Journées d'analyse statistique des données sur R
Edition 2020

CEREGE, 2e étage porte ouest, salle 301 (géomatique)
16, 17, 18 & 20 mars 2020

Inscriptions : Joëlle Cavalieri, FR ECCOREV, tel 0442971521 - mel : cavalieri@cerege.fr
La formation est ouverte en priorité aux doctorants et jeunes chercheurs des UMR affiliées à ECCOREV.

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. On dispose au CEREGE de 20 ordinateurs en réseau WIFI. La salle étant relativement petite, on ne pourra accueillir qu'une trentaine d'étudiants par module. Le public visé sera en priorité: les thésards, post-docs et jeunes chercheurs des institutions membres d'ECCOREV, selon la place disponible. L'objectif est de rendre l'utilisateur suffisamment autonome dans les domaines les plus « populaires » de la statistique, de manière à ce qu'il soit capable de pratiquer directement les cas standards et d'être capable de personnaliser son approche.

J1 – Lundi 16 mars 2020 – Introduction & Initiation à la syntaxe de R

J1 – 9h30-12h30
Introduction à R
Par Franck TORRE, IMBE-AMU & Maxime LOGEZ, INRAE
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 1/2
Par Franck TORRE, IMBE-AMU
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. Des exemples provenant d'échantillonnage en écologie serviront d'illustration.

J2 – Mardi 17 mars 2020 – Analyse multivariée en environnement

J2 – 9h30-12h30
Analyse de données environnementales multivariées 2/2
Par Franck TORRE, IMBE-AMU
Les méthodes de couplage de tableaux permettent d'étudier le lien entre deux tableaux. On présentera l'analyse de coinertie qui permet d'étudier la structure commune à deux tableaux contenant différents descripteurs sur les mêmes individus. On présentera également les analyses multivariées explicatives type analyse de redondances (RDA) ou analyse des correspondances sous contrainte (CCA). 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 ….

J2 – 13h30-16h30
Notions avancées sur R
Par Maxime LOGEZ, INRAE
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 18 mars – Graphiques elaborés & Analyse spatiale

J1 – 9h30-12h30
Graphiques ggplot
Par Maxime LOGEZ, INRAE
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. 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 matin – Jeudi 19 ou vendredi 20 mars – Analyse spatiale,

J4 – 9h30-12h30
Analyse spatiale 2/2
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. 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
Modèle linéaire généralisé
Par Maxime LOGEZ, INRAE

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. Les modèles linéaires classiques, tels que la régression linéaire, l'ANOVA ou encore l'ANCOVA font l'hypothèse que la variable à expliquer (Y) suit une loi normale et que la moyenne de Y (l'espérance) varie selon une équation de la forme $E(Y_i|X_i) = a_iX_i+b$ avec X_i la ou les variable(s) explicative(s). 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.