UNE CONDITION NÉCESSAIRE ET SUFFISANTE POUR MINIMISER LE RISQUE D'INFÉRENCE PRÉDICTIVE DANS LES MÉTHODES DE PERTURBATION DE DONNÉES

Cédric WANKO(*), Jean-Marie BERTON (**)

(*) Laboratoire TRIS (Traitement et Recherche sur l'Information Statistique), Montpellier Université
(**) Membre Rattaché TRIS

cedric.wanko@free.fr

***Mots-clés** : distribution bivariée, indépendance, perturbation, inférence prédictive, orthonormalisation de Gram-Schmidt*

Résumé

Dans la nécessaire régulation liée au "Contrôle de la Divulgation Statistique"- Statistical Disclosure Control - les méthodes de perturbation des données à l'aide d'un bruit aléatoire permettent de réduire les pertes informationnelles et prennent en compte les données non confidentielles favorisant les risques de divulgation prédictive dans les modèles de codage. En effet, Il est important dans les différents modèles de codage de tenir compte des variables non confidentielles face aux problèmes de ré-identification des données confidentielles. On ne peut pas tout masquer, il faut de fait prendre en compte l'existence de variables publiques non confidentielles dans le processus de codage et des risques qui leurs sont afférents. Dans le but de lutter contre cette perte d'information tout en conservant les caractéristiques statistiques essentielles, notre attention s'oriente naturellement vers une méthode de perturbation des données à l'aide d'un bruit aléatoire ; une méthode utilisée par Burridge (2003) sous l'appelation de « Information Preserving Statistical Obfuscation (IPSO) » et généralisée via un coefficient de similarité par Muralidhar and Sarathy (2008) puis Domingo-Ferrer and Gonzàlez-Nicolàs (2010) et enfin Calviño (2017) sous les appelations respectives de « Sufficiency Based Noise Addition (SBNA) », « MicroHybrid method (MH) » et « Principal Component Analysis method (PCA) ». Par ailleurs, on sait que tenir compte des données non confidentielles favorisant les risques de divulgation prédictive dans le modèle de codage contribue à renforcer la sécurité du modèle vis à vis de ces risques de divulgation prédictive. Malgré tout, ces risques sont toujours sensibles aux différentes méthodes de prédiction relatives au *Data Mining* (Donoho (2015)) car même si les variables confidentielles sont supprimées du modèle, les variables non confidentielles sont encore présentes dans le processus de codage et constituent malgré tout un lien entre les données masquées et publiquement divulguées et les données confidentielles. Par conséquent, pour minimiser les risques de divulgation prédictive, on propose de minimiser l'influence des variables non confidentielles tout en les conservant au sein du modèle. Les variables non confidentielles doivent donc être peu ou pas significatives dans la spécification du modèle tout en conservant les propriétés théoriques de base du modèle. Soient $\mathbf{f(.)}$ la densité de probabilité et $\mathbf{F(.)}$ sa primitive. Dans ce papier, on définit une condition nécessaire et suffisante montrant que lorsque les variables non confidentielles sont orthonormalisées et perturbées en leurs structure alors elles satisfont la "predictive inference risk requirements"

$$f(\mathcal{Y}) = f(\mathcal{X}) \quad et \quad f(\mathcal{Y},\mathcal{S}) = f(\mathcal{X},\mathcal{S})$$

qui minimise les risques de divulgation prédictive tout en conservant respectivement les "data utility

requirements" et "disclosure risk requirements"

$$f(Y) = f(X) \text{ et } f\left(Y, \tilde{S}\right) = f\left(X, \tilde{S}\right)$$

$$f\left(X/\tilde{S}, Y\right) = f\left(X/\tilde{S}\right)$$

(Muralidhar and Sarathy (2003)) avec

$$\tilde{S} = \left(\tilde{S}_1, \tilde{S}_2\right)$$

le vecteur orthonormalisé et transformé des données non confidentielles $S = (S_1; S_2)$, les variables confidentielles $X = (X_1;X_2)$ et les variables masquées et publiquement divulguées $Y = (Y_1;Y_2)$. On fournit une approche théorique puis calculatoire avant de confronter les résultats à différentes méthodes de perturbation linéaires et non linéaires puis de fournir un exemple empirique.

Bibliographie

[1] Burridge J. , « Information preserving statistical obfuscation ». *Statistics and Computing* , vol 13, pp 321-327, 2003.

[2] Calvi˜no A., « A Simple Method for Limiting Disclosure in Continuous Microdata Based on Principal Component Analysis ». *Journal of Official Statistics*, 33, 1, pp 15-41, 2017.

[3] Domingo-Ferrer, J. and U. Gonzalez-Nicol´as, « Hybrid Microdata Using Microaggregation. » *Information Sciences*, 180, pp 2834-2844, 2010.

[4] Donoho D., « 50 years of Data Science ». Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf, 2015.

[5] Muralidhar K. and R. Sarathy, « A theoretical basis for perturbation methods ». *Statistics and Computing*, 13, pp 329-335, 2003.

[6] Muralidhar, K. and R. Sarathy, « Generating Sufficiency Based Nonsynthetic Perturbed Data ». *Transactions on Data Privacy*, 1, pp 17-33, 2008.