Journées d'analyse statistique des données sur R
17-20 mai 2021

Marseille Saint Jérôme
Salle 212
50 places (25 en demi-jauge)
Bâtiment Dépt Langues Vivantes
Faculté Saint Jérôme - Marseille
43.33592003633092, 5.411042144178019

Journées R ECCOREV 2016 à Barcelonnette

Inscription en ligne sur le site ECCOREV
La formation est ouverte en priorité aux doctorants et jeunes chercheurs des UMR affiliées à ECCOREV.

Contexte COVID-19 et considérations pratiques. L'accès en voiture est autorisé sur la base d'un enregistrement préalable (plaque minéralogique, modèle et couleur du véhicule) à préciser au moment de l'inscription. L'inscription est gratuite. La restauration type snacking sur place par vente à emporter est possible sur le campus. L'organisation ne prévoit pas de prise en charge. Veillez à disposer d'un ordre de mission établi par votre employeur. La participation assidue à la formation est éligible au plan de formation de l'Ecole Doctorale ED251. *Le contexte sanitaire avait compromis l'édition 2020. Pour cette édition 2021, nous envisageons des sessions retransmises par visio-conférences en fonction de l'évolution de ce contexte mais aussi en cas d'affluence au-delà de la capacité d'accueil présentielle sur le campus de Saint-Jérôme à Marseille.*

Ces journées se dérouleront sur ordinateur, avec le logiciel R. 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 20 participants présentiel. Chaque participant amène son ordinateur portable personnel équipé wifi. La configuration logicielle sera précisée avant le début des journées et nous anticiperons l'installation des packages pour chaque

session. Le public visé sera en priorité : les thésards, post-docs et jeunes chercheurs des institutions membres d'ECCOREV ou ITEM, 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.
L'équipe de formateurs comprend : Alberte Bondeau (CNRS-IMBE), Joël Guiot (CNRS Emérite – Cerege), Maxime Logez (INRAE – RECOVER), Laura March (IRD – LPED) & Franck Torre (AMU-IMBE).

J1 – Lundi 17 mai 2021

J1 – 9h30-12h30
Introduction à R
Par Franck TORRE, AMU-IMBE & Maxime LOGEZ, INRAE-RECOVER
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.

J1 – 13h30-16h30
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 – Mardi 18 mai 2021

J2 – 9h30-12h30
Modélisation de données environnementales
Par Franck TORRE, AMU-IMBE & Maxime LOGEZ, INRAE-RECOVER
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.

J2 – 13h30-16h30
Notions avancées sur R
Par Maxime LOGEZ, INRAE-RECOVER
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 – Mercredi 19 mai 2021

J1 – 9h30-12h30
Graphiques ggplot
Par Maxime LOGEZ, INRAE-RECOVER
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.

J3 – 13h30-16h30
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 – Jeudi 20 mai 2021

J4 – 9h30-12h30
Analyse spatiale 2/2
Par Alberte Bondeau, CNRS-IMBE

Dans la continuité de la première demi-journée :
Visualisation et description de données spatiales. Hypothèses générales et modèles utilisés en géostatistique (utilisation de méthodes de simulations pour visualiser le potentiel et les limites du cadre théorique). Outils d'analyse de la variabilité spatiale: variogramme expérimental, fonction de covariance spatiale, choix de modèles et ajustement (présentation autour d'exemples). Méthodes d'interpolation par Krigeage (ordinaire et universel) dans des cas simples et univariés. Influence du choix du modèle et réflexion sur les types d'échantillonnage.

J4 – 13h30-16h30
Retour d'expérience et application à un cas d'étude sur le changement climatique, Analyses statistiques et rédaction de rapport dynamique sous R (notebook)
Par Joël GUIOT, CNRS-Cerege émérite

Le cas d'étude consiste à étudier les tendances des évolutions des températures depuis 1900 dans différentes parties du monde et à la comparer à la tendance globale. Ce cas permettra d'utiliser des fonctions de régressions, d'analyse spatiale et des graphiques sur ggplot2, mais le point le plus important consistera à utiliser Rstudio pour écrire des rapports basés sur les résultats des analyses grâce à son notebook. C'est une facilité très utile pour partager les résultats avec d'autres, mais également pour introduire de la flexibilité dans les analyses.

Affiliations des conférenciers

cerege
CENTRE EUROPÉEN DE RECHERCHE ET D'ENSEIGNEMENT DES GÉOSCIENCES DE L'ENVIRONNEMENT

imbe

INRAE

LPED
Laboratoire Population Environnement Développement
UMR 151 AMU-IRD