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@cerege.fr

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

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

Aix en Provence, le mardi 26 mai 2011

Journées d'analyse statistique des données sur R

Quand ? 14-15-23 juin 2011

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

Inscription : Joëlle Cavalieri, FR ECCOREV, tel 0442971521,
mel : cavalieri@cerege.fr

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 module. Les participants pourront sélectionner les modules qui les intéressent, ce qui permettra d'optimiser les présences.
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 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.

_____Mardi 14 juin _____

J1 mardi 14 juin 9h-12h30 : **Introduction** à R par F Torre (UPCAM, IMEP)

MODULE 1: 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 mardi 14 juin 13h30-17h : **Introduction à l'Analyse de Données Fonctionnelles** par A Malkassian (UM, LMGEM)

MODULE 2 : En océanologie et plus généralement en Sciences de l'Environnement, nombre de données peuvent être considérées comme des courbes indicées par le temps ou tout autre variable (données fonctionnelles). Citons par exemple des séries temporelles, des profils verticaux de salinité, de conductance dans un sol, des spectres de taille d'organismes vivants (phytoplancton), etc ... La plupart des méthodes permettant de comparer entre eux des individus d'un échantillon statistique fait appel aux méthodes classiques d'analyses multivariées déclinées suivant plusieurs variantes qui dépendent du type de données à traiter (ACP pour des variables quantitatives, AFC pour des tableaux de contingence, Classifications,…). Dans le cas de données fonctionnelles, ces méthodes ne

sont pas immédiatement applicables. La raison principale est que les méthodes classiques sont invariantes si l'on permute les variables du tableau à analyser. Il faut donc tenir compte de cette contrainte lorsque les variables sont ordonnées. Il arrive également que l'on ne puisse pas constituer initialement le tableau sur lequel va porter l'analyse, par exemple, lorsque des courbes constituant un ensemble d'individus à comparer, n'ont pas été échantillonnées aux mêmes points.
Nous proposons lors de cette séance d'apporter des solutions aux deux problèmes soulevés précédemment à partir d'exemples de profils physico-chimiques échantillonnés lors de campagnes océanographiques. Il s'agira dans un premier temps, d'introduire des méthodes de lissage (LOESS, splines) permettant de boucher des trous dans des séries incomplètes. Puis nous aborderons simultanément des méthodes d'analyse multivariée et de classification dans le cas de ces profils via l'utilisation du package FDA.

_____Mercredi 15 juin _____

J2 mercredi 15 juin 9h-12h30: **Analyse à un tableau** par F. Torre (UPCAM, IMEP)

MODULE 3a: 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 mercredi 15 juin 13h30-17h : **Couplage de tableaux** par F. Torre (UPCAM, IMEP)

MODULE 3b : Les méthodes de couplage de tableaux permettent d'étudier le lien entre deux tableaux. On présentera l'analyse de co-inertie 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).

_____Jeudi 23 juin _____

J3 jeudi 23 juin 9h-12h : **Programme libre** par F . Torre (UPCAM, IMEP)

MODULE 4 : Cette matinée pourra être consacrée au traitement de jeux de données proposés par les participants. A défaut, une analyse-type sera présentée : importation du jeu de données depuis une base de données – graphiques – premières analyses statistiques en insistant sur la méthode de gestion des données, l'écriture de scripts, l'utilisation de l'aide en ligne et plus généralement des ressources du web. Un temps sera consacré à des questions/réponses sur les séances déjà écoulées.

J3 jeudi 23 juin 13h30-17h : **Statistiques spatiales** par P. Monestiez (INRA Avignon)

MODULE 5 : 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.