Introduction

La compréhension de la vie peut être grandement améliorée par l'utilisation d'outils mathématiques, informatiques et statistiques. De plus avec l'avénement de méthodes plus puissantes, la biologie génére un grand volume de données qui doivent être correctement analysée afin de parvenir à une meilleur compréhension de notre environnement.

Cette formation a pour but d'outiller les élèves ingénieurs en sciences agronomiques ou biologiques en fin de cycle avec les outils de statistique appliquée couramment utilisés en biologie. Le focus est volontairement mis sur la modélisation des données, les autres aspects de l'analyse étant abordés dans d'autres formations pour une question de temps mais aussi et surtout de pédagogie pour éviter d'aborder un trop grand nombre de concepts en une seule formation.

Ce que vous allez apprendre

Au cours de cette formation nous aborderons les concepts suivants:

- Exploration des données: Ce module met l'accent sur les valeurs aberrantes et les transformations de données. En effet, la majorité des méthodes statistiques ont des conditions d'application qu'il faut toujours vérifier avant de les utiliser au risque de se tromper soi-même.
- Regression: Nous mettrons l'accent dans ce module sur une compréhension profonde des conditions d'application de la regréssion linéaire et des outils d'évaluation de ces conditions. La majorité des étudiants ont l'habitude d'affirmer comprendre la regression linéaire mais échoue à identifier ses conditions d'application.
- Modélisation linéaire généralisée (GLM) et modélisation additive généralisée (GAM): La majorité des données biologiques ne respectent pas les conditions d'application de la regression. Des outils plus appropriés tels que les modèles linéaires généralisées et les modèles additifs peuvent alors être appropriés.
- Modélisation mixte et les moindres carrés généralisés (GLS): Les modèles mixtes sont également connus sous le nom de modèles à effets mixtes ou de modèles multiniveaux et sont utilisés lorsque les données ont une forme hiérarchique, comme dans les données longitudinales ou de panel, les mesures répétées, les séries temporelles et les expériences bloquées, qui peuvent avoir des coefficients fixes et aléatoires ainsi que des termes d'erreur multiples.

Ce livre est un manuel de formation à destination des ingénieurs en sciences agronomiques ou biologiques en fin de cycle (3e année).

Il se veut pratique et facile à aborder tout en n'hésitant pas à revenir sur les concepts mathématiques avancés pour solidifier les fondements mathématiques des participants.

Ce manuel est utilisé lors de la formation organisée par le biais du service innovation et transfert de technologie des fablabs agritech de l'Institut National Polytechnique Félix Houphouët-Boigny de Côte d'Ivoire.

Ce que vous n'allez pas apprendre

Cette formation n'est pas une formation à l'utilisation de R. Nous n'aborderons donc pas les notions basiques de son utilisation. Nous ferons cependant l'effort d'apporter de l'aide ou des informations pour chaque fonction utilisée. Nous n'aborderont pas aussi les thèmes des tests d'hypothèses, de la conception d'expérience, de la manipulation des données, de la visualisation des données ou de la programmation littérale avec R.

Ces thèmes sont abordés dans d'autres formations que nous organisons.

Nous avons choisi de faire cette formation en utilisant les fonction R basiques et non sous l'aspect de la philosophie tidyverse (pour laquelle nous consacrons une autre formation). Ce choix est fait pour garder l'attention des étudiants dirigées exclusivement sur la compréhension des concepts statistiques présentés et à leur mise en pratique avec R.

Pré-requis

Nous avons fait quelques suppositions sur ce que vous savez déjà pour tirer le meilleur parti de cette formation. Vous devez avoir des connaissances générales en calcul et il est utile que vous ayez déjà une certaine expérience de la programmation de base. Si vous n'avez jamais programmé auparavant, Hands on Programming with R pourrait être un outil précieux.

Vous avez besoin de quatre éléments pour exécuter le code de ce livre : R, RStudio, un certains nombre de jeu de données. Les packages sont les unités fondamentales du code R reproductible. Ils comprennent des fonctions réutilisables, de la documentation décrivant comment les utiliser et des exemples de données.

R

Pour télécharger R, rendez-vous sur CRAN, le réseau complet d'archives R, https://cloud.r-project.org. Une nouvelle version majeure de R est publiée une fois par an, et il y a 2 à 3 versions mineures par an. C'est une bonne idée de faire des mises à jour régulièrement. La mise à jour peut être un peu fastidieuse, en particulier pour les versions majeures qui vous obligent à réinstaller tous vos paquets, mais la remettre à plus tard ne fait qu'empirer les choses. Nous recommandons R 4.2.0 ou une version ultérieure pour cette formation.

RStudio

RStudio est un environnement de développement intégré (IDE) pour la programmation R, que vous pouvez télécharger à partir de https://posit.co/download/rstudio-desktop/. RStudio est mis à jour plusieurs fois par an et vous informe automatiquement de la sortie d'une nouvelle version. C'est une bonne idée de faire des mises à jour régulières pour profiter des dernières et meilleures fonctionnalités. Pour ce livre, assurez-vous d'avoir au moins RStudio 2022.02.0.

Lorsque vous démarrez RStudio , vous voyez deux zones clés de l'interface : le volet de console et le volet de sortie. Pour une exécution du code ligne par ligne, il faut taper le code R dans le volet de la console et appuyer sur la touche Entrée pour l'exécuter. Cependant si l'on veut créer un fichier pour y saisir le code il est possible d'utiliser l'éditeur de texte incorporé.