L'APPARIEMENT D'ENQUÊTES AVEC DES DONNÉES ADMINISTRATIVES SOCIALES OU FISCALES

Patrick JABOT, Pierre-Éric TREYENS

Insee, Direction régionale de Bretagne, Pôle revenus fiscaux et sociaux

patrick.jabot@insee.fr
pierre-eric.treyens@insee.fr

Mots-clés : appariement (« data matching », « record linkage », « object identification »), enquêtes, fichiers administratifs

Résumé

Depuis plusieurs années, l'exploitation conjointe des données d'enquête et des fichiers administratifs se développe. Elle permet en effet de conjuguer les qualités respectives des deux modes de collecte de l'information tout en limitant la charge de réponse pour les individus et les entreprises. En France, pour des raisons juridiques notamment, le rapprochement des enquêtes et des données administratives est rarement réalisé grâce à des identifiants personnels qui faciliteraient la fusion des fichiers. Le rapprochement des sources nécessite alors la construction de clés, créées à partir de variables discriminantes disponibles dans les deux bases, permettant l'identification des observations.

Ce processus, qualifié **d'appariement**, est notamment utilisé à l'Insee par le pôle Revenus fiscaux et sociaux pour enrichir les enquêtes du système statistique public (SSP) avec des données de revenus. Le taux de pauvreté monétaire (Enquête Revenus fiscaux et sociaux, ERFS) ou la part des dépenses de logement dans le budget des Français (Enquête Budget de Famille, BDF) sont des indicateurs qui découlent de ces opérations d'appariement.

En première approche, **l'appariement** (dénommé « data matching », « record linkage » ou «object identification » en anglais) consiste donc dans l'association d'individus présents dans plusieurs fichiers distincts mais comportant des variables communes. Cette opération requiert différentes étapes : la standardisation des variables, la création de clés identiques, la recherche plus ou moins systématique de ces clés dans les fichiers et la suppression éventuelle des doublons lorsqu'une recherche produit plusieurs échos. Chacun de ces objectifs peut être poursuivi en adoptant des techniques et méthodes distinctes qui influent sur la qualité des résultats et la durée des traitements. Cet exposé se propose de décrire certains choix adoptés au pôle Revenus fiscaux et de présenter des expérimentations récentes destinées à améliorer les appariements réalisés.

Ces expérimentations ont porté sur trois étapes qui concentrent généralement les difficultés de l'opération d'appariement : le mécanisme de création des clés, la recherche systématique des clés dans les fichiers et l'algorithme de sélection des enregistrements lorsque plusieurs observations pourraient être retenues pour l'appariement. La normalisation initiale du fichier, souvent gourmande en temps, a donc été ignorée dans ce travail, notamment parce qu'elle s'avère extrêmement dépendante de la qualité des données. La **création des clés** est actuellement très largement manuelle et fondée sur une connaissance des sources : une option consiste dès lors à produire ces identifiants de manière industrielle en distinguant notamment des clés fortes et faibles. Une approche naïve suggère ensuite **une recherche systématique des clés** d'enquête dans le répertoire administratif. Cette démarche s'avère peu efficace et se heurte à une contrainte de ressources informatiques. Une amélioration possible du process consiste alors à découper le répertoire en sous-

blocs (« blocking ») afin de ne comparer que des observations qui peuvent potentiellement s'apparier. Enfin, **la sélection des enregistrements** s'opère en deux temps. Lorsque deux enregistrements sont identifiés par l'égalité de clés fortes, ils sont immédiatement écartés de la recherche. En revanche, une distance entre des clés moins discriminantes, définie préalablement, ouvre la possibilité d'appariements vraisemblables entre les autres observations. Un seuil de décision, fixé a priori, distingue entre les observations appariées et les non appariées. Une comparaison des résultats de ces tests avec les enrichissements déjà réalisés esquissera des pistes de réflexion pour des améliorations futures.