Validation de données en phénotypage végétal

A. Schorgen[a] and **N. Hilgert**[a]

[a] UMR 729 MISTEA, INRA
2, place Pierre Viala
34060 Montpellier Cedex 2
antoine.schorgen@supagro.inra.fr
nadine.hilgert@supagro.inra.fr

Mots clefs : Statistique, Visualisation, Régression, Agriculture, Phénotypage.

Dans un contexte de changement climatique et de respect de l'environnement, l'agriculture porte une partie de ses efforts de recherche dans le phénotypage, notamment pour aider à la sélection de variétés végétales moins consommatrices d'intrants. Le projet ANR-PHENOME, dans lequel notre travail s'inscrit, a permis un investissement dans des plateformes de phénotypage à haut débit dans lesquelles sont étudiés simultanément un grand nombre de génotypes, soumis à des scénarios de stress environnementaux (manque d'eau, manque d'azote...). Une très grande quantité de données sortent automatiquement de ces plateformes, et nous nous intéressons particulièrement aux photographies des plantes prises au cours du temps.

A partir de ces images, l'enjeu est de reconstituer des informations fiables sur des variables biologiques telles que la biomasse ou la surface foliaire des plantes. Cette étape nécessite forcément une procédure de validation des données sur laquelle nous nous concentrerons dans notre présentation. En effet, toute expérience automatique présente des failles (absence de pot dans la cabine photo, poussière qui perturbe le système ou capteur défaillant...) se traduisant par des données manquantes ou aberrantes; malheureusement le grand nombre de données récoltées empêche une correction manuelle au cas par cas. L'objectif est donc de fournir aux utilisateurs de la plateforme des méthodes statistiques et informatiques de nettoyage automatique des données.

Dans un premier temps, des routines R sont mises en place pour organiser les données (packages *data.table* et *dplyr*) et pour visualiser facilement les résultats (package *ggplot2*). Ensuite, nous utilisons des méthodes statistiques de type régression paramétrique non linéaire (avec fonction logistique ou gompertz) ou régression non-paramétrique (lissage polynomial, splines...) pour détecter les valeurs aberrantes et ainsi annoter les données.