QUALITÉ

EXPLORATION ET DÉVELOPPEMENT DE MÉTHODES

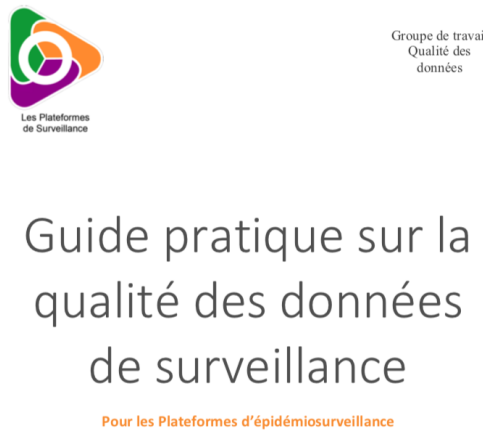
QUALITÉ DES DONNÉES

La qualité des données est essentielle pour la fiabilité des analyses et des résultats. Évaluer la qualité des données utilisées permet de prendre en compte les erreurs et les sources d'incertitudes liées aux données dans les analyses, et d'identifier les leviers d'amélioration de la qualité des productions de la Plateforme ESV. Une fois la qualité des données évaluée, leur nettoyage vise à corriger au moins une partie des données qui peuvent alors être intégrées aux analyses et ainsi apporter des informations supplémentaires sur les phénomènes étudiés.

GUIDE PRATIQUE INTERPLATEFORME

► Description

Les 3 plateformes d'épidémiosurveillance en santé animale (ESA), santé végétale (ESV), et chaîne alimentaire (SCA) sont en train de produire un guide pratique sur la qualité des données (prochainement disponible sur le site de la Plateforme ESV). Celui-ci sert de référence dans le développement des traitements relatifs à la qualité des données sur la Plateforme ESV.



► Exemple d'application

La Plateforme ESV centralise des données de surveillance sur le nématode du pin en France issues de multiples acteurs (FREDON, DSF, SRAL, ANSES ...). Ces données présentent des configurations variées (multiples bases de données ou fichiers excel) et une qualité fluctuante (données incomplètes, mal renseignées, erronées ...). Les données brutes récupérées par la Plateforme ESV suivent une démarche qualité présentée par le guide. Cette démarche permet d'homogénéiser les différentes façons de procéder entre les acteurs et également entre les 3 plateformes (ESV, SCA, ESA). Le schéma de qualité se

base sur 3 étapes : 1) la qualité des données est évaluée ; 2) les données sont nettoyées ; et 3) les données sont prêtes à être utilisées dans les analyses. Les données utilisées dans les analyses de la pression d'échantillonnage en France, passent par ces 3 étapes pour assurer la fiabilités des résultats.

► Notes et références

Palussiere, M. (2013). Thèse VETAGRO SUP : Évaluation de la qualité des données collectées dans le cadre d'un dispositif de surveillance en santé animale : proposition d'un guide élaboré à partir du dispositif de déclaration obligatoire des avortements bovins en France.

INDICATEURS POUR ÉVALUER LA QUALITÉ DES DONNÉES

► Description

Différents indicateurs permettent d'évaluer la qualité des données :

- complétude : la données est présente ;
- format : la données est dans le format attendu (par exemple, une date est au format JJ/MM/AAAA) ;
- validité : la donnée appartient à un domaine de valeurs plausibles (par exemple, les coordonnées géographiques d'un échantillon prélevé sur une plante terrestre doivent appartenir au domaine des terres émergées) ;
- cohérence : la donnée est cohérente avec ce qui est attendu (par exemple pour un plan de surveillance en 2018, on attend des dates de prélèvement en 2018).

L'évaluation de la qualité des données avec ces 4 indicateurs permet de détecter les marges de progression lors de la production de ces données. Ainsi, elle permet aux acteurs de la surveillance de faire évoluer les méthodes de production et de récupération des données.

► Exemple d'application

Les acteurs du plan de surveillance du nématode du pin sont évalués régulièrement sur la qualité des données issues de 4 indicateurs ; la complétude, le format, la validité et la cohérence.

Une visualisation graphique de cette analyse apporte à chacun des acteurs un repère temporel de l'amélioration de la production de leurs données et un focus sur leurs points faibles.

► Notes et références

Palussiere, M. (2013). Thèse VETAGRO SUP : Évaluation de la qualité des données collectées dans le cadre d'un dispositif de surveillance en santé animale : proposition d'un guide élaboré à partir du dispositif de déclaration obligatoire des avortements bovins en France.

Palussiere, M. et al. (2014). ANSES. Guide d'évaluation de la qualité des données d'un dispositif de surveillance épidémiologique en santé animale Version «bêta».

NETTOYAGE DES DONNÉES

► Description

Le nettoyage des données est l'action de détecter et corriger les données de mauvaise qualité et/ou comportant des erreurs. Soit la donnée est supprimée du jeu de données car elle est inutilisable et fausserait les analyses, soit la donnée peut être modifiée avec un niveau d'erreur acceptable pour permettre de garder une partie de l'information. Par exemple, si les coordonnées sont absentes mais que l'adresse exacte est notifiée, les coordonnées peuvent être reconstruites et les données complétées. Des données de meilleure qualité diminuent l'incertitude des résultats d'analyses liées à ces données et apporte une plus grande précision.

► Exemple d'application

Les données récupérées dans le cadre du plan de surveillance du nématode du pin en France, proviennent d'un jeu de données complété par de multiples acteurs qui ne travaillent pas avec les mêmes systèmes de géoréférencement. Les coordonnées du jeu de données sont parfois en Lambert 93, WGS 84 ou d'autres systèmes de géoréférencement. Le nettoyage consiste alors à mettre toutes les coordonnées dans le même système de géoréférencement pour obtenir un jeu de données homogène.

Voici un exemple de données brutes :

Coordonnée X	Coordonnée Y
2°02'35.4"E	44°31'51.6"N
-0,854453	44,172861
699709,38	2497132,62
514259,12	2412040,57

Voici un tableau résultant du processus de vérification de la qualité des données et de leur nettoyage :

Coordonnée X brute	Coordonnée Y brute	Projection	Coordonnée X WGS 84	Coordonnée Y WGS 84
2°02'35.4"E	44°31'51.6"N	degrés	2,0431667	44,5310000
-0,854453	44,172861	WGS 84	-0,854453	44,172861
699709,38	2497132,62	Lambert II étendu	3,710734	49,463680
514259,12	2412040,57	Lambert 93	1,172169	48,701192

► **Notes et références**

European Commission. (2020). e-learning Module 11 : Comment nettoyer les données.

N'dry N'Goran, A.O. (2016). Principes et méthodes de nettoyage de données. SUPINFO International University.