Traitement des données manquantes dans un projet participatif sur la biodiversité des sols agricoles.

Mario Cannavacciuolo*[1]

[1]UR LEVA (Légumineuses, Ecophysiologie Végétale, Agroécologie) – SFR 4207 QUASAV – École Supérieure d'Agricultures – Angers Loire – Univ Bretagne Loire – 55 rue Rabelais – BP 30748 – 49007 Angers Cedex 01, France, France

Résumé

Le sol, à travers sa composante biologique, joue un rôle essentiel dans le fonctionnement des agrosystèmes. Dans un contexte de développement durable, il apparaît indispensable de se doter d'outils permettant d'appréhender les impacts des systèmes de culture sur les organismes du sol. AgrInnov est un projet participatif français associant chercheurs et agriculteurs ; il vise à développer des outils opérationnels de caractérisation de la biodiversité dans les sols ainsi qu'à la construction de référentiels d'interprétation de cette biodiversité. Les bioindicateurs retenus ciblent trois grands groupes d'organismes du sol : les vers de terre, les nématodes et les communautés microbiennes. En parallèle, des indicateurs d'évaluation agronomique sont mis en place : analyses physicochimiques du sol, observation de la structure du sol, évaluation de l'activité de décomposition de la matière organique. Le projet s'appuie sur un réseau de 248 parcelles viticoles et de grandes cultures qui intègre une grande diversité de situations pédoclimatiques et culturales à l'échelle du territoire français. Les agriculteurs ont réalisé eux-mêmes l'échantillonnage de leur sol.
Les projets de sciences participatives présentent souvent des données erronées ou manquantes. Dans la cadre du projet AgrInnov, les données manquantes constituent un frein au diagnostic de l'impact des pratiques agricoles sur les caractéristiques biologiques et agronomiques du sol. Il est donc nécessaire de définir une stratégie de traitement des données manquantes (élimination de l'individu ou remplacement de la donnée manquante) afin de limiter leur impact dans l'analyse des données.

Les méthodes d'imputation de données manquantes sont préférées à la suppression d'individus ou de variables car toute l'information disponible est conservée [4]. Ainsi, les méthodes d'imputation multiple [2], les méthodes basées sur les forêts aléatoires et les méthodes d'imputation factorielles [1] induisent généralement de bonnes estimations des données manquantes.

Deux méthodes d'imputation factorielles (fonctions imputePCA et imputeMFA du package missMDA [3]), une méthode d'imputation multiple par " Predictive Mean Matching " (fonction mice du package mice [6]) et une méthode d'imputation par le plus proche voisin (fonction missForest du package missForest [5]) sont comparées, par le biais de simulations numériques sur un jeu de données sans données manquantes, afin de sélectionner celle qui induit les meilleures estimations.
Lorsque les données manquantes sont de type MAR (Missing At Random), une imputation

*Intervenant

multiple par PMM semble la plus pertinente. En effet, en prenant en compte le pourcentage de données manquantes par variables, cette méthode d'imputation minimise le NRMSE (Normalized Root Mean Square Error) et le critère de la moyenne des écarts à la matrice de corrélation par rapport aux trois autres méthodes testées (imputations factorielles et imputation par le plus proche voisin). Cette méthode reconstitue mieux les données et déforme moins les relations linéaires entre les variables.