

# Modélisation de la variabilité génotypique des paramètres d'un modèle de culture chez *Arabidopsis thaliana*

Stage niveau M2 début 2021

Unité Maïage INRAE Maïage Jouy-en-Josas

Pour postuler, merci d'envoyer un CV, une lettre de motivation et vos derniers relevés de notes à [estelle.kuhn@inrae.fr](mailto:estelle.kuhn@inrae.fr).

## Contexte

Dans le contexte actuel, l'agriculture doit proposer de nouvelles innovations pour faire face aux effets du changement climatique. Pour mieux appréhender ces effets, il est primordial de mieux comprendre les interactions existantes entre la plante et son environnement, en particulier les interactions entre différentes variétés et différentes conditions environnementales. Il est ainsi crucial de développer de nouvelles méthodes permettant d'identifier les processus biologiques clés qui sont à l'origine de la variabilité observée chez les plantes. De telles approches permettront de mieux comprendre les interactions entre les différentes variétés (génotypes) et différents environnements et d'identifier des leviers d'action biologiques afin de sélectionner les variétés les mieux adaptées.

S'agissant de modéliser des observations répétées sur plusieurs individus, les modèles à effets mixtes sont largement utilisés dans de nombreux domaines tels que l'agronomie, la pharmacologie, l'écologie ou encore l'épidémiologie. Ces modèles permettent de modéliser différents niveaux de variabilité en prenant en compte à la fois la variabilité existante entre deux individus et la variabilité existante entre plusieurs mesures faites sur un même individu. On parle de variabilités inter-individuelle et intra-individuelle. Les différents niveaux de variabilité sont modélisés par deux types d'effets : d'une part les effets fixes, communs à tous les individus de la population, d'autre part les effets aléatoires qui varient d'un individu à l'autre. Les effets fixes sont des paramètres inconnus du modèle qu'on cherchera à estimer. Les effets aléatoires sont des variables aléatoires latentes du modèle qui ne sont pas observées. Les objectifs sont d'estimer les paramètres de la loi de probabilité des effets aléatoires et également de prédire ces effets aléatoires.

Modéliser au plus juste les différents types d'effets est primordial, à la fois pour l'inférence des paramètres et pour pouvoir utiliser le modèle à des fins prédictives. Ainsi, être en mesure de distinguer parmi tous les effets du modèle ceux qui peuvent être modélisés par des effets fixes permet de réduire le nombre de paramètres du modèle, et peut également aider à mieux identifier les processus qui seraient à l'origine de la variabilité observée dans la population. Dans notre contexte, l'objectif est d'identifier les processus biologiques de développement de la plante fortement dépendant de son génotype. Pour distinguer parmi les effets aléatoires ceux qui pourraient être modélisés par des effets fixes au regard des données considérées, on peut considérer la matrice de covariance des effets aléatoires. Plus précisément, un effet aléatoire de variance nulle ne présente aucune variabilité et pourrait être modélisé comme un effet fixe. Du point de vue statistique, cela revient à tester si la matrice de covariance associée à ces effets aléatoires est nulle. La difficulté réside dans le fait que les valeurs des paramètres à tester sont sur la frontière de l'espace des paramètres. Des travaux sur ce sujet ont déjà été menés, en particulier la loi asymptotique de la statistique du test de rapport de vraisemblance a été identifiée [1]. Cependant, dans de nombreux cas pratiques, particulièrement en sciences des plantes, le nombre d'individus observés est petit, rendant la mise en œuvre des résultats asymptotiques discutables.

L'objectif du stage est de proposer des méthodes adaptées à des échantillons de petite taille et de les mettre en oeuvre sur des données réelles en collaboration avec Céline Richard-Molard (INRAE, EcoSys). On s'intéressera notamment au modèle de croissance de plante *Arnica*, qui décrit les échanges de flux de carbone et d'azote chez *Arabidopsis thaliana*. Ce modèle comporte une dizaine de paramètres, tous reliés à des processus biologiques précis. L'objectif sera d'identifier les paramètres et par extension les processus biologiques impliqués dans la variabilité phénotypique observée. Un jeu de données portant sur 48 écotypes d'*Arabidopsis thaliana*, représentant 100 % de la diversité allélique de l'espèce, permettra de tester sur données réelles les approches développées au cours du stage.

## Objectifs du stage

- \* proposer des procédures de test alternatives non asymptotiques.
- \* implémenter les méthodes sur des données simulées pour comparer les performances avec le test asymptotique en utilisant un modèle de croissance simple, par exemple logistique.
- \* implémenter la méthode sur les données réelles de *Arabidopsis thaliana* en utilisant le modèle de culture *Arnica*.

**Aspects mathématiques** Deux approches sont envisageables pour construire une procédure de test non asymptotique, d'une part une approche basée sur une version bootstrap du test de rapport de vraisemblance, d'autre part une approche basée sur la construction d'une statistique de test ajustée, incluant un terme correctif pour les petits échantillons en s'inspirant des travaux de [2].

## Profil recherché

Formation niveau BAC+5 (Master 2 ou école d'ingénieurs), connaissance en statistiques théoriques et appliquées, ayant un fort intérêt pour les applications en sciences du vivant ; maîtrise d'un langage de programmation indispensable ; rigueur scientifique, curiosité intellectuelle, facilité de communication.

## Modalités pratiques

Le stage se déroulera à Jouy-en-Josas. La durée du stage sera de cinq ou six mois, entre février et septembre 2021. La gratification mensuelle est d'environ 550 euro (taux légal). L'encadrement sera réalisé par Estelle Kuhn (INRAE, MaIAGE) et Charlotte Baey (Université de Lille, laboratoire Painlevé). Le stage pourra possiblement déboucher sur un sujet de thèse.

## Références bibliographiques

- [1] C. Baey, P.H. Cournède, E. Kuhn. Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, Vol. 135, pp 107–122, 2019.
- [2] A. C. Monti, M. Taniguchi. Adjustments for a class of tests under nonstandard conditions. *Statistica Sinica*, Vol. 28, pp 1437–1458, 2018.