# Journées d'analyse statistique des données sur R
# Edition 2015

**CEREGE, 2e étage porte ouest, salle 301 (géomatique)**
**18-19 mai & 26-27 mai 2015**

**Inscriptions : Joëlle Cavalieri, FR ECCOREV, tel 0442971521 - mel : cavalieri@eccorev.fr**
**La formation est ouverte en priorité aux doctorants et chercheurs des laboratoires affiliés 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 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 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 18 mai 2015 – Initiation à R et régression spatiale

J1 – 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 – 13h30-16h30
*Régression spatiale*
Par Ghislain GENIAUX, ECODEVELOPPEMENT-INRA Avignon
L'usage de méthode d'analyse statistique de données géolocalisées à partir de régression linéaire demande d'adapter les estimateurs à l'éventuelle présence d'autocorrélation spatiale (corrélation des résidus et /ou de la variable à expliquer avec ses voisins) et/ou d'hétérogénéité spatiale (variation des relations entre variables dans l'espace). Ce module traitera des méthodes de régression en présence de dépendance spatiale et/ou d'hétérogénéité spatiale, avec un focus sur les problèmes liés à l'estimation sur gros échantillons. On présentera dans un premier temps le concept de matrice de poids spatial qui est au cœur de ces modèles et comment les créer manuellement et à partir des packages existants. On abordera ensuite l'estimation et les tests de spécification des principaux modèles de régression avec dépendance spatiale, à partir des packages « spdep » et « sphet », par maximum de vraisemblance et par la méthode des moments généralisés. Enfin, pour estimer les modèles à coefficients variables dans l'espace nous aborderons les méthodes de régression semi-paramétriques qui permettent aux coefficients estimés de varier selon leurs coordonnées spatiales, soit par régression locale (Geographically Weighted Regression *du package « GWmodel ») ou par modèles géo-additifs (Generalized Additive Model du package « mgcv »). Diverses techniques d'import de données spatiales (rgdal, RPostgreSQL, ...) seront également proposées.*

## J2 – Mardi 19 mai 2015 – Analyse multivariée en environnement –

J2 – 9h30-12h30
*Analyse multivariée*
Par Franck TORRE, IMBE-AMU
L'analyse de données permet de mettre en évidence l'information contenue 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 – 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 26 mai 2015 – Notions avancées sur R et analyse spatiale

J3 – 9h30-12h30
*Notions avancées*
Par David NERINI, 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, …).

J3 – 13h30-16h30
*Analyse spatiale*
Par Denis ALLARD, BIOSP-INRA Avignon

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 – Mercredi 27 mai 2015 – Régression, optimisation & modèle linéaire généralisé

J4 – 9h30-12h30

*Régression et optimisation*

Par David NERINi & 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 – 13h30-16h30

*Modèle linéaire généralisé*

Par Maxime LOGEZ, IMBE-AMU

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 $Y = a_iX_i+b$ 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.