

Sujets de projet M2 science des données

Rémi MAHMOUD, Anne Laperche

October 2025

Le colza est l'une des principales cultures oléagineuses en Europe. Il constitue une source majeure d'huile végétale pour l'alimentation humaine, la production de biocarburants et l'alimentation animale sous forme de tourteaux riches en protéines. Compte tenu de ces usages multiples et de l'importance économique du colza, l'amélioration de ses performances agronomiques (rendement, vigueur en début de cycle) est un enjeu majeur pour la compétitivité et la durabilité de la filière.

Je propose ici trois sujets en lien avec le colza pour le projet de M2 sciences des données, en collaboration avec Anne Laperche (UMR IGEPP, INRAE / Institut Agro). J'encadrerai ces projets et Anne Laperche en suivra l'évolution.
Je précise que je ne pourrai encadrer que deux sujets sur les trois proposés.

En statistique, il est toujours bon de comprendre comment les données ont été générées. Dans cette optique, un déplacement d'une demi-journée à l'INRAE du Rheu, pour participer et/ou observer des prélèvements, pourra être prévu durant le projet.

Projet 1: Apport des données fonctionnelles en agronomie: le cas des données météo et du rendement

Actuellement, la plupart des études de prédition du rendement à partir de la météo en agronomie s'appuient sur des résumés de variables météorologiques (température, humidité du sol, indice foliaire, etc.) calculés sur des intervalles définis entre stades phénologiques majeurs (ex. de la levée à la floraison, ou de la floraison à la maturité), ou par périodes fixes (semaines, mois, etc.). Ces valeurs agrégées servent ensuite de prédicteurs dans des modèles multivariés comme la régression PLS (Partial Least Squares).

Cependant, cette approche par résumés risque de perdre des informations dynamiques contenues dans la trajectoire complète des variables au cours du cycle de culture. L'utilisation de méthodes de statistique fonctionnelles, qui traitent les variables comme des fonctions continues dans le temps, pourrait

permettre de capturer plus finement les variations temporelles et leur lien avec le rendement.

Toutefois, un défi important de ce type d'approche sera de réaligner les séries temporelles entre essais, en particulier en fonction des stades phénologiques, car les vitesses de développement des cultures varient selon les conditions environnementales ou les génotypes. Réaligner les données sur une échelle temporelle relative aux stades phénologiques pourrait ainsi améliorer la comparabilité des dynamiques fonctionnelles et, potentiellement, la capacité prédictive des modèles.

Nous avons à notre disposition un jeu de données contenant des séries temporelles de différentes variables météorologiques dans 22 essais incluant du colza d'hiver et les rendements liés.

Les enjeux de ce projet sont multiples:

1. Familiarisation avec les données et les méthodes de statistique fonctionnelle
2. Mise au point / recherche d'une (ou plusieurs) méthodes d'alignement des stades phénologiques (ex. landmark registration)
3. Application de ces méthodes aux séries temporelles
4. Mise en place d'algorithme de prédictions du rendement
5. Comparaison des résultats avec les méthodes plus classiques

Projet 2: Utilisation de données spectrales pour la prédiction de traits d'intérêt du colza

Ce projet s'inscrit dans la continuité des recherches menées par Marianne Laurençon au cours de sa thèse. L'un des objectifs de ces travaux était de prédire des traits phénotypiques associés à la vigueur précoce du colza, à partir de spectres NIRS (Near Infrared Spectroscopy) enregistrés sur les graines ou les feuilles. Marianne Laurençon a ainsi mis en évidence la possibilité de prédire certains de ces traits avec un degré de précision intéressant. En phénomique, la prédiction de tels caractères repose généralement sur des modèles exploitant les spectres comme des données multivariées (PLS, LASSO etc.), chaque longueur d'onde étant considérée comme une variable indépendante.

Or, un spectre NIRS représente en réalité une courbe d'absorbance mesurée sur plusieurs centaines de longueurs d'onde, ce qui en fait une donnée intrinsèquement continue et fonctionnelle. Tirer parti de cette nature fonctionnelle pourrait permettre d'améliorer la qualité des prédictions et de mieux capturer les signaux complexes portés par les spectres.

L'enjeu de ce projet est donc double : il s'agit de se familiariser d'une part avec les données de spectroscopie NIRS (issues de graines et de feuilles), et d'autre part avec les différentes méthodes de statistique fonctionnelle (ACP

fonctionnelle, modèles de régression fonctionnelle, techniques de machine learning pour données fonctionnelles, etc.). Ces approches seront ensuite appliquées aux jeux de données disponibles, afin de comparer leurs performances prédictives et leur capacité d'interprétation à celles des méthodes plus classiques utilisées en génétique.

Projet 3: Combinaison de données NIRS et génomique pour la prédition de traits