Journées de méthodologie statistique de l'Insee 2022

REVUE DU COUPLAGE D'ENREGISTREMENTS À STATISTIQUE CANADA

Abdelnasser SAÏDI *(*)*

(*) Statistique Canada, Division des Méthodes et de l'Intégration Statistique

Abdelnasser.saidi@statcan.gc.ca

Mots-clés : couplage probabiliste, déterministe, méthodologie de Fellegi-Sunter, faux positifs, faux négatifs

Domaine concerné : ***Combinaison des sources***

Résumé

Les situations nécessitant de coupler plusieurs fichiers sont courantes. Le couplage a alors pour but d'apparier des enregistrements provenant de deux fichiers différents. Il arrive aussi parfois qu'on veuille apparier de façon interne les enregistrements d'un seul fichier dans le but d'y détecter d'éventuels duplicates. Souvent, les enregistrements qu'on tente d'appariés n'ont pas d'identificateur unique. L'appariement hiérarchique, le couplage probabiliste et l'appariement statistique sont alors des méthodes qui peuvent être utilisées.

À Statistique Canada, apparier des données sans identificateurs uniques est fait en utilisant aussi bien un couplage déterministe hiérarchique ad-hoc que le couplage probabiliste selon Fellegi-Sunter [1].

Nous présentons une revue des défis du couplage probabiliste et des outils développés pour en déterminer les paramètres optimaux. Un système généralisé codé en SAS appelé G-coup inclut une implémentation complète de la théorie de Fellegi-Sunter dans une interface graphique. La dernière version de G-Coup permet de faire aussi en complément un couplage déterministe fondée sur les profils d'agrément. Les spécificités de G-coup ont trait à des comparateurs de champs comportant la notion d'accord partiel, des comparaisons matricielles, conditionnelles et définissables par l'utilisateur. Le critère de classification fondée sur des seuils comporte des défis dus à la difficulté d'estimer la distribution des composantes du score de Fellegi-Sunter. Nous discuterons les différentes approches pour déterminer ces seuils.

L'analyse des données liées est aussi un problème crucial. Traiter naïvement les données liées comme si elles ne comportaient pas d'erreurs de couplage peut conduire à une inférence biaisée. On présente un cadre de travail qui exploite les données liées tout en tenant compte des faux négatifs.

Bibliographie

[1] Fellegi I.P., Sunter A.B., « A Theory of Record Linkage », *Journal of American Statistical Association* JASA, vol 64, pp 1183-1210, 1969.