Journées d'analyse statistique des données sur R
Edition 2024 délocalisée à l'Observatoire de Haute-Provence
(25-29 mars 2024)

Observatoire de Haute-Provence - 1912 Route de l'Observatoire
OSU Institut Pythéas - CNRS - AMU
04870 St.Michel l'Observatoire France
Téléphone : +33 (0)492 70 6540 - 43.93533075948819, 5.710592911138732

Photo : Oak Observatory at OHP (O3HP)

Inscriptions : à partir du site ECCOREV
La formation est ouverte aux chercheurs, enseignants chercheurs, doctorants des UMR affiliées à la fédération ECCOREV, aux Instituts d'établissement ITEM & OCEANS.

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. L'objectif est de rendre l'utilisateur débutant 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.

L'équipe de formateurs comprend : Alberte Bondeau (CNRS-IMBE), Claire Della Vedova (indépendante – société DellaData), Maxime Logez (INRAE – RiverLy), Laura March (IRD – LPED), Mathieu Santonja (AMU-IMBE) & Franck Torre (AMU-IMBE).

J1

J1 – 11h30-13h00
Accueil OHP et service de paniers repas (Maison Jean Perrin)

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.

Pause-café/biscuit

J1 – 17h00-18h30
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.

J1 – 18h30-19h30
Tidyverse
Par Laura March, IRD-LPED
Le terme tidyverse est une contraction de tidy (qu'on pourrait traduire par "bien rangé") et de universe. Il s'agit en fait d'une collection d'extensions conçues pour travailler ensemble et basées sur une philosophie commune. Elles abordent un très grand nombre d'opérations courantes dans R (la liste n'est pas exhaustive) : visualisation, manipulation des tableaux de données, import/export de données, manipulation de variables, extraction de données du Web, programmation …. Un des objectifs de ces extensions est de fournir des fonctions avec une syntaxe cohérente, qui fonctionnent bien ensemble, et qui retournent des résultats prévisibles. Elles sont en grande partie issues du travail d'Hadley Wickham, qui travaille désormais pour RStudio.

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 la biodiversité des prairies alpines et ii) sur l'impact de la gestion forestière sur les services écosystémiques rendus par les forêts méditerranéennes.

Pause-café/biscuit

J2 – 17h00-19h30
Consultation statistique sur des jeux de données proposés par les participants

J3

J3 – 9h30-12h30
Analyse spatiale 1/2
Par Alberte Bondeau, CNRS-IMBE

Cette demi-journée est une introduction au traitement des données géospatiales avec R. R permet de manipuler des données géoréférencées sous forme matricielle ou vectorielle, cette session présentera les différentes commandes permettant de traiter et de visualiser ces données sous formes de cartes. Les étapes nécessaires à la combinaison de données matricielles issues de sources différentes (parfois avec des projections différentes) avec des données vectorielles seront illustrées avec l'analyse de l'évolution du climat et du fonctionnement des écosystèmes sur la région Méditerranéenne. Quelques méthodes d'interpolation spatiale seront discutées.

J3 – 13h30 – 16h30
Après-midi récréative OHP

- Visite de la grande coupole (13h45-14h45)
- Visite O3HP (15h-16h)

A noter qu'un lâcher du ballon sonde ozone (mardi 26/02 à 11h) et d'autres animations agrémenteront notre semaine en mode geek.

Pause-café/biscuit

J3 – 17h00-19h30
Analyse spatiale 2/2
Par Alberte Bondeau, CNRS-IMBE

Cette session à pour but de permettre à tous les participants qui travaillent avec des données géoréférencées de réaliser des cartes pour visualiser « spatialement » leurs données, et de découvrir quelles analyses spatiales il est possible de réaliser.

J3 – Soirée
Consultation statistique sur des jeux de données proposés par les participants

J4

J4 – 9h30-12h30
Modélisation Dose réponse et SSD
Par Claire Della Vedova (société DellaData)
Dans cet atelier axé sur l'écotoxicologie, nous nous concentrerons sur la modélisation des relations dose-réponse au niveau des organismes individuels pour établir des paramètres d'écotoxicité tels que EC10 et EC50. Ensuite, nous appliquerons ces paramètres pour calculer une concentration de référence protectrice, la Hazard Concentration, pour les communautés écologiques à travers l'approche des distributions de sensibilité des espèces (SSD). Cet apprentissage permettra aux écologistes et gestionnaires environnementaux d'approfondir leurs compétences en modélisation environnementale et de renforcer leur capacité à prendre des décisions basée sur des données.

J4 – 13h30-16h30
Outils cartographiques
Par Claire Della Vedova (société DellaData)
Dans cette session les stagiaires se familiariseront avec les fondamentaux de la création de cartes administratives sous R. En exploitant des packages tels que ggplot2, ils apprendront à superposer des emplacements spécifiques sur des cartes, tout en personnalisant titres, couleurs, et autres éléments graphiques. L'avantage d'utiliser R pour cartographier des emplacements réside dans la possibilité d'intégrer la cartographie dans un continuum d'analyses environnementales et écologiques, assurant une transition fluide entre les différentes étapes de l'analyse sans quitter l'environnement logiciel R.

Pause-café/biscuit

J4 – 17h00-19h30
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.

J4 – Soirée
Consultation statistique sur des jeux de données proposés par les participants

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