Journées de méthodologie statistique de l'Insee 2022

GESTION DU SECRET POUR LA DIFFUSION GRAND PUBLIC DE CUBES MULTIDIMENSIONNELS : UNE EXPÉRIMENTATION AU SSM AGRICULTURE

Michael LEVI-VALENSIN
Service de la Statistique et de la Prospective -
Bureau des Méthodes et de l'Informatique Statistiques

michael.levi-valensin@agriculture.gouv.fr

Mots-clés : secret statistique, optimisation, automatisation

Domaine concerné : Institutionnel, open science- confidentialité, diffusion

---

Résumé

Comme pour toutes les sources, la mise en ligne des données agricoles sur le site Agreste exige de mettre sous secret certaines valeurs non diffusibles selon les méthodes dites suppressives historiquement appliquées dans les services statistiques français.
Cette obligation est régie par le principe 5 du Code de Bonnes Pratiques de la Statistique Européenne sur le secret statistique et la protection des données.

Une difficulté supplémentaire survient lors de la publication de tableaux interactifs qui croisent plusieurs variables à différents niveaux de nomenclature. Il est nécessaire de considérer tous les croisements possibles et les éventuels secrets induits générés.
Ce travail était jusqu'alors réalisé de manière heuristique par une fonction conçue en interne qui appliquait les règles de secret induit mais sans optimisation de la perte d'information ni la certitude absolue de ne pas retrouver les valeurs des cellules masquées.

Le package **sdcTable** développé par Statistics Austria (INS d'Autriche), regroupe les fonctionnalités de τ-Argus pour appliquer les secrets primaire et induit et permet d'appliquer les méthodes suppressives dans une chaîne de traitements statistiques sous R.
Rappelons que la gestion du secret induit s'apparente à un programme d'optimisation sous contraintes : minimiser une perte d'information tout en respectant certains sous-totaux.

$$
\begin{aligned}
\min \quad & \sum_{i=1}^{n} w_i y_i \\
\text{s.t.} \quad & \\
& \left.\begin{aligned} Ax^{l,p} &= 0 \\ -a_i y_i \leq\ x_i^{l,p} &\leq M y_i \qquad i = 1 \ldots n \\ x_p^{l,p} &\leq -lpl_p \end{aligned}\right\} \\
& \left.\begin{aligned} Ax^{u,p} &= 0 \\ -a_i y_i \leq\ x_i^{u,p} &\leq M y_i \qquad i = 1 \ldots n \\ x_p^{u,p} &\geq upl_p \end{aligned}\right\} \\
& y_i \in \{0, 1\},
\end{aligned}
$$

Cette présentation est un retour d'expérience sur l'utilisation de packages R pour appliquer les méthodes suppressives à des bases de données agricoles. Par la suite, ces dernières sont intégrées dans une chaîne de traitement jusqu'à la mise en ligne sur le site Agreste sous forme de tableaux dynamiques.

Plusieurs tests ont ainsi été réalisés sur des fichiers comme l'enquête SRI sur les sciages de bois par essence ou les Surfaces Agricoles Utiles par commune au Recensement Agricole 2010, voire sur le millésime 2020. La mise en forme des tables et la construction des fichiers hiérarchiques sont des étapes préalables mais indispensables avant toute application des règles.

Des méthodes perturbatrices de type « cellule clé » ont également été envisagées et testées grâce aux packages ***ptable*** *pour construire les tables de perturbation et* ***cellkey*** *pour les appliquer. Mais l'absence de cadre législatif fixant des règles en matière de perturbation des données ainsi que des difficultés techniques n'ont pas permis d'aboutir.*

Bibliographie

[1] CSO Best Practice for Statistical Disclosure Control of Tabular Data - Timothy Linehan and Karina Dineen : Office statistique irlandais
https://www.cso.ie/en/search/?addsearch=cso%20best%20practice%20for%20statistical%20disclosure%20control%20of%20tabular%20data

[2] Privacy in Statistical Databases. CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004, Proceedings

[3] CSO Best Practice for Statistical Disclosure Control of Tabular Data