



LES RENCONTRES DE STATISTIQUE APPLIQUÉE

Institut National d'Études Démographiques
(Salle Alfred Sauvy)
Lundi 3 Décembre 2012
9h – 12h30



Récolter, explorer et visualiser ses données avec R : applications en sciences sociales

Séminaire organisé conjointement par le Service des méthodes statistiques de l'Ined et le séminaire Semin-R (<http://rug.mnhn.fr/semin-r/>)

9h30 : **Elisabeth MORAND** (Ined) et **Christophe POUZAT** (Laboratoire de Physiologie Cérébrale-CNRS UMR 8118-UFR biomédicale Paris V)

Introduction de la séance

9h45 : **Eric MATZNER-LOBER** (Université Rennes II)

Pourquoi R devient incontournable en recherche, enseignement et développement

Nous aborderons le logiciel R sous trois aspects : son utilisation en enseignement, en recherche et dans le monde de l'entreprise. A chacune de ces trois utilisations, il est nécessaire d'évaluer les demandes spécifiques du domaine et les réponses qu'apporte `\textsf{R}`.

Rappelons brièvement que `\textsf{R}` est multi-plateforme et multi-OS. Il est entièrement gratuit, très complet et offre à la fois des commandes mais aussi des menus déroulants. Il est donc raisonnable de penser que ce logiciel fera partie des logiciels de statistique les plus enseignés.

Sa facilité de programmation et sa forte utilisation dans le monde de la recherche en font dès aujourd'hui un langage omniprésent. On peut donc s'interroger sur les futures évolutions de `\textsf{R}` vis à vis de la recherche en statistiques.

Dans la troisième partie, nous comparerons R avec ses différents concurrents et analyserons les points qui gouvernent les choix de logiciels en entreprise (prix, interface graphique, intégration dans les bases de données...). Bien évidemment le logiciel est loin d'être parfait mais il comporte dès à présent des avantages qui lui valent d'être adopté par un nombre croissant d'entreprises.

10h15 : **Stéphane DRAY** (Laboratoire de Biométrie et Biologie Evolutive-Université Lyon I)

Ade4 : outils statistiques et graphiques pour l'analyse de données multivariées

Dans de nombreux domaines (écologie, sciences humaines et sociales, marketing, santé, etc.), de l'information est collectée et stockée sous la forme de tableaux contenant les mesures de plusieurs variables pour un grand nombre d'individus. Les méthodes d'analyse multivariée fournissent un ensemble d'outils permettant de résumer l'information contenue dans de grands volumes de données en identifiant les relations entre les variables, et les similarités entre les individus (Lebart et al., 1995). Les résultats issus de ces méthodes sont présentés sous la forme de graphiques, pour un nombre réduit de dimensions, permettant une exploration des principales structures identifiées dans les données.

Depuis 2002, le package `ade4` (Dray and Dufour, 2007) est développée au Laboratoire de Biométrie et Biologie Evolutive (Université Lyon 1). Il contient des procédures d'analyse multivariée pour le traitement d'un seul (Chessel et al., 2004), de deux mais aussi de K tableaux (Dray et al., 2007).

A ce jour, une quarantaine de méthodes ont été implémentées, dont près de la moitié ont été développées par les auteurs du package. Une quarantaine

de fonctions graphiques sont également disponibles afin de représenter les résultats issus de ces analyses. Au cours de cet exposé, le cadre théorique, les principales fonctionnalités et les futurs développements du package seront présentés et illustrés au travers d'un certain nombre d'exemples.

10h45 : Pause

11h00 : **Timothée GIRAUD** (UMS 2414 Riate) et **Marta SEVERO** (Geriico - Université Lille III)

Visualisation de données géographiques et médiatiques avec R

La combinaison d'informations géographiques et d'informations médiatiques peut s'avérer particulièrement fructueuse pour l'analyse des événements médiatiques internationaux et notamment pour l'étude de leur propagation spatio-temporelle. L'objectif de cette communication est de présenter les caractéristiques de la donnée médiatique et de montrer les possibilités que le logiciel R offre pour le traitement et la visualisation de ce type de données.

Nous présenterons quelques exemples faisant appel à l'interactivité, à l'animation et à la cartographie à travers l'analyse de la couverture médiatique internationale du conflit de Wukan, un village du sud de la Chine, où après des mois de protestations contre l'accaparement illégal de terres les habitants ont obtenu des élections libres.

11h30 : **Pierre RATINAUD** (Département des Sciences de l'Education et de la Formation - Laboratoire LERASS-Université Toulouse II)

Visualisation de données textuelles avec R et IRaMuTeQ

Nous présenterons les différentes possibilités offertes par le logiciel iramuteq et quelques bibliothèques de R pour représenter graphiquement des données textuelles. Plusieurs méthodes d'analyses seront évoquées. La représentation simple des effectifs d'apparition des formes se fera sous forme de courbes ou de nuages de mots (Fellows, 2012). Nous développerons particulièrement les graphiques liés à la représentation des classifications, notamment dans le cas de la méthode ALCESTE (Reinert, 1983, 1990 ; Ratinaud & Déjean, 2009) et soulignerons l'intérêt des graphes de similitude et de distance dans ce champ (Marchand & Ratinaud, 2012). Les apports de la 3D à ces représentations seront signalés.

12h00 : Discussion et conclusion

Chaque trimestre, le Service Méthodes Statistiques de l'Ined propose un séminaire de statistique appliquée. Le séminaire est ouvert à tous, sans frais de participation.
http://www.ined.fr/fr/rendez_vous/rencontres_statistique_appliquee/