

{qdd} : un package R de nettoyage et de qualité des données pour les Plateformes d'Epidémiosurveillances

Marine Marjou¹³, Marie Grosdidier¹³, Charlotte Rüger²³⁴⁵, Pauline Bres²⁵

Résumé (max 300 mots)

Les trois Plateformes d'Epidémiosurveillance françaises : en santé animale (ESA), en santé végétale (ESV) et sécurité de la chaîne alimentaire (SCA) centralisent et travaillent sur des données issues de divers plans de surveillance nationaux. Ces données sont multiples et variées et un travail de nettoyage en amont est nécessaire afin de pouvoir les analyser et les exploiter. Le package {qdd} peut s'utiliser selon deux angles différents mais complémentaires. Il permet notamment de décrire d'un point de vue quantitatif la qualité d'un jeu de données en générant automatiquement un rapport (par exemple le pourcentage de données manquantes, de coordonnées géographiques en dehors du périmètre d'étude, etc.). Ce rapport permet de mieux connaître les données et de faire un retour aux fournisseurs de données dans le but d'améliorer le recueil des données des plans de surveillance. Dans un deuxième temps, le package propose également des fonctions qui permettent de nettoyer facilement et rapidement les données de types calendaire, caractères ou des coordonnées géographiques.

Les trois grands principes qui servent de base au développement du package sont la complétude, le format et la cohérence des données.

Le package est en cours de conception.

Mots-clefs : data – package – qualité des données – nettoyage – épidémiologie – épidémiosurveillance –

Développement

Les données issues des différents plans de surveillance sont souvent soit des données récoltées *à la main* via des logiciels spécifiques soit par simple remplissage de tableaux Excel. Dans les deux cas, bien souvent, il n'y a pas de vérification dans le remplissage des informations. Il n'est donc pas rare de se retrouver avec, dans un même fichier, des doublons, des coordonnées géographiques dans des systèmes géographiques différents ou en dehors de la zone d'étude, des dates incohérentes, etc.

Pour ce faire nous avons établis, avec un groupe restreint de personnes travaillant sur les données des trois plateformes d'épidémiosurveillances, une liste de concepts à évaluer dans notre package :

- La complétude (existe-t-il des données manquantes ?)
- La cohérence (la date de prélèvement est-elle antérieure à la date de réception du prélèvement ? Les coordonnées géographiques de mon prélèvement correspondent-elles à la commune qui est renseignée dans mes données ?)
- Le format (le code INSEE de la commune du prélèvement contient-il 5 chiffres ? les coordonnées géographiques sont-elles renseignées selon le système géographique de référence ?)

¹³ INRAE, UR BioSP : marine.marjou@inrae.fr – marie.grosdidier@inrae.fr

²⁵ Anses, Unité Epidémiologie et appui à la surveillance (EAS) : charlotte.ruger@anses.fr – pauline.bres@anses.fr

³ Plateforme d'Epidémiosurveillance en Santé Végétale (ESV)

⁴ Plateforme d'Epidémiosurveillance en Santé Animale (ESA)

⁵ Plateforme d'Epidémiosurveillance Sécurité de la Chaîne Alimentaire (SCA)

À partir de ces trois concepts et d'une liste relativement exhaustive des types de données que nous rencontrons souvent, nous avons établis (et continuons d'établir en fonction des nouveaux datasets) une liste de fonctions qui nous permet de calculer des indicateurs de qualité (définis par les trois concepts précédent) et des fonctions qui permettent de nettoyer les données. Un rapport sur la qualité des données peut également être généré automatiquement en faisant appelle à une fonction générique de l'ensemble des fonctions de qualité des données.

Ce package est encore en cours de conception. Son objectif est d'homogénéiser les pratiques de nettoyage des données entre les trois plateformes et les agents des trois plateformes. Le deuxième objectif étant d'avoir un suivi de la qualité des données dans le temps afin de pouvoir faire un retour aux fournisseurs de données dans le but d'améliorer les différents plans de surveillance.

Références

Plateforme ESV : <https://plateforme-esv.fr/>

Plateforme ESA : <https://plateforme-esa.fr/fr>

Plateforme SCA : <https://www.plateforme-sca.fr/>