# Modélisation jointe de données longitudinales et de survie en grande dimension. Application à la prédiction des effets des attaques de pyrale sur la floraison du maïs

**Stage niveau M2 début 2022**

**Unité MaIAGE Mathématiques et informatique appliquées du génome à l'environnement, INRAE Jouy-en-Josas**

Pour postuler, merci d'envoyer un dossier complet contenant **votre CV, votre lettre de motivation et vos derniers relevés de notes de Master 1 ou équivalent** à estelle.kuhn@inrae.fr et à sarah.lemler@centralesupelec.fr.

## Contexte

Dans le contexte actuel de changement climatique, l'agriculture est au cœur des préoccupations, à la fois comme l'une des causes de ce processus, mais aussi du fait des bouleversements majeurs qu'elle subira et auxquels elle devra s'adapter. L'une des notions clés qu'il est nécessaire de mieux comprendre pour appréhender ces questions est celle de l'interaction entre la plante et son environnement au sens large (conditions météorologiques, conditions de sol, présence de ravageurs, conduite de culture, ...). Comment adapter les variétés cultivées aux effets du changement climatique, en particulier aux situations de stress? Peut-on identifier des leviers d'action biologiques afin de sélectionner les variétés adaptées aux nouvelles conditions?

La pyrale du maïs (*Ostrinia nubilalis* est un des ravageurs majeurs du maïs en Europe. Elle peut entraîner des baisses considérables de rendement et favoriser le développement de pathogènes sur les plantes attaquées. Le présent projet vise à développer un modèle prédictif de la date de floraison du maïs intégrant simultanément les variabilités de l'environnement et génétiques, ainsi que la dynamique des attaques de pyrales.

Les données disponibles proviennent d'une expérience de sélection divergente menée depuis plus de 20 ans par l'unité GQE-Le Moulon (Université Paris-Saclay/INRAE/ CNRS/AgroParisTech). A partir de deux lignées ancestrales, des lignées contrastées pour la date de floraison ont été dérivées. Pour chaque fond génétique, chaque année, 4 à 6 représentants (lignées) de deux familles à floraison précoce et deux familles à floraison tardive sont semés puis sélectionnés selon le caractère précoce ou tardif de la date de floraison. L'année suivante, les graines produites par ces plantes sont semées et la sélection est de nouveau opérée. Après 20 ans de sélection divergente, plusieurs semaines séparent les populations précoces et tardives en terme de date de floraison. Pour les années 2018 et 2019, nous disposons pour les plantes de chaque génotype d'une estimation de la date moyenne de floraison et d'observations de l'évolution de la présence de symptômes d'attaques de pyrale au cours du temps. Des données climatiques sont également disponibles et des données de génotypage ont par ailleurs été acquises sur les lignées parentales des plantes observées en 2018 et en 2019.

## Objectifs du stage

* proposer un modèle statistique joint pour analyser les données de survie (date de floraison) et les données longitudinales (dynamique de la proportion de plantes attaquées par la pyrale), en intégrant des covariables de grande dimension (marqueurs génétiques),

* proposer un estimateur pénalisé des paramètres du modèle

* développer et implémenter un algorithme d'estimation et le valider sur données simulées,
* proposer un prédicteur pour la date de floraison et le valider sur données simulées,
* ajuster le modèle proposé aux données réelles,
* valider les prédictions obtenues sur les données réelles.

**Aspects mathématiques** L'approche envisagée pour la modélisation reposera d'une part sur les modèles de survie à effets aléatoires pour la date de floraison incluant un effet du génotype, des variables explicatives environnementales et génétiques et une variable décrivant la dynamique du phénotype intermédiaire, la proportion (prévalence) de plantes attaquées par la pyrale, d'autre part sur les modèles non linéaires (e.g croissance logistique) à effets mixtes pour la prévalence des plantes attaquées par la pyrale. L'estimation des paramètres pourra se faire via des procédures adaptées à la grande dimension telles que des critères de vraisemblance pénalisées de type LASSO ou des approches bayésiennes de type spike and slab. Des algorithmes stochastiques seront mis en œuvre pour l'inférence. Un prédicteur de la date de floraison sera construit à partir du modèle ajusté ou de méthodes d'apprentissage statistique telles que les forêts aléatoires de survie étendues aux modèles joints. Une étude comparative sur la performance relative des prédictions de date de floraison pourra être menée incluant d'autres méthodes de la littérature.

# 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 s'inscrit dans le cadre du projet ANR Stat4Plant. Il se déroulera principalement au centre INRAE de Jouy-en-Josas dans l'unité MaIAGE et de façon ponctuelle dans l'unité INRAE GQE Le Moulon située à Gif-sur-Yvette. La durée du stage sera de cinq ou six mois, entre février et septembre 2022. La gratification mensuelle est d'environ 550 euro (taux légal). L'encadrement sera réalisé par Estelle Kuhn (INRAE, MaIAGE), Judith Legrand, Marchadier Elodie (INRAE, GQE Le Moulon) et Sarah Lemler (CentraleSupélec, MICS). Le stage pourra possiblement déboucher sur un sujet de thèse combinant de la modélisation mathématique, des statistiques et de l'analyse de données expérimentales.

# Références bibliographiques

[1] Oodally A., Kuhn E., Goethals K., Duchateau L., Modeling dependent survival data through random effects with spatial correlation at the subject level. arXiv, (2020).
[2] Kuhn E., Matias C., Rebafka T., Properties of the Stochastic Approximation EM Algorithm with Mini-batch Sampling. Statistics and Computing, Vol. 30, pp 1725–1739, (2020).
[3] Guilloux A. et al., Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates. Journal of Multivariate Analysis, Vol. 148, pp 141–159, (2016).
[4] Durand E., Tenaillon M., Ridel C., Coubriche D., Jamin P., Jouanne S., et al. Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds. BMC Evol Biol. 10:2(2010).