

# Stage M2 : modélisation d'un espace latent via des méthodes type *embeddings* dans un modèle à effets mixtes.

Jean-Benoist Leger – Estelle Kuhn

2026

Pour candidater, merci d'envoyer un CV, une lettre de motivation et vos relevés de notes de M1 et de M2 (si disponibles) à Jean-Benoist Leger ([jbleger@hds.utc.fr](mailto:jbleger@hds.utc.fr)) et Estelle Kuhn ([estelle.kuhn@inrae.fr](mailto:estelle.kuhn@inrae.fr)).

## Motivation biologique

On considère le cas pratique d'une population de variétés de la plante modèle *Arabidopsis Thaliana* pour lesquelles on mesure au cours du temps une caractéristique du processus de vieillissement des feuilles (sénescence) en comptant le nombre de feuilles fanées au cours du temps. Ce processus physiologique est complexe et résulte d'un compromis entre deux processus dynamiques, d'une part le maintien de la photosynthèse, d'autre part la remobilisation de l'azote des feuilles vers les graines. Du point de vue biologique, il existe plusieurs stratégies pour réaliser ce compromis, conduisant à une variabilité complexe et importante entre les différentes variétés. L'objectif appliqué est de caractériser cette variabilité à partir des observations du processus de sénescence.

## Objectifs du stage

On souhaite caractériser la variabilité existante au sein d'une population d'individus à partir de mesures répétées : un caractère d'intérêt est mesuré pour chaque individu plusieurs fois dans des conditions différentes. Lorsque ce caractère est mesuré plusieurs fois au cours du temps pour chaque individu, il s'agit de données longitudinales. Les modèles à effets mixtes sont des modèles à variables latentes bien adaptés pour ce type de données car ils permettent de modéliser à la fois la variabilité présente au sein des mesures répétées d'un individu et la variabilité entre les individus de la population. Du point de vue de la modélisation statistique, ces modèles comportent des paramètres communs à l'ensemble des individus de la population, appelés effets fixes, et des paramètres individuels spécifiques à chaque individu, appelés effets aléatoires. Ces derniers sont modélisés par des variables aléatoires non observées. Ainsi caractériser la variabilité existante entre les individus de la population équivaut à caractériser la distribution des paramètres individuels.

Pour modéliser cette distribution, une approche paramétrique peut être adoptée en choisissant une famille paramétrique de lois bien adaptée au contexte. Caractériser la loi revient alors à estimer les paramètres. En l'absence d'information sur la forme de la distribution ou dans le cas d'une distribution très complexe, une approche non paramétrique peut permettre une modélisation des effets aléatoires moins contraintes. Différentes approches ont été proposées, comme par exemple dans le cas de modèles linéaires (Comte and Samson, 2012). Plusieurs approches sont comparées par Antic et al. (2009).

L'objectif du stage est de développer une nouvelle approche non paramétrique en capitalisant sur les méthodes récentes de type *embeddings* issues de l'apprentissage profond. On pourra par exemple considérer des *normalizing flows* (Kobyzev et al., 2020). Il s'agit d'une classe de modèles génératifs probabilistes flexibles qui transforment une distribution simple en une distribution complexe via une série de transformations simples, différentiables et réversibles. Les modèles *normalizing flows* permettent de représenter efficacement un large panel de distributions complexes. Pour atteindre cet objectif, ils mobilisent souvent des réseaux de neurones constitués de transformations élémentaires composables.

Ce stage vise à développer et évaluer une méthodologie combinant ces outils pour mieux capturer des caractéristiques telles que la multimodalité et les asymétries. Après une phase de conception et d'implémentation adaptée aux spécificités des modèles non linéaires, une évaluation sera menée sur des données synthétiques et réelles d'*Arabidopsis Thaliana*. L'approche sera comparée aux modèles traditionnels pour quantifier ses avantages dans la représentation des distributions latentes.

## Profil recherché

Formation niveau BAC+5 (Master 2 ou école d'ingénieurs), connaissance en statistiques théoriques et computationnelles, maîtrise d'un langage de programmation indispensable ; rigueur scientifique, curiosité intellectuelle, facilité de communication.

## Modalités pratiques

Le stage s'inscrit dans le cadre du projet ANR AgriCoolTools. Il se déroulera au centre INRAE de Jouy-en-Josas dans l'unité MaIAGE. La durée du stage sera de cinq ou six mois, entre février et septembre 2026. La gratification mensuelle est d'environ 650 euro (taux légal). L'encadrement sera réalisé par Estelle Kuhn (INRAE, MaIAGE) et Jean-Benoist Leger (UTC, Heudiasyc). L'analyse des données réelles *Arabidopsis thaliana* se fera en étroite collaboration avec Fabien Chardon (INRAE, IJPB). Le stage pourra possiblement déboucher sur un sujet de thèse combinant des statistiques mathématiques, computationnelles et des applications en sciences du vivant.

## Références

- Antic, J., Laffont, C. M., Chafai, D., and Concordet, D. (2009). Comparison of nonparametric methods in nonlinear mixed effects models. *Computational statistics & data analysis*, 53(3):642–656.
- Comte, F. and Samson, A. (2012). Nonparametric estimation of random-effects densities in linear mixed-effects model. *Journal of Nonparametric Statistics*, 24(4):951–975.
- Kobyzev, I., Prince, S., and Brubaker, M. A. (2020). Normalizing flows : An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:3964–3979.