Journées de méthodologie statistique de l'Insee 2022

IMPACT DU NETTOYAGE DES DONNÉES SUR LA QUALITÉ D'UN APPARIEMENT

Heidi KOUMARIANOS (*)

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

heidi.koumarianos@insee.fr

Mots-clés : appariement, nettoyage de données, qualité

Domaine concerné : Combinaison de sources, données administratives

Résumé

L'Insee souhaite favoriser l'utilisation massive et plus industrialisée des sources administratives et ce notamment dans la sphère des statistiques socio-démographiques. C'est dans ce cadre notamment que se place le programme de Répertoires Statistiques d'Individus et de Logements (RéSIL) qui vise à construire un système de répertoires d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses.
Les appariements entre sources vont être encore plus au cœur du système d'information de l'Insee.
Il convient donc de mettre en place les processus les plus automatisés et les plus robustes possibles afin d'obtenir les gains d'efficience et de qualité souhaités.

Un processus d'appariement se décompose en plusieurs phases .

- le nettoyage des données ;
- la réduction de la taille du problème (blocage ou indexation) ;
- l'appariement des unités statistiques proprement dit ;
- la comparaison des unités au sein des paires ;
- la classification des paires retenues.

Cet article se concentrera sur la première de ces phases, en cherchant à mesurer son impact sur les suivantes.

Lors d'un processus d'appariement, l'étape de nettoyage des données en entrée revêt une grande importance : elle s'assure du respect des formats de données, de la comparabilité des informations, et permet d'adapter les règles d'appariement aux caractéristiques des jeux de données.
Identifiée comme cruciale et la plus chronophage par les statisticiens en charge d'opérations d'appariement, elle fait rarement l'objet de mesures d'efficacité.

On présentera d'abord les enjeux de l'étape de standardisation et de nettoyage de données individuelles. Cette étape s'appuie sur une analyse préalable de la qualité des données, notamment par le biais de contrôles formels, ou de conformité à un référentiel (géographique par exemple) et/ou d'appartenance à des nomenclatures.

Après ce rappel théorique, ces traitements seront effectués sur différents jeux de données individuelles (données fiscales et données de l'enquête annuelle de recensement notamment), pour essayer de différencier l'impact de la normalisation sur des données de qualité et d'origine (données administratives versus données d'enquêtes) différentes. Plusieurs scénarios seront comparés, en mobilisant différents niveaux de nettoyage : une première normalisation minimale pour rendre possible les comparaisons, une normalisation plus classique (suppression des caractères non attendus, conformité à un référentiel géographique), et un nettoyage de données plus poussé (exclusion d'enregistrements de trop faible qualité, modification plus importante des chaînes de caractères pour éliminer civilités, relations…).

Dans chacun des cas étudiés, on essaiera de mesurer la charge de travail liée à la normalisation, et son impact sur les données (part de données modifiées, enregistrements exclus…).
On analysera ensuite son impact sur les phases ultérieures du processus d'appariement.
On s'attachera enfin à essayer de quantifier l'impact de la normalisation et du nettoyage des données sur le résultat de l'appariement (nombre d'enregistrements appariés, de façon exacte ou floue, impact sur la déformation des distributions, estimations de faux positifs et faux négatifs), en lien avec la méthode d'appariement retenue (probabiliste versus déterministe).

Bibliographie

[1] RANDALL S. M., FERRANTE A. M., BOYD J. H., SEMMENS J. B, « The effect of data cleaning on record linkage quality », *BMC Medical Informatics and Decision Making,* Vol. 13, n° 1, décembre 2013
[2] DOIDGE J., CHRISTEN P. et HARRON K., « Quality assessment in data linkage », published by the UK Government Analysis Function and Office for National Statistics as part of the National Statistician's Quality Review : Joined up data in government: the future of data linking methods, dernière mise à jour juin 2021
[3] CHRISTEN P., « Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection », Springer, 2012
[4] CHRISTEN P., CHURCHES T. et XI ZHU J., 2002. « Probabilistic Name and Address Cleaning and Standardisation », 2002
[5] SANMARTIN C., « *Modèle du processus d'un projet de couplage d'enregistrements* », Statistique Canada, 2017
[6] TUOTO T., CIBELLA N., FORTINI M. et SCANNAPIECO M., « From theory to practice: the software RELAIS as a solution for record linkage », 2010
[7] GILL L., « Methods for Automatic Record Matching and Linkage and their Use in National Statistics », National Statistics Methodological Series no. 25, 2001