LES RENCONTRES DE STATISTIQUE APPLIQUÉE

Palette d'applications sous R

Jeudi 28 avril 2011 de 9h15 à 17h30

Institut Henri Poincaré (amphithéâtre Hermite)

PROGRAMME	
9h15	Accueil
9h30	Élisabeth Morand (Ined) et Jérôme Sueur (MNHN, UMR CNRS 7205 OSEB Paris) • Introduction
9h50	Maxime Hervé (UMR INRA - Agrocampus Ouest – Univ. Rennes I BiO3P) • Débuter avec R L'objectif de cette présentation est de donner envie d'utiliser R. Nous verrons que (i) malgré les apparences R n'est pas difficile à utiliser et (ii) qu'on y trouve beaucoup d'intérêts. Après une brève description de ce qu'est R, nous donnerons quelques exemples pour montrer que la syntaxe du langage est très intuitive et facile à prendre en main. Puis nous montrerons toute la puissance de R en parlant des packages additionnels. L'accent sera mis sur les aides disponibles, à la fois hors-ligne et sur internet. Et nous verrons enfin que R est vraiment de plus en plus utilisé dans la communauté scientifique.
10h20	Jérôme Sueur (MNHN, UMR CNRS 7205 OSEB Paris) • Optimiser ses représentations graphiques L'analyse de données est inconcevable sans un support graphique de qualité. La visualisation des résultats est un exercice (art) qu'il est (assez) facile de pratiquer avec R. Les fonctions graphiques de R autorisent une grande liberté pour la construction de figures ou de planches de figures. Nous verrons plus particulièrement les possibilités offertes par la librairie ggplot2 au succès grandissant. Reposant sur une grammaire graphique originale, ggplot2 permet de produire des graphiques complexes et originaux en quelques lignes de code.
	Pause
11h10	Raymond Baudoin (MNHN, Inventaire et Suivi de la Biodiversité Paris) • Accéder à une base de données Le chargement des données est la phase incontournable pour l'utilisation des puissants outils de traitements qu'offre la majorité des librairies R. Mais bien souvent ces données sont créées et gérées par une autre application, tableur MS Excel, logiciel de base de données… Le passage par un fichier intermédiaire issu d'une extraction qui est ensuite lu dans R peut être, dans bien des cas, évité par une connexion directe à l'application gérant les données. Avec un serveur de données comme Oracle, Postgres, MySQL, MS-Access, respectant la norme SQL (Structured Query Language - langage normalisé de contrôle et interrogation des bases de données relationnelles), l'utilisation de la commande SELECT offre la possibilité de ne prendre en compte que certaines variables (champs) et de sélectionner sur les contenus (valeurs) pour obtenir le tableau souhaité (data.frame) et unique même dans le cas d'un schéma relationnel complexe (n tables). Les librairies utilisées: · pour la connexion à des bases de données : RODBC, SQLiteDF, RMySQL, RPostgreSQL, ROracle, DBI · pour accéder aux feuilles de classeurs MS-Excel : RODBC, xlsReadWrite, xlsx · pour utiliser la commande SELECT sur des fichiers ou des data frames : sqldf
11h40	Christophe Genolini (Univ. Paris X) • Créer son propre package R Programmer, c'est bien. Partager, c'est mieux ! En R, le partage de vos routines se fait via un format particulier : le package. Ecrire un package est une étape obligatoire dans le processus de diffusion de votre programme. Hélas, c'est une procédure assez obscure et fort peu documentée. Dans cette présentation, nous nous attacherons donc à baliser un peu le long et épineux chemin conduisant à la création d'un package. De votre package ?
12h00	Discussion

14h15	Alexis Gabadinho (Dept.d'Econométrie et Laboratoire de Démographie et d'Etudes Familiales de l'Université de Genève) • L'analyse de séquences dans R avec la librairie TraMineR TraMineR (contraction de Life Trajectory Miner for R) est une librairie R destinée à l'analyse et la visualisation de séquences d'états ou d'événements. Cette librairie est plus particulièrement développée pour l'analyse de données biographiques longitudinales issues des sciences sociales, telles que des trajectoires professionnelles ou familiales, mais les outils proposés sont susceptibles de s'appliquer à tous types de séquences catégorielles. TraMineR propose un ensemble unique de fonctionnalités utilisables par des personnes non-spécialistes de R: - Prise en charge de différents formats de données séquentielles; - Fonctions graphiques pour la représentation d'ensembles de séquences ("index plot", "frequency plot", "distribution plot", "representative sequence plot", etc ...); - Caractéristiques longitudinales de séquences individuelles (entropie, turbulence, indice de complexité, etc ...); - Séquences de mesures transversales par position (distributions des états, séquence des états modaux, entropie transversale, etc ...); - Calcul de distances entre séquences avec un choix de métriques ("Optimal Matching", "Longest Common Subsequence", "Longest Common Prefix", "Hamming", "Dynamic Hamming"); - Analyse de type ANOVA et arbres de régression à partir de matrices de distances entre séquences; - Analyse de séquences d'événements (sous-séquences fréquentes, règles d'association, sous-séquences discriminantes, etc...).
14h45	Yves Tillé (Institut de statistique, Univ. de Neuchâtel) • Sondages avec le package Samplig Le package « sampling » écrit en langage R a été originellement développé dans le cadre de cycles de formation de statisticiens publics. Il a ensuite été soumis et accepté comme package officiel du « R Project for Statistics Computing ». Ce package contient un ensemble de fonctions permettant se sélectionner des échantillons avec des plans de sondages complexes, de réaliser des calages et des calages généralisés, d'estimer des paramètres et leur variances, de traiter les problèmes de non-réponses. Il offre donc un ensemble d'outils modernes pour réaliser un traitement d'enquête.
15h20	Michel Baylac (MNHN, UMR CNRS 7205 OSEB Paris) • Exploration et discrimination des formes L'analyse des formes biologiques (morphométrie) s'est fortement développée depuis les années 80 au travers de la morphométrie géométrique. Cette dernière analyse les formes représentées par des coordonnées de points homologues ou de contours, par opposition à la morphométrie classique qui traite exclusivement de distances ou de ratios. Sous R, il existe plusieurs packages (Shapes, Rmorph) ou ensembles de fonctions (Claude, 2008) de morphométrie géométrique. Rmorph sera utilisé pour montrer la puissance de ces méthodes, appliquées à des structures 2D ou 3D et dans des contextes variés de la biologie et de la systématique : allométries, symétries, évolution des formes, partitions et discriminations, exploration des espaces morphologiques.
	Pause
16h00	Franck Picard (CNRS, laboratoire Biometrie et Biologie Evolutive, UCB Lyon I) • Modèles de détection de ruptures et applications avec le logiciel R Les modèles de détection de ruptures permettent de dater ou de localiser des changements abrupts dans des données bruitées organisées au cours du temps ou dans l'espace. C'est un sujet ancien qui connait un nombre d'applications considérable (industrie, biologie, sciences du climat) et qui a fait l'objet de nombreux développements méthodologiques. Lorsqu'on s'intéresse à ces modèles, la démarche statistique se décompose généralement en trois étapes: quelle(s) caractéristique(s) du signal sont concernées par les changements abrupts ? Combien y-a-t-il de changements ? Et enfin où sont-ils ? L'utilisation pratique de ces modèles repose sur des développements logiciels permettant de les appliquer à des données spécifiquement. Nous avons développé un package R, cghseg, pour l'application de ces modèles à la biologie moléculaire. Au cours de l'exposé j'expliquerai les choix faits lors de la construction du package ainsi que les limites de R s'agissant de l'étape de localisation des ruptures. En effet cette étape nécessite l'utilisation d'algorithmes dont la complexité est raisonnable en théorie, mais limitante en pratique. Se pose alors la question d'interfacer R avec d'autres langages comme C++ pour des étapes nécessitant plus de puissance de calcul.
16h30	Élisabeth Morand (Ined) et Jérôme Sueur (MNHM, UMR CNRS 7205 OSEB Paris) • Synthèse : R
17h00	Débat et discussion

Chaque trimestre, le Service Méthodes Statistiques de l'Ined propose un séminaire de statistique appliquée ; cette session spéciale est organisée en étroite relation avec la Société Française de Statistique et le groupe d'utilisateurs de R du Muséum national d'Histoire naturelle. Le séminaire est ouvert à tous, sans frais de participation.