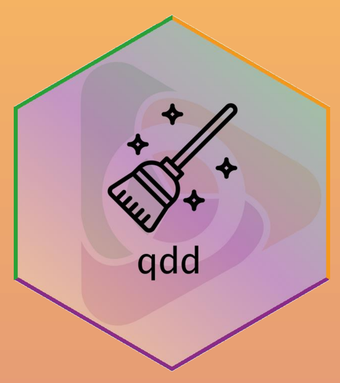
{qdd} : un **package R** de contrôle de la **qualité** et de **nettoyage des données** pour les Plateformes d'Epidémiosurveillance



Marine Marjou[13], Marie Grosdidier[13], Charlotte Rüger[2345], Pauline Bres[25]

Contact : marine.marjou@inrae.fr

Les Plateformes d'Epidémiosurveillance ?

Trois plateformes d'épidémiosurveillance existent dans les domaines de la **santé animale**, de la **santé végétale** et de la **chaîne alimentaire**. Les plateformes fonctionnent avec une **gouvernance partagée** entre public-privé. Elles visent à associer dans ses différents groupes de travail l'ensemble des partenaires impliqués dans la **surveillance des dangers sanitaires** concernés : État, organismes d'appui scientifique, laboratoires, représentants des agriculteurs, des professionnels des filières de production et de transformation, acteurs impliqués dans la faune sauvage et l'environnement. L'objectif principal des plateformes est d'améliorer l'efficience de la surveillance, chacune dans son champ de compétence.

Qualité des données : Les trois plateformes centralisent et travaillent sur des données issues de divers plans de surveillance nationaux en provenance de divers partenaires. Ces données sont multiples et variées et un travail de nettoyage en amont est nécessaire afin de pouvoir les analyser et les exploiter.

Objectif du package : Harmoniser et faciliter le travail de nettoyage des données au sein des trois plateformes en proposant un ensemble de fonctions regroupées dans le package {qdd}.

Jeu de données brut	Description
Données géographiques	Coordonnées géographiques, commune, département, région, pays
Données calendaires	Date de prélèvement, date d'analyse
Données quantitatives	Nombre de prélèvements…
Données qualitatives	Résultat d'analyse…

Référentiels :
- Géographiques
- Descripteurs de modalités
- Normes

Package qdd

Fonctions de qualité:
- **Complétude** : qualité d'un espace métrique sans valeur manquante
- **Validité** : concordance entre la donnée collectée et la valeur réelle
- **Cohérence** : organisation logique ou absence de contradiction entre plusieurs données liées entre elles
- **…**

➔ Calcul d'indicateurs

Fonctions de nettoyage :
- Homogénéité
- Interopérabilité
- Fiabilité

Rapport de qualité

Jeu de données propre

Jeu de données brut

ID	Longitude	Latitude	Ville
001	43.95431	4.804438	Paris ✗
002	5.278445	44.17397	Avignon ✓
003	6318912	844830.1	Avignon ✗
004	4.806610		Avignon ✗

Jeu de données propre

ID	Longitude	Latitude	Ville
001	4.804438	43.95431	Avignon ✓
002	5.278445	44.17397	Avignon ✓
003	4.804433	43.95431	Avignon ✓
004	NA	NA	Avignon ✗

Rapport de qualité du jdd brut

Indicateurs	Longitude	Latitude
Complétude : % de données manquantes	0 %	25 %
Format : % de données au format wgs84	75 %	50 %
Validité : % de données dans notre zone d'étude (PACA)	25 %	
Cohérence : % de données ayant une correspondance entre ville et coordonnées (longitude/latitude)	25 %	

Rapport de qualité jdd propre

Longitude	Latitude
25 %	25 %
75 %	75 %
75 %	
75 %	

1 INRAE, UR BioSP, Avignon
2 Anses Laboratoire de Lyon, Unité Epidémiologie et Appui à la Surveillance
3 Plateforme d'Epidemiosurveillance en Santé Végétale (ESV) https://plateforme-esv.fr
4 Plateforme d'Epidémiosurveillance en Santé Animale (ESA) https://plateforme-esa.fr
5 Plateforme de Surveillance de la Chaîne Alimentaire (SCA) https://plateforme-sca.fr

Guide pratique sur la qualité des données de surveillance