## Journées d’analyse statistique des données sur R

## Edition 2025 délocalisée à l’OHP (St Michel l’O.)

## 31 mars au 4 avril

Comité d’organisation : F Torre & J Cavalieri

Equipe pédagogique : Alberte Bondeau (CNRS-IMBE), Claire Della Vedova ([www.delladata.fr](http://www.delladata.fr)), Maxime Logez (INRAE – RiverLy), Laura March (IRD – LPED), Mathieu Santonja (AMU-IMBE) & Franck Torre (AMU-IMBE).

## Contexte

La fédération ECCOREV (2009-2013) soutient depuis de nombreuses années cet événement visant l’initiation à la pratique de la statistique des données environnementales sous logiciel R (syntaxe de base et notions plus avancées , analyse multivariée, statistiques spatiales).

Les plus récentes éditions – sauf en 2021 - ont été délocalisées au centre Séolane à Barcelonnette, er depuis 2024 à l’OHP (St Michel l’O.).

**La délocalisation permet une plus grande disponibilité des formateurs et des participants, pour de la consultation statistique sur une plage horaire étendue (sessions nocturnes offrant l’occasion de travaux sur les données des participants). Le coût additionnel engendré par cette délocalisation est donc amplement justifié.**

## Modalités de cours

Ces journées se dérouleront sur ordinateur, avec le logiciel R, qui est un logiciel libre d’usage très répandu dans la communauté scientifique « Eccorev ». Le conférencier présentera les techniques en les illustrant avec des exemples reproductibles par les étudiants immédiatement sur leur ordinateur. Il y aura donc à la fois vidéo-projection et TP en parallèle.

L’inscription se fera pour l’ensemble des sessions. La salle peut accueillir 18 participants en présentiel. Chaque participant amène son ordinateur portable personnel équipé du wifi. La configuration logicielle sera précisée avant le début des journées et nous anticiperons l’installation des packages pour chaque session. Le public visé (niveau débutant) sera en priorité : les thésards, post-docs et jeunes chercheurs des UMRs membres d’ECCOREV, des Instituts ITEM & OCEAN, en particulier ceux inscrits à l’ED251 selon la place disponible. L’objectif est de rendre l’utilisateur suffisamment autonome dans les domaines les plus « populaires » de la statistique, de manière qu’il soit capable de pratiquer directement les cas standards et d’être capable de personnaliser son approche.

## Nombre de participants : 16

## Considérations pratiques

**La prise en charge ne comprend pas le transport depuis Aix Marseille qui reste à la charge des unités de recherche**. L’hébergement en chambres collectives et la restauration sur place sont pris en charge du lundi midi au vendredi midi (chambres doubles, panier repas le midi et plateau repas le soir). Des navettes vers le village de St Michel l’O. pourront être organisées. Une après-midi récréative ainsi qu’un repas. Un engagement à participer à l’ensemble des sessions conditionne l’éligibilité au plan de formation doctorale. Un protocole sanitaire sera établi pour la durée de l’événement.

## Programme Edition 2025

J1

J1 – matin

*Transport vers OHP*

J1 – 13h30-16h30

*Introduction à R*

Par Franck Torre, AMU-IMBE & Maxime Logez, INRAE-RiverLy

Le but de cette session est de permettre à l’utilisateur novice de naviguer au sein l’environnement et d’utiliser les outils offerts par R pour l’analyse de données. Quelques applications graphiques et statistiques simples seront travaillées.

J2

J2 – 9h30-12h30

*Analyse de données environnementales multivariées*

Par Franck Torre, AMU-IMBE

L’analyse de données permet de mettre en évidence l’information contenu dans un tableau de données multivariables. En fonction de la nature de ces variables, différentes méthodes ont été proposées et leur présentation est au programme de cette séance : analyse en composantes principales normées ou centrées, analyse factorielle des correspondances, analyse des correspondances multiples, analyses de données mixtes. Les méthodes de couplage de tableaux (coinertie, RDA/CCA) permettent d'étudier le lien entre deux tableaux. Ces dernières permettent de décomposer la variance d’un tableau à expliquer selon différents compartiments de variables explicatifs. Des exemples provenant d’échantillonnage en écologie serviront d’illustration : tableaux biologiques, mésologiques, météorologiques, intentions expérimentales …. Des exemples provenant d’échantillonnage en écologie serviront d’illustration.

J2 – 13h30-16h30

*Analyses de données par équations structurelles*

Par Mathieu Santonja, AMU-IMBE

Après un cours permettant d’acquérir les bases conceptuelles et pratiques pour l’utilisation des équations structurelles, la demi-journée/journée sera à consacrée à l’analyses de deux jeux de données portant i) sur l’impact du changement climatique sur le réseau trophique du sol en prairie alpine et ii) sur l’impact de la gestion forestières sur la biodiversité et les services écosystémiques rendus par les forêts méditerranéennes.

J2 – 17h30-19h30

*Présentation de travaux & problématiques quantitatives*

Par les participants volontaires

*Un temps – en soirée - pourra être réservé à la consultation statistique.*

J3

J3 – 9h30-12h30

*Notions avancées sur R*

Par Maxime Logez, INRAE-RiverLy

Ce module a pour but de familiariser les utilisateurs avec la programmation en R, avec d’une part l’usage et la création de fonctions, l’utilisation d’outils de programmations classiques et très utilisés que sont les boucles et leurs pendants (fonctions de la famille des apply) ainsi que les différents éléments de langages indispensables. Nous montrerons les possibilités du logiciel en termes de lecture de données (lecture conditionnelle de tableau, …) ainsi que sur l’utilisation de représentations graphiques interactives.

J3 – 9h30-12h30

*Après-midi récréative OHP*

Lâcher ballon sonde (selon météo)

Visite guidée du T193

…

J3 – 17h30-19h30

*Graphiques ggplot*

Par Maxime Logez, INRAE-RiverLy

La librairie ggplot2 offre de très nombreuses possibilités de représentation graphiques simples (nuages de points, histogramme, courbe de densité, …) et complexes (multi-panneau). Elle s’intègre pleinement dans l’univers « tydiverse ». Le but de cette session sera d’utiliser les fonctions de mise en forme des tableaux des librairies *dplyr* et *tidyr* pour ensuite réaliser des représentations graphiques avec ggplot2 et les customiser.

*Un temps – en soirée - pourra être réservé à la consultation statistique.*

J4

J4 – 9h30-12h30

*Analyse spatiale 1/2*

Par Alberte Bondeau, CNRS-IMBE

Après une présentation rapide des types de questions et de données auxquelles s'appliquent les méthodes des statistiques spatiales (processus ponctuels, analyses sur réseaux et sur grille, géostatistique), la demi-journée sera consacrée à une introduction des concepts et méthodes de la Géostatistique au travers d’exemples et de petits programmes sous R.

J4 – 13h30-16h30

*Outils de reporting (Quarto)*

Par Claire Della Vedova (société DellaData)

Formation sur les thèmes du reporting avec l'outil Quarto sous l'environnement RStudio et sur l'utilisation des outils d'intelligence

artificielle avec R.

J4 – 17h30-19h30

*Présentation de travaux & problématiques quantitatives*

Par les participants volontaires

*Un temps – en soirée - pourra être réservé à la consultation statistique.*

J5

J5 – 9h30-12h30

*Modélisation de données environnementales*

Par Franck Torre, AMU-IMBE & Maxime Logez, INRAE-RiverLy

Après un rappel sur tests d’hypothèses et le modèle linéaire, l’objectif de cette session est d’initier les utilisateurs aux modèles linéaires généralisés, GLM, à travers des exemples pratiques pris soit en sciences médicales soit en sciences environnementales. Très souvent de par la nature de la variable expliquée, l’hypothèse de sa normalité ne peut être envisagée et il convient d’utiliser d’autres outils statistiques que les modèles linéaires classiques. Les GLMs sont des extensions des modèles linéaires à des distributions non normales comme la loi de Poisson ou la loi Binomiale, adaptées à des variables de comptage ou des données de présence-absence (proportions). Pour pouvoir modéliser des variables avec de telles distributions nous aborderons au cours de cette session la régression de Poisson et la régression logistique.

J5 – après-midi

*Retour à Marseille*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Affiliations des conférenciers* | | | |  |
|  | Une image contenant texte, clipart  Description générée automatiquement |  | Une image contenant texte  Description générée automatiquement | Blog - DellaData |