pensées et construites de la même façon, elles ne poursuivent pas le même objectif et ne couvrent pas la même population. C'est dans ce contexte que l'appariement de données individuelles prend tout son sens. Il consiste à rapprocher deux bases de données d'origine distincte partageant des unités statistiques communes mais contenant des informations différentes.

La tâche est aisée si les deux bases disposent d'un identifiant unique commun pour tous les enregistrements, comme un numéro de sécurité sociale ou un code statistique non signifiant. En l'absence d'un tel identifiant ou lorsque celui-ci n'est pas de bonne qualité, l'appariement se fait alors sur une combinaison d'autres champs (état civil, adresse, etc.) qui eux-mêmes peuvent présenter des défauts. Il s'agit ainsi d'utiliser une méthode permettant de repérer, parmi toutes les paires possibles du produit cartésien des deux bases, lesquelles correspondent à un seul et même individu ; et ce malgré des informations plus ou moins erronées sur les champs servant à l'identification.

Les enjeux sont multiples, allant de la construction de répertoires au repérage de doublons en passant par l'enrichissement de données. L'appariement de données individuelles trouve particulièrement sa place dans le cadre d'une exploitation généralisée des sources administratives, permettant notamment de remplacer en partie les données d'enquêtes par des données administratives.

L'article sera structuré autour d'un découpage possible des principales étapes d'un processus d'appariement.

Préparation des données

Cette première phase consiste en une succession d'étapes de nettoyage et de normalisation pour préparer les deux bases à la comparaison. Il s'agit de se débarrasser des fioritures, qui n'apportent pas particulièrement d'information mais nuisent à l'appariement, comme les accents

ou la capitalisation des lettres dans un nom. Le résultat d'un appariement est intimement lié à la qualité des données utilisées. Porter un soin particulier à cette étape est un investissement indispensable pour réaliser un bon appariement.

Indexation
L'idéal pour effectuer un appariement entre deux fichiers serait de comparer l'ensemble des paires du produit cartésien. Malheureusement, il est souvent impossible de procéder de la sorte en raison du temps de calcul associé aux comparaisons de champs textuels. L'objectif de l'indexation est alors de réduire la dimension en ne considérant que les paires qui ont une chance raisonnable de correspondre à la même entité. La méthode la plus répandue est celle du blocage.

Comparaison des champs
Cette étape consiste à calculer des mesures de similarité pour chaque champ intervenant dans l'appariement et pour chaque paire potentielle conservée à l'issue de l'indexation. De nombreuses mesures existent pour chaque type de champ (texte, nombre, date, etc.) et leur choix est étroitement lié à celui de la méthode de classification choisie.

Comparaison des paires : classification
Les mesures de similarité calculées à l'étape précédente sont mobilisées pour classer les paires en deux ou trois catégories : les paires liées, les paires non liées, et parfois des paires laissées en suspens pour un éventuel examen manuel. De nombreuses méthodes existent et se rangent en général en deux groupes, les méthodes déterministes et les méthodes probabilistes. Ces dernières se distinguent par l'utilisation d'une probabilité : celle que les deux éléments d'une paire correspondent à un même individu. Aucune méthode ne s'impose réellement face aux autres, le choix doit être guidé par le type de données à apparier ainsi que les objectifs poursuivis.

Résolution des conflits
Cette étape n'a lieu que s'il existe des restrictions sur les combinaisons de paires qui peuvent être liées, comme dans le cas d'un appariement *one-to-one* ou *many-to-one*. Des conflits peuvent apparaître, comme le fait de lier deux individus différents du premier fichier à un même individu du second fichier. Il est alors nécessaire de trancher, en s'appuyant sur les distances ou les probabilités issues des étapes précédentes.

Évaluation de la qualité
Une fois l'appariement terminé, l'évaluation de sa qualité est primordiale, à la fois pour procéder à d'éventuels ajustements dans le processus et pour déterminer si son résultat satisfait les critères nécessaires à une utilisation dans une étude statistique par exemple. L'objectif est de rassembler autant d'informations que possible sur la qualité de l'appariement. La difficulté réside dans le fait que le vrai statut de chaque paire n'est en général pas connu. Il est nécessaire dans ce cas d'annoter manuellement un échantillon de paires pour être en mesure d'en déduire différents indicateurs.

L'article couvre également des aspects pratiques de mise en oeuvre des méthodes présentées et propose un ensemble d'enseignements issus de différents cas d'usage d'appariements au sein du SSP.

Bibliographie

[1] Christen, Peter. 2012. Data matching : concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media.

[2] Fellegi, Ivan P and Alan B Sunter. 1969. "A theory for record linkage." Journal of the American Statistical Association 64(328) :1183–1210.