Anonymisation de données individuelles : bien calées, bien protégées ?

L'objectif de cette communication est de présenter une comparaison entre différentes méthodes d'anonymisation d'un fichier de données individuelles, et d'introduire une méthode originale de protection fondée sur un sous-échantillonnage et une correction des pondérations, permettant d'obtenir *in fine* un fichier anonymisé. L'expérimentation est réalisée sur l'enquête « Vols, violences et sécurité » (VVS), qui est le volet Internet de l'enquête « Cadre de vie et sécurité » (CVS), menée par l'Insee auprès des ménages.

Dans un premier temps, les données utilisées pour le test sont présentées, en distinguant les variables indirectement identifiantes et les variables d'intérêt sensibles. L'objectif de réduction du risque de divulgation est introduit : le but de cette démarche d'anonymisation est d'obtenir au final un fichier *k*-anonymisé. Un fichier est considéré comme *k*-anonymisé si, pour toute combinaison des variables indirectement identifiantes choisies, *k* répondants au moins sont représentés ; ils sont par conséquent indistinguables par un attaquant.

Différentes méthodes pour atteindre cet objectif sont ensuite présentées :

- Réduction du niveau de détail utilisé pour les variables indirectement identifiantes ;
- Suppressions de modalités pour certaines variables indirectement identifiantes permettant la ré-identification ;
- Une nouvelle méthode, qui s'appuie sur un sous-échantillonnage et un calage sur marges. Les individus dont les caractéristiques identifiantes sont rares sont supprimés (sous-échantillonnage ciblé sur les individus dont le risque de ré-identification est jugé faible). Pour corriger le biais de sélection engendré par le sous-échantillonnage, des calages sur marges sont réalisés *a posteriori*.

L'efficacité des différentes méthodes est enfin comparée en considérant des mesures d'utilité (en termes d'information statistique) des fichiers obtenus. En particulier, des statistiques multivariées et des modèles de régression logistique sont calculés à partir des différentes bases de données obtenues avec les méthodes d'anonymisation, et comparés aux résultats obtenus à partir des données originales, afin de quantifier la perte d'information engendrée par les mécanismes de protection.

Auteur de la proposition de communication : Maxime Bergeat, Département des méthodes statistiques, Insee
maxime.bergeat@insee.fr
01 41 17 64 86