Fédération de Recherche ECCOREV n° 3098

CNRS/Université Paul Cézanne Aix-Marseille

Europôle Méditerranéen de l'Arbois
Bâtiment du CEREGE BP 80
13545 Aix en Provence cedex 4

Direction : Joël Guiot
Tél : 04 42 97 15 32
guiot@eccorev.fr

Administration : Joëlle Cavalieri
Tél : 04 42 97 15 21 Fax : 04 42 97 15 47
cavalieri@eccorev.fr

Site internet : http://www.eccorev.fr/

Journées d'analyse statistique des données sur R
Edition 2014

CEREGE, 2e étage porte ouest, salle 301 (géomatique)

11-12 & 17-18 juin 2014

Inscriptions : Joëlle Cavalieri, FR ECCOREV, tel 0442971521 - mel : cavalieri@eccorev.fr

La formation est ouverte en priorité aux doctorants et chercheurs des laboratoires fédérés dans 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 par session. Les participants pourront sélectionner les sessions qui les intéressent, ce qui permettra d'optimiser le programme de chacun et d'accepter un maximum de participants.
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 chercheurs des laboratoires 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 de personnaliser son approche.

J1 – Mercredi 11 juin 2014 – Prise en main du logiciel R – Session 1

Packages requis : base !

J1 – Mercredi 11 juin 2014 – 9h30-12h30
Introduction à R
Par Franck Torre, IMBE-AMU
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 – Mercredi 11 juin 2014 – 13h30-16h30
Notions avancées
Par David Nérini & Clément Aldebert, MIO-AMU
Ce module s'adresse à des utilisateurs de R ayant déjà pratiqués les bases du logiciel. Au travers d'exemples pris en sciences de l'environnement, nous montrerons l'utilisation de certains packages ainsi que la création de fonctions sous R. 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. Toutes ces notions seront abordées en utilisant des méthodes statistiques classiques (régression multiple, ACP, …).

J2 – Jeudi 12 juin 2014– Analyse multivariée en environnement – Session 2

Packages requis : ade4, adegenet, vegan …

J2 – Jeudi 12 juin 2014 – 9h30-12h30
Analyse multivariée
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 – Jeudi 12 juin 2014 – 13h30-16h30
Analyse multivariée
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 ….

J3 – Mardi 17 juin 2014 – Analyse spatiale – Session 3

J3 – Mardi 17 juin 2014 – 9h30-12h30
Analyse spatiale
Par Pascal Monestiez, BIOSP-INRA Avignon (sous réserve)
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.

J3 – Mardi 17 juin 2014 – 13h30-16h30
Régression spatiale
Par Ghislain GENIAUX, ECODEVELOPPEMENT-INRA Avignon
Ce module traitera des régressions spatiales en présence de 1) dépendance spatiale et/ou 2) d'hétérogénéité spatiale, avec un focus sur les problèmes liés à l'estimation sur gros échantillons. 1) Estimation et tests de spécification des modèles avec dépendance spatiale (SAR, SEM, SDM et SARAR) à partir des packages spdep et sphet par maximum de vraisemblance et par la méthode des moments généralisés (GMM). La création et l'utilisation (estimation par GMM linéarisée) de matrices de voisinage spatial sur larges échantillons seront traités avec le package SLD; 2) Prise en compte de l'hétérogénéité spatiale non observée et de modèles à coefficient variable spatialement à partir de modèles géo-additifs (mgcv) ou de modèles localement pondérés (SLD).+ Diverses techniques d'import de données spatiales (rgdal, RPostgreSQL, ...).

J4 matin – Mercredi 18 juin 2014 – Régression, optimisation – Session 4

J4 – Mercredi 18 juin 2014 – 9h30-12h30
Régression et optimisation
Par David Nérini & Clément Aldebert, MIO-AMU
Dans le domaine des sciences de l'environnement, beaucoup de données peuvent être traitées comme des courbes (profil de résistivité, profils de température, évolution d'une population au cours du temps, …). On s'intéresse dans ce module à un ensemble de méthodes d'optimisation permettant de reconstituer ces courbes à partir d'échantillons de données ponctuelles. Il s'agira d'aborder des méthodes de régression non-paramétriques comme les splines ou des méthodes de régression à noyaux mais également les méthodes nécessaires à l'ajustement d'un modèle d'équations différentielles ordinaires à des données échantillonnées. Les travaux seront illustrés à partir de données océanographiques de température dans l'Océan austral mais aussi à partir de données expérimentales de suivi de populations zooplanctoniques en chémostat.

J4 ap midi – Mercredi 18 juin 2014 – Bioindicateurs et paléoécologie – Session 5

J3 – Mercredi 18 juin 2014 – 13h30-16h30

Méthodes de calcul de bioindicateurs environnementaux, application aux assemblages (paléo-)biologiques

Par Joël Guiot, CEREGE-CNRS

La recherche d'indicateurs environnementaux à partir de données biologiques implique l'utilisation de méthodes d'analyse multivariées, linéaires ou non-linéaires. Même si d'autres packages sont également disponibles, nous nous focaliserons sur le package bioindic. Nos applications seront tirées de la paléoclimatologie, paléoécologie et dendroclimatologie. Nous traiterons des problèmes suivants :

1) méthodes de synthétisation des assemblages (ACP) et standardisation des données

2) fonctions de réponse de la végétation (de la croissance des arbres) au climat : régression linéaire, réseaux de neurones

3) fonctions de transfert permettant de reconstruire le climat (et tout paramètre environnemental) à partir d'assemblages paléoécologiques sur les continents (pollen, insectes, mollusques, diatomées) et les océans (foraminifères, coccolithes, dinocystes, diatomées). Seront abordées les méthodes de calibration (régression, PLS, WA-PLS, GAM, réseaux de neurones) et les méthodes de similarité (meilleurs analogues, surfaces de réponse).