5324 EUROS DE L'HEURE : OUTLIER OU FOOTBALLEUR ? MÉTHODES D'APPRENTISSAGE NON SUPERVISÉ POUR LA DÉTECTION D'ANOMALIES : APPLICATION AU CAS DE LA DÉCLARATION SOCIALE NOMINATIVE

MARIE CORDIER-VILLOING (**), THOMAS DEROYON(*), JULIE DJIRIGUIAN(*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale
(**) Insee, Direction des Statistiques Démographiques et Sociales

Mots-clés *(6 maximum)* : Machine learning, big data, données administratives

Résumé (entre 350 et 800 mots maximum)

Cet article s'inscrit dans le projet global de refonte des Déclarations Annuelles de Données Sociales (DADS) qui vise à générer automatiquement des fichiers mensuels et un ensemble de fichiers annuels de Déclarations Sociales Nominatives (DSN), analogues pour ces derniers aux Déclarations Annuelles de Données Sociales qui servent de source de référence pour l'analyse des salaires en France. Cette refonte, qui se traduit par une reconstruction complète des chaînes de production des données sur les emplois et les salaires, est l'occasion de repenser la méthodologie des différents processus et traitements statistiques qui leur sont appliqués. Cet article porte sur un de ces traitements particuliers, la détection des erreurs, qui constitue la première étape du contrôle et redressement (ou *data editing*) des données de la DSN ou des DADS. Ce sujet a en effet fait l'objet de nombreux développements méthodologiques, notamment issus du champ de l'apprentissage statistique. La disponibilité de nouvelles données mensuelles massives incite à tester des méthodes de *machine learning* pour détecter automatiquement les anomalies.

Cette présentation a pour but de rendre compte des travaux qui ont été réalisés dans ce cadre, notamment des difficultés que nous avons rencontrées. Ces difficultés tiennent aux données elles-mêmes, et à la nature du problème que nous essayons de traiter.

Nous ne disposons en effet pas d'un échantillon de données de la DSN ou des DADS dans lesquelles les erreurs et les observations valides sont identifiées. À partir des données actuellement disponibles, il est seulement possible de rechercher des anomalies, *i.e.* des observations atypiques, *via* un apprentissage non supervisé. Cet exercice est cependant beaucoup plus difficile et très dépendant de la manière dont est défini ce qui constitue ou pas une anomalie. Chaque algorithme disponible repose en effet sur une représentation spécifique de ce qu'est une donnée normale et une donnée atypique et cherche, sur la base de cette définition, à identifier les secondes. Il est de plus impossible, sans échantillon d'observations dont la qualité est connue, de juger de la pertinence de la détection opérée par les différentes méthodes : quelle part des anomalies qu'elles identifient correspond

effectivement à des erreurs ? Quelle part des erreurs dans les données arrivent-elles à détecter ?

Les travaux que nous avons réalisés nous ont permis de comparer les caractéristiques des anomalies identifiées par trois méthodes issues de l'apprentissage statistique avec les méthodes actuellement utilisées dans la chaîne de traitement des DADS, reprises dans celle de la DSN. Ils montrent que les algorithmes détectent en grande partie des anomalies différentes, en fonction de la manière dont ils définissent et identifient les erreurs présumées. L'utilisation combinée de plusieurs algorithmes de détection d'erreurs permettrait ainsi de couvrir un spectre plus large d'erreurs potentielles.

Bibliographie