Analyse multivariable
Université Laval, Québec, mars-avril 2008

Dr. Daniel Borcard
Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre ville
Montréal QC H3C 3J7
Canada
daniel.borcard@umontreal.ca

Forme du cours
L'ensemble du cours prend 7 jours, qui comprennent des périodes de théorie couvrant les grands domaines de l'écologie numérique, des démonstrations de programmes d'ordinateur et des applications. Ces dernières se feront d'une part sur des jeux de données de démonstration communs à tous les étudiants (période d'apprentissage), et d'autre part sur des jeux de données apportés par les participants (période d'application).

Table des matières du cours théorique (5 matinées de 8h30 à 11h30)

Chapitre 1: les données
Programmes d'ordinateur
Définitions
Les matrices de données: catégories et transformations

Chapitre 2: association
Matrices d'association: mode Q, mode R
Types de descripteurs
Le problème du double zéro
Coefficients de similarité
Coefficients de distance

Chapitre 3: groupement
Vue d'ensemble
Groupements à liens
Groupements moyens
Groupement de Ward
Partitionnement *K*-means

Chapitre 4: ordination en espace réduit
Première partie (4a): ordination simple
Généralités
Analyse en composantes principales (PCA)

Pré-transformation de données
Analyse factorielle des correspondances (CA)
Analyse en coordonnées principales (PCoA)
Cadrage multidimensionnel non-métrique (NMDS)

Deuxième partie (4b): ordination canonique

Analyse de redondance (RDA)
Analyse canonique des correspondances (CCA)
Partitionnement de la variation (Borcard et al. 1992)
Analyse canonique partielle
db-RDA
ANOVA multivariable par RDA

Chapitre 5: les tests statistiques pour données multivariables

Tests par permutation
Tests de la RDA et de la CCA
Test de Mantel

Chapitre 6: analyse spatiale des données écologiques

Introduction
La mesure de l'autocorrélation spatiale
La modélisation des structures spatiales
Trend surface analysis
Coordonnées principales de matrices de voisinage (PCNM, Borcard & Legendre 2002)

Jour 1 (26 mars): chapitres 1 et 2
Jour 2 (27 mars): chapitre 3
Jour 3 (28 mars): chapitre 4a
Jour 4 (1er avril): chapitres 4b et 5
Jour 5 (2 avril): chapitre 6
Jour 6 (3 avril): présentations des étudiants
Jour 7 (4 avril): travaux pratiques sur données personnelles

Pratique

La pratique sur ordinateur sera axée sur l'apprentissage de l'utilisation du langage statistique R (http://cran.r-project.org/), qui allie souplesse et universalité. Ce langage est en train de devenir le standard mondial en statistique, et est adopté par un très grand nombre d'institutions universitaires. Il est gratuit et existe pour toutes les plateformes principales (PC, Max OSX, Linux…). Le langage R comprend des librairies permettant de réaliser la quasi-totalité des analyses démontrées au cours. Son apprentissage ne nécessite aucune connaissance préalable en programmation.

Afin d'assurer un encadrement de qualité aux étudiants, le professeur sera secondé par Marie-Hélène Ouellette, doctorante au département de sciences biologiques de l'Université de Montréal, une démonstratrice qualifiée, chercheuse en écologie numérique, connaissant bien le langage R et les analyses multivariables enseignées au cours.

Phase d'apprentissage des méthodes: 5 après-midis correspondant aux 5 matinées de théorie.

1. Présentation du langage R; manipulation de données; statistiques de base en R.
2. Matrices d'association et groupement en R.
3. Ordination simple en R.
4. Ordination canonique en R
5. Analyse spatiale en R.

Le 6^{e} jour est consacré à des présentations préparées par chaque étudiant. En 10 minutes, l'étudiant doit présenter brièvement sa problématique, ses hypothèses, ses données, et les analyses qu'il envisage de réaliser. Pour 20 étudiants, l'ensemble des présentations se ferait en 4 périodes de 50 minutes. Le reste de la journée sera consacré à la mise en place technique des analyses des données personnelles, avec discussion avec le professeur et la démonstratrice.
Le dernier jour sera consacré aux analyses de données personnelles. Au plus tard 2 semaines après la fin du cours, l'étudiant devra remettre (par courrier électronique) un court rapport d'analyse présentant sa problématique, ses hypothèses et ses données (1 page), suivies des analyses qu'il a réalisées durant et après le cours, ainsi que l'interprétation des résultats (nombre de pages dépendant des analyses; maximum 6).
La formule proposée nécessite un travail de préparation de l'étudiant, à la fois pour la présentation et pour les données. Cette option a le mérite de permettre à l'étudiant d'appliquer à ses propres données les méthodes vues au cours, sans court-circuiter les nécessaires étapes d'apprentissage des programmes.

En support au cours et aux travaux pratiques, le professeur monte une page web fournissant un document de notes de cours, les données, les scripts d'analyse en langage R, et tout autre matériel (p.ex. tirés-à-part en format pdf) jugé utile.

Le professeur

Daniel Borcard est titulaire d'un doctorat en écologie de l'Université de Neuchâtel (Suisse). Après un stage post-doctoral en analyse multivariable (écologie numérique) d'un an chez le Professeur Pierre Legendre (Université de Montréal), il a enseigné cette discipline en Suisse durant 7 ans. Depuis 1999 il enseigne la biostatistique à l'Université de Montréal. La Faculté des Arts et des Sciences de l'U de M lui a décerné son Prix d'Excellence en enseignement (2005). Il a donné de nombreux cours intensifs d'écologie numérique au Canada et à l'étranger, dont un à l'Université Laval à titre de Professeur invité en 2006. Depuis 1989, à part ses recherches en écologie des communautés, il poursuit aussi des recherches en écologie numérique. Il est l'auteur principal de deux techniques d'analyse qui connaissent un très grand succès en écologie numérique: le partitionnement de variation de matrices de données multivariables et l'analyse spatiale par coordonnées principales de matrices de voisinage (PCNM). Son Science Citation Index en écologie numérique est d'environ 900.